이학박사학위논문

드노보 전장유전체 조립 및 분석을 통한 해양 절지동물 진화 사례 연구

The case studies on the evolution of marine arthropods through *de novo* genome assemblies and analyses

2020년 8월

서울대학교 대학원

생명과학부

정 진 협

ABSTRACT

The case studies on the evolution of marine arthropods through *de novo* genome assemblies and analyses

Jin-Hyeop Jeong
School of biological sciences
The graduate school
Seoul National University

The *de novo* genome assembly has become an essential approach for studying non-model organisms since the post-genome era arrived. The reported cases of *de novo* genome assemblies of non-model arthropods have increased dramatically in recent days. The marine arthropod, however, is one of the least sequenced animal groups despite of their surprisingly high taxonomic and morphological diversity. The *de novo* genome studies on these marine arthropods remain mostly limited in terms of their cases and quality of assemblies up to now. This study therefore conducted the first case of *de novo* genome research focusing to the under-sampled marine arthropod

groups, the Class Pycnogonida and the Infraorder Brachyura in Korea. In this study, one mitochondrial genome and four whole-genomes were *de novo* assembled and their genomic characteristics were discussed. While the two cases of *de novo* genomes assembled by using short read-length sequencing showed limited assembly quality, the long read-length based assemblies of *Nymphon striatum* and *Chionoecetes opilio* provided significantly informative, high-qualitied genomes. The preliminary phylogenomic research of this study which firstly included the representative genomes of pycnogonid and brachyuran decapod, also implied that recent hypothesis of xiphosuran nested in the most derived clade, Arachnopulmonate, is indeedly plausible. Furthermore, the limitations of *de novo* genome researches on the laboratory experiment lacking bioinformatics background were discussed to establish an optimized research workflow for the genomic study on non-model marine arthropod.

Key words: comparative genomics, *de novo* genome assembly, marine arthropods, mitochondrial genome, whole-genome, phylogenomics

Student number: 2015-22644

CONTENTS

LIST OF FIGURES

LIST OF TABLES

BACKGROUNDS

BACKGROUNDS

General Backgrounds

Whole-genome sequencing is defined as a procedure which determines the entire genomic nucleotides of an organelle or an organism by connecting relatively short, fragmented shotgun genomic reads (Paszkiewicz and Studholme, 2010). This concept had greatly affected on the entire fields of biology ever since the completion of the draft human genomic map in 2003, which was declared by the Human Genome Project team. Also, the development of the Next-generation sequencing technologies is another great milestone for the genomic researches. These early technologies of "Next-generation sequencing", such as the Pyrosequencing of 454 Life Sciences and Illumina's sequencing-by-synthesis, had enabled to generate massive amounts of decoded nucleotides in parallel, which was impossible for the Sanger sequencing used in Human Genome Project (Mardis, 2008). In past decades, the project cost per a three billion bases, or 3Gb-sized genome has dramatically reduced from 100 million US dollars in around 2001 to only 1,000 US dollars in around 2016, according to the data provided by the National Human Genome Research Institute (National Human Genome Research Institute, 2019). Nevertheless, sequencing nearly human-sized genomes with such low cost is only available for a few completed model-organisms by "Resequencing" technology, which greatly reduce the minimum required coverage for determining a genome via mapping against the already completed, "reference genome" of the same species. Therefore, the *de novo* genome sequencing and assembly are required to decode the whole-genome of non-model organisms which assembles the genome from fragmented sequencing reads without any reference genome (Ellegren, 2013).

There are two major differences between reference genomes of model organisms and *de novo* genomes of non-model organisms; the quality of genomic assembly and annotation. The quality of genomic assembly is often measured with “3 Cs” criteria, which contains contiguity, completeness, and correctness of the assemblage (Studholme, 2015; https://www.pacb.com/blog/beyond-contiguity/). Due to the limitation of read-length of Next-generation sequencing technologies, *de novo* assembled genomes may have too much number of fragmentary genomic sequences; or low contiguity, parts of coding or non-coding structural genomic regions not assembled; or low completeness (Narzisi and Mishra, 2011; Studholme, 2015). Moreover, false positive cases of indels or translocations might be resulted from miss-assembled genomic sequences with low correctness which can misinform genomic annotation or further evolutionary or comparative genomics studies (Phillippy et al., 2008; Meader et al., 2010). The genomic annotation is the another challenge of *de novo* genome research, which is conducted by computer-based prediction of the genomic structure, such as genes, repetitive elements, single nucleotide variations (SNPs), and the functions from these components (Stein, 2001; Iliopoulos et al., 2003; Reese et al., 2003). Needless to say, both the quality of genomic assembly and annotation are critical for *de novo* researched non-model organismal genomes, however they are considered as the bottleneck of the workflow of those researches which cost enormous cost and time (Phillippy et al., 2008).

In 2013, there were 215 genomes of non-model animal species which had been reported (Ellegren, 2013). According to this study, only two marine arthropods were reported with their genomes sequenced, among the 77 sequenced arthropod species. In addition, the Wikipedia article which is titled as “List of sequenced animal genomes” (https://en.wikipedia.org/wiki/List_of_sequenced_animal_genomes/, Lastest update at

2020.05.24., Retrieved at 2020.06.01.) was referred for detailed further investigation on the statistics of sequenced animal genomes. From its list of the animal species, over 505 animal species have had their genomes sequenced *de novo* with literature publications (**Table 1**, **Appendix Table 1**).

Table 1. List of Sequenced animal genomes with taxonomical and habital information, modified from Wikipedia article (Latest update at 2020.05.24. Retrieved at 2020.06.03.)

Clades	**Phylum / Subphylum**	**Class**	**No. of genomes**	**"Marine genomes"**
Porifera	Porifera	Demospongiae	3	3
Eumetazoa	Ctenophora	Tentaculata	2	2
Parahoxozoa	Placozoa	N/A	2	2
	Cnidaria	Anthozoa	7	7
		Cubozoa	1	1
		Hydrozoa	2	2
		Scyphozoa	4	4
		Staurozoa	1	1
Deuterostomia	Hemichordata	Enteropneusta	2	2
	Echinodermata	Asteroidea	1	1
		Echinoidea	1	1
		Holothuroidea	1	1
	Chordata / Urochordata	Ascidiacea	2	2
		Appendicularia	1	1
	Chordata / Cephalochordata	Leptocardii	1	1
	Chordata / Vertebrata	Hyperoartia	1	1
		Chondrichthyes	5	5
		Actinopterygii	40	24
		Sarcopterygii	1	1
		Amphibia	8	0
		Reptilia	21	4
		Aves	96	25
		Mammalia	113	10

Table 1. Continued from the previous page

Protostomia	Arthropoda / Hexapoda	Insecta	102	0
	Arthropoda / Crustacea	Hexanauplia	2	2
		Branchiopoda	2	0
		Malacostraca	4	3
	Arthropoda / Chelicerata	Merostomata	2	2
		Arachnida	9	0
	Arthropoda / Myriapoda	Chilopoda	1	0
	Tardigrada	Eutardigrada	1	0
	Mollusca	Bivalvia	14	13
		Cephalopoda	5	5
		Gastropoda	6	3
	Platyhelminthes	Cestoda	7	0
		Rhabditophora	2	0
		Trematoda	1	0
	Nematoda	Chromadorea	21	0
		Enoplea	4	0
Protostomia	Annelida	Polychaeta	1	1
		Clitellata	2	0
	Brachiopoda	Lingulata	1	1
	Rotifera	Eurotatoria	1	0
Total			504	131

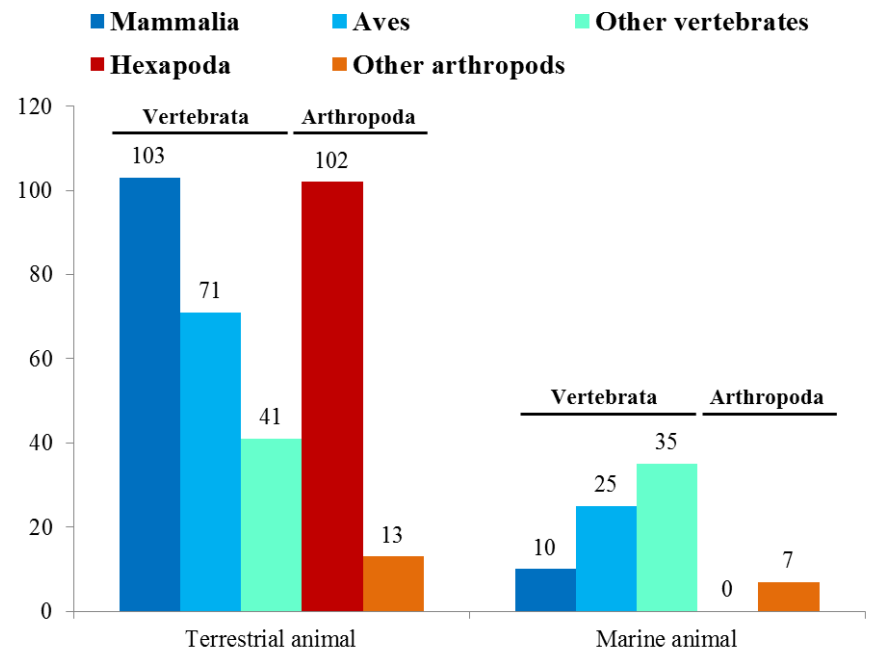

Figure 1. The numbers of terrestrial and marine species of Subphylum Vertebrata and Phylum Arthropoda whose genome sequences were *de novo* assembled.

Among those cases, species of Class Aves and Class Mammalia, which are the two most thoroughly sequenced animal clades, occupied 96 and 113 cases of sequenced genomes, respectively. The species of Subphylum Hexapoda record as 103 cases of sequenced genomes, thus the sum of sequenced birds, mammals and hexapods occupies 61.78% of the entire cases of sequenced animal genomes. On the other hand, when those animal species are categorized by their known habitats, terrestrial species including the species inhabiting in fresh water environments record as 374 cases, so that they occupied 74.06% of the total known cases. There are 131 cases of marine animals with euryhaline animal species considered as marine animals, and amongst these marine taxa, 61 species are marine invertebrates when marine vertebrates are excluded, and there are only 7 species of marine arthropods (**Table 1**, **Figure 1**). With further literature investigation on

the published non-hexapod arthropod genomes, an additional horseshoe crab species was reported with fully sequenced genome (Nossa et al., 2014; Kenny et al., 2016). In cases of the crustacean species, there are 6 cases of sequenced species of Subclass Copepoda (Polechau et al., 2015; Madoui et al., 2017, Barreto et al., 2018; Jørgensen et al., 2019a; Jørgensen et al., 2019b; Kang et al., 2017), 3 additional cases of species of Subclass Branchipoda other than *Daphnia pulex* (Coulbourne et al. 2011; Baldwin-Brown et al., 2018; Savojardo et al., 2018; Lee et al., 2019). For crustacean class, Malacostraca, a terrestrial isopod species, *Armadillidium vulgare* (Chebbi et al., 2019) and two amphipod species, *Parhyale hawaiensis* (Kao et al., 2014) and *Hyaella azteca* (Poynton et al., 2018) were reported with fully sequenced genomes. In Order Decapoda, the most famous and economically important malacostracan taxa, 6 species of shrimps, *Penaeus japonicus* and *P. monodon* (Yuan et al., 2018), *P. vannamei* (Zhang et al., 2019), *Caridina multidentata* (Sasaki et al., 2017), *Neocaridina denticulata* (Kenny et al., 2014), *Exopalaemon carinicauda* (Yuan et al., 2017), a crayfish, *Procambrus virginalis* (Gutekunst et al., 2018), and two species of true crabs, *Eriocheir sinensis* (Song et al., 2016; Tang et al., 2020a), *Portunus trituberculatus* (Lv et al., 2017; Tang et al., 2020b) were reported with decoded genomes. These cases of less than 30 *de novo* researched genomes of non-hexapod arthropod indicate that non-hexapod arthropod species are far under-sampled compared to hexapod species, despite of their taxonomical diversity and importance.

Literature Reviews and General Introduction

It is widely known as the marine ecosystem covering more than 70% of the total surface of the Earth and occupying more than 97% of its water mass volume. The marine ecosystems extend from the intertidal zones to the abyssal zone reaching up to 6000m in terms of their depths, and from coastal regions to the open ocean in terms of their distances from the nearest landmass. Therefore, they support vast range of diverse marine species, with more than 190,000 documented species and more than 2 million species yet to be described (Mora et al., 2011). In terms of taxonomical diversity, 12 animal phyla are found in exclusively marine habitats amongst extent 35 phyla while there is no known animal phylum whose species inhabit only in terrestrial habitats (Boeuf, 2011). From the historical view of the life, the great part of major animal clades has been evolved at the marine ecosystems for more than 540 million years, since the Cambrian explosion which triggered the massive adaptive radiation of extent animal phyla and the evolution of their enormously diverse body plans (Marshall, 2006).

For exemplar, Class Insecta from Subphylum Hexapoda solely contributes more than 80% of the total number of described arthropod species with relatively limited variety of their body plans, which is contrasted to that species richness of marine arthropods are widely distributed along more than 6 marine arthropod Classes, such as Pycnogonida (Subphylum Chelicerata), Branchiopoda, Hexanauplia, Icthyostraca, and Malacostraca (Subphylum Crustacea) with their extremely differentiated body plans (Oakley et al., 2012). While insects share their common body plan (the head capsule, thorax consisted with three segments bearing legs, and abdomen), marine arthropods have vast diverse patterns of their body plans in terms of the pattern and number of segments and their appendages, and the degree of fusion of segments in each tagma (Deutsch and Mouchel-

Vielh, 2003; Grimaldi, 2009). Furthermore, Class Cephalocarida and Remipedia, archaic crustacean clades with only a few reported species, were reported as the most probable candidate of the sister taxa of Subphylum Hexapoda by recent phylogenomic researches (Reiger et al., 2010; Andrew, 2011; Reumont et al., 2012). Thus, these cases imply the fact that marine ecosystems serve as the reservoirs of archaic animal taxa which are crucial to understand the early evolutionary histories of modern crown animal groups and emphasize the necessity of *de novo* genome researches targeting on these marine animals.

As mentioned in the previous subsection, however, marine invertebrates are the least researched animal group in the field of *de novo* genome research. Moreover, the qualities of their assembly and annotation are usually much worse than those of reference model organismal genomes (Ellegren, 2013). I retrieved detailed statistics data for qualifying some published arthropod genomes from the "NCBI Genome List" webpage (https://www.ncbi.nlm.nih.gov/genome/browse#!/, Latest update at 2020.05.03. Retrieved at 2020.05.03.) for comparing the qualities of genomic assemblages of model arthropods and non-model, marine arthropods (**Table 2**). In Subphylum Hexapoda, the statistic values of two thoroughly studied model organisms, *Drosophila melanogaster* and *Anopheles gambiae* (Class Insecta, Order Diptera) were retrieved with those of a non-model insect, Folsomia candida (Class Entognatha, Order Isotomidae). The previously referred 2 amphipods (Class Malacostraca, Order Amphipoda), 4 branchiopods (Class Branchiopoda), 6 copepods (Class Hexanauplia, Subclass Copepoda), and an isopod (Class Malacostraca, Order Isopoda) were also investigated. Finally, to the best current knowledge, the genomic statistics data of 10 decapods species was also investigated.

Table 2. Statistics of some published arthropod genomes with the parameters assessing quality of their genomes, retrieved and modified from the "NCBI Genome List" (Latest update at 2020.05.03. Retrieved at 2020.05.03.).

Taxa	Scientific name	Model-organism	Genome size (Mb)	Ambiguous bases (%)	Contig No. (ea)	Scaffold No. (ea)	Contig N50 (bases)	Scaffold N50 (bases)	Accessible genes (ea)	Public data availability
Insecta	*Drosophila melanogaster*	Yes	143.726	0.802	2,442	1,870	21,485,538	25,286,936	17,874	NCBI, Refseq
	Anopheles gambiae	Yes	265.027	4.744	16,825	8,145	85,548	12,309,988	14,102	NCBI, Refseq
Entognatha	*Folsomia candida*	No	221.703	0.113	228	162	4,885,648	6,519,406	22,100	NCBI, Refseq
Amphipoda (Malacostraca)	*Hyalla azteca*	No	550.886	0.476	23,426	18,000	114,415	215,427	20,022	NCBI, Refseq
	Parhyale hawaiensis	Yes	2752.561	20.289	610,812	278,189	10,438	20,228,728	N/A	NCBI
Branchiopoda	*Daphnia pulex*	Yes	189.551	2.113	2,150	493	194,489	1,160,003	N/A	NCBI
	Daphina magna	Yes	122.953	6.746	16,818	14,486	4,193	10,124,675	21,539	NCBI, Refseq
	Lepidurus apus	No	87.970	1.066	20,545	7,908	15,890	42,769	N/A	NCBI
	Lepidurus arcticus	No	73.106	0.462	6,908	3,152	82,939	118,958	N/A	NCBI
Copepoda (Hexanauplia)	*Acartia tonsa*	No	989.163	0.311	383,038	351,850	3,244	3,610	N/A	NCBI
	Apocyclops royi	No	262.264	1.814	143,521	97,072	2,108	3,257	N/A	NCBI
	Eurytemora affinis	No	389.033	0.649	14,526	6,171	67,724	252,275	23,789	NCBI, Refseq
	Oithona nana	No	85.010	3.463	7,437	4,626	38,620	400,614	N/A	NCBI
	Tigriopus californicus	No	191.143	2.084	11,341	459	44,438	15,806,032	15,577	NCBI, Refseq
	Tigriopus kingsejongensis	No	338.547	0.161	1,097	938	1,293,995	1,473,880	N/A	NCBI

Table 2. Continued from the previous page

	Caridina multidentata	No	1948.953	0.002	2,751,313	2,750,712	819	819	N/A	NCBI
	Eriochier sinensis	No	1118.180	0.213	111,755	17,553	2,066	22,400	N/A	GigaScience
	Exopalaemon carinicauda	No	6699.724	1.441	11,890,323	9,470,451	696	962	N/A	NCBI
	Neocaridina denticulata	No	1284.468	N/A	3,346,358	N/A	400	N/A	N/A	NCBI (SRA archives only)
	Penaeus japonicus	No	1660.270	2.132	2,891,064	2,434,740	700	912	N/A	NCBI
Decapoda (Malacostraca)	*Penaeus monodon*	No	1447.416	5.542	3,492,929	2,525,346	431	769	N/A	NCBI
	Penaeus vannamei	No	1663.581	2.737	33,020	4,683	86,864	605,555	30,733	NCBI, Refseq
	Portunus trituberculatus, (Lv et al., 2017)	No	842.129	4.300	1,268,724	898,300	756	1,154	N/A	NCBI (SRA archives only)
	Portunus trituberculatus, (Tang et al., 2020b)	No	1005.046	0.096	2,446	523	4,121,416	21,793,880	N/A	GigaScience
	Procambrus virginalis	No	3290.471	50.507	2,332,443	1,187	1,942,826	39,275	N/A	NCBI
Isopoda (Malacostraca)	*Armadillidium vulgare*	No	1725.108	0.427	52,740	43,541	38,359	51,088	19,051	NCBI, Refseq

To validate the contiguity of *de novo* assembled genomic sequences, 3 criteria of contiguity, completeness, and correctness are famously used as I mentioned in the previous subsection. There are two mostly used concepts to assess the contiguity of genomic assemblage, one is the N50, and the other is the number of genomic contigs or scaffolds (Meader et al., 2010). The N50 value is defined as the length of the smallest genomic contigs or scaffolds when its length summed with those of entire bodies of smaller contigs or scaffolds firstly reaches at least 50% of the total length of assemblage (Miller et al., 2010). With the consideration of the total length of assembly and the contig or scaffold number, N50 provide intuitive understanding on the quality of genomic contiguity. The completeness of the assembly can be verified from various features, such as the ratio between the finally assembled and initially estimated genome size, the ratio of ambiguous bases resulted from the scaffolding process (Pop et al., 2004), and the relative number of core orthologous genes predicted from the genomics sequences (Parra et al., 2007; Simão et al., 2015). On the other hand, the correctness of genomic assembly is relatively hard to be measured (Miller et al., 2010; Earl et al., 2011). The relative content of ambiguous bases can be used to infer the correctness of assembly indirectly, since the lower the computational threshold for connecting contigs into a scaffold becomes, the more miss-assemblies (such as collapsing repetitive regions or introducing false translocatons between distantly located contigs) happen with increased amount of unambiguous bases introduced (Meader et al., 2010).

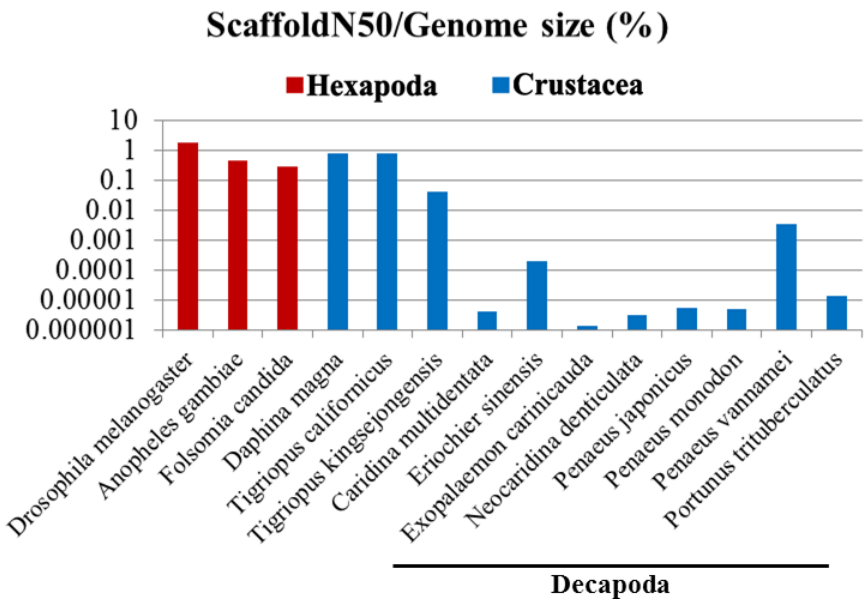

Figure 2. The assembly quality of sequenced genomes of some hexapod and copepod, decapod crustaceans described in **Table 2**.

There are five model-organisms in **Table 2**, *D. melanogaster* and *A. gambiae* belonging to the Hexapoda and *Daphnia magna*, *D. pulex*, and *Parhyale hawaiensis* belonging to the Crustacea. The genomes of these well-studied arthropods show high genomic contiguity, which are indicated with the number and N50 values of their genomic contigs and scaffolds. Moreover, the percentages of ambiguous bases are recorded generally low considering their contiguity parameters, such as more than 1Mb (1 million bases) long N50 values of their genomic scaffolds. In contrast, the majority of crustacean genomes, except two *Daphnia* species, demonstrate much lower genomic contiguity than those of model arthropods (**Table 2, Figure 2**). Except for two copepod species of Genus *Tigriopus*, their scaffold N50 value are less than 1 Mb, and as the other extreme cases, 2 copepods (*Acartia tonsa*, *Apocyclops royi*), 4 decapods (*Caridina multidentata*, *Exopalaemon carinicauda*, *Neocaridina denticulata*, *Penaeus japonicus*, and *P. monodon*)

genomes show scaffold N50 value even shorter than 1,000 bases indicating their substantially low genomic contiguities. Moreover, non-model arthropod genomes bigger than 1Gb (1 billion bases) demonstrate ambiguous bases ratios substantially high (up to 50.507% in *Procambrus virginalis*) considering their genomic contiguities.

In terms of the quality of genomic annotation, differences between non-model and model arthropod genomes are more drastic. Amongst 5 model arthropods with highly contiguous genomes, *Daphnia pulex* and *Parhyale hawaiensis* are only cases without reviewed genomic annotation published in the NCBI Refseq database (**Table 2**). In contrast, irrelevant to their published articles, 16 species out of 19 total non-model arthropods do not contain publically accessible genomic annotations in the NCBI Refseq, except for *Folsomia candida*, *Hyalla azteca*, and *Armadillidium vulgare*. Lastly, 3 decapod species (*Neocaridina denticulata*, *Eriocheir sinensis*, and *Portunus trituberculatus* reported from Lv et al., 2017) are found to be currently inaccessible from NCBI Genome List webpage, leaving their accession numbers of Bioproject and Biosample only. Therefore, these cases imply insufficient quality assessment against their genomic assemblies and annotations.

The non-hexapod marine arthropods are the least researched animal groups despite of their great necessities for *de novo* genome research in order to understand the early evolutionary history of this phylum. Moreover, publically accessible genomic annotations of these marine arthropods are much more limited in number compared to assembled genomic sequences of the very same species, which is another great obstacle for conducting comparative genomic analyses to understand their evolution.

This study therefore conducted *de novo* genome researches on 3 species of marine arthropods (*Chionoecetes opilio*, *Nymphon striatum*, and *Portunus trituberculatus*) which

belong to the undersampled arthropod taxa, Class Pycnogonida and Infraorder Brachyura. A marine ray-finned fish (*Liparis tanakae*) genome was also researched in this study that provided a unique insight on its genomic characteristics and the quality control criteria to verify the quality of assembled 3 arhtropod genomes. In addition, using these *de novo* assembled arthropod genomes, a preliminary case of evolutionary genomic study was conducted in a laboratory without high level computering resources. The comprehensive approaches of this study aim to provide unique insights on the Chelicerate phylogeny with publically available data of the assembled *de novo* genomes deposited to the NCBI.

The following contents of each chapter are summarized here:

1) Chapter 1 describes the pilot researches for establishing *de novo* genome research workflows with a marine ray-finned fish genome assembly (*L. tanakae*) and basic phylogenetic analyses with full mitochondrial genome datasets (*C. opilio*).

2) Chapter 2 demonstrates three arthropod *de novo* genome assemblies (*C.* opilio, *N. striatum*, and *P. trituberculatus*) and their quality improvement procedures. This chapter also provides discussion on the optimization methods of *de novo* genome researches for non-model marine arthropods.

3) Chapter 3 provides a preliminary case of comparative genomic study which applys the *de novo* genomes of *C. opilio, N. striatum* and *P. trituberculatus* with 16 selected species representing major arthropod taxa.

Chapter 1. THE PILOT RESEARCHES FOR EVOLUTIONARY STUDIES ON MARINE ARTHROPOD GENOMES

1.1. The preliminary genomic studies on *Liparis tanakae* and its genomic characteristics

1.1.1. Introduction

It is known that fishes occupy more than half of the all known vertebrate species (Koepfli et al., 2015), and within them, they also show great diversity in body plans ranging from those of jawless fishes (Superclass Cyclostomata), to those of the most specious group, ray-finned fishes (Class Actinopterygii). Thanks to their diversities and often unusually small, a few hundred Mb sized genomes, numerous model and non-model actinopterygiian fishes have been sequenced since the early era of genomics (Brenner et al., 1993; Aparicio et al., 2002; Jaillon et al., 2004). In addition to their convenience of obtaining relatively high-qualitied genomes, these actinopterygiian genomes also enabled one of the first true comparisons between large, interspecies genomic structures which revealed the series of lineage-specific whole-genome duplication events in early vertebrate history (Christoffeles et al., 2004; Jallion et al., 2004).

The Family Liparidae is one of the most specious actinopterygiian families (Chyung, 1977; Knudsen et al., 2007) including 29 genera and about 345 species in the world (Chernova et al., 2004; Chernova et al., 2005; Stein, 2006). Liparid fishes are known with peculiar morphological characteristics, such as thin and loose gelatinous skin without a scale. From the view from marine arthropod genomics field, their mucous rich tissues can be applied as models for extracting high molecular-weighted genomic DNA suitable for *de novo* genome sequencing, which is one of the greatest obstacles in many marine animals, such as mollusks and crustacenas with slimy tissues (Bitencourt et al., 2007; Panova et al., 2016; Schultzhaus et al., 2019).

Liparis tanakae is a common snailfish species in the coastal water of Korea, China and Japan (Tomiyama et al., 2013; Chen et al., 1997; Rhodes, 1998; Jin et al., 2003). It has also been reported economically important species as one of a major predator of both wild and hatchery-released juveniles of a famous edible fish, Japanese flounder (*Paralichthys olivaceus*) in East Asian countries (Tomiyama et al., 2009). In addition, it is also commercially caught as an edible fish in some localities of Korea and Japan, and used as the main ingredient of a local winter season tonic soup in Korea (Ustadi et al., 2005). As an exemplar monitoring case of its population, a Korean Institute has started to release artificially fertilized and raised juvenile *L. tanakae* to promote the protection of its population since 2013, which resulted the annual amount of released juveniles increased from 2 million to 79 million in 5 years (Korea Fisheries Resources Agency, unpublished data, 2019).

In addition to its economic importance, *L. tanakae* shows typical morphological characteristics as in other liparid fishes. There are some preliminary genetic and proteomic researches focusing on biochemical natures of its tissues. A study suggested five novel candidate genes rich in its skin and muscle tissues of high glycoprotein contents which might contribute to evolution of the liparid-specific morphological characteristics by histochemical analyses (Song et al., 2000). The following research which had conducted an interactive *in situ* hybridization and immunohistochemistry reported specific expression patterns of these five novel candidates in *L. tanakae* tissues and one specific clone with unique expression patterns shared with *L. tanakae* tissues and human salivary tissues (Song et al., 2002). In addition, these researches provided a hypothetical 3D-protein structures of these novel candidates from tissues of *L. tanakae* which shared similar predicted function and structure with those of human aPRPs (acidic

proline-rich-proteins) rich in human salivary glands (Song et al., 2002). Their interactive genomic study, however, does not exist until the year 2019, despite of peculiar morphological traits of liparid fishes.

The aim of this study is to assemble and annotate *de novo* sequenced genome of *Liparis tanakae* for the first time, and provide a genomic resource verified, deposited to the NCBI data reservoir that is available open to public. The annotated *de novo* genome assembly of *L. tanakae* was used to discuss the evolution of its liparid specific morphologies by comparing the collagen family of structural genes with four model vertebrate genomes. In addition, the methodologies used in this study are used as a pilot research to establish workflows of marine arthropod *de novo* genome researches in Chapter 2 and 3. Futhermore, the statistics of *de novo* assembled *L. tanakae* genome was used as a verified control group to assess the assembly quality of three arthropod genomes researched in Chapter 2. Here, I report the first *de novo* draft genome of *L. tanakae* which was researched since the year 2016.

1.1.2. Materials and Methods

Sample collection and Whole-genome sequencing

A juvenile female *L. tanakae* with its body length 21.01cm and mass 50.32g, was collected at around 400 meters deep from the East Sea of South Korea (38.76ºN, 130.85ºE) (**Figure 3**). In order to prevent the degradation of genomic and transcriptomic nucleic acids, the sample was immediately placed in the liquid nitrogen and brought to the laboratory. The tissue preparation and lysis procedures were conducted according to described protocols suitable for various types of animal tissues (Zhang et al., 2013). Its muscular tissue (approximately $1cm^3$) was isolated from the frozen individual and then homogenized by grinding with liquid nitrogen immersion. The resulted tissue powders were then followed by a manual phenol/chloroform DNA extraction. The transcriptomic RNA was extracted from these powdered tissues using TRIzol® RNA Reagent (Thermo Fisher Scientific, MA, USA) following the manufacturer's instruction. In order to obtain sufficient amount and quality for *de novo* genome sequencing, both of the extracted genomic DNA and transcriptomic RNA were verified using a NanoDrop 1000 spectrometer (Thermo Fisher Scientific, MA, USA) and a 2100 Bioanalyzer (Agilent Technologies, CA, USA). The validated DNA and RNA extracts from the *L. tanakae* specimen were about approximately 5μg, respectively. Finally, the specimen information was deposited at the NCBI with following accession numbers (PRJNA523297, SAMN10970109).

TruSeq DNA Nano DNA Library Preparation Kit and Nextera Mate Pair Library Preparation Kit V2 (Illumina, CA, USA) were used to construct the genomic DNA libraries for Illumina paired-end (PE) sequencing. To generate the transcriptomic cDNA libraires for RNA sequencing, TruSeq RNA library preparation kit v2 (Illumina, CA, USA) was applied. To construct 6 different insert-sized genomic libraries and a

transcriptomic library (**Table 3**), the nucleic acid extracts were sheared with Covaris instrument (Covaris®, MA, USA) with 200 cycles of running at 6°C to obtain optimized insert-sizes of each library in **Table 3**. These sheared extracts then underwent modifications of end repair (paired-end), circularization (mate pair), adapter ligation and enrichment, following the protocols in each respective manufacturer's instruction. The resulted sequencing libraries were validated by 2100 Bioanalyzer before the *de novo* genome sequencing. Finally, HiSeq 4000 insrument (Illumina, CA, USA) was applied to sequence these libraries using HiSeq 4000 SBS Kit.

Figure 3. The juvenile female Liparis tanakae used in this study

Table 3. The statistics of libraries and *de novo* sequenced reads of *Liparis tanakae* genome and transcriptome, after the quality control.

Library type	Insert-size (bp)	Read length (bp)	Total reads bases (bp)	No. of reads	GC (%)	Reads Q20 (%)	Reads Q30(%)
DNA, paired-end	350	151	55,246,020,081	423,247,172	42.16	97.87	92.36
DNA, paired-end	550	151	55,847,817,195	444,495,594	41.90	97.56	91.84
DNA, mate pair	3,000	151	9,634,828,376	72,884,446	43.49	92.72	83.61
DNA, mate pair	5,000	151	8,156,715,114	60,728,058	42.96	92.09	82.33
DNA, mate pair	8,000	151	40,076,007,643	304,166,366	43.14	92.99	84.25
DNA, mate pair	10,000	151	60,424,370,494	400,161,394	41.94	95.91	89.91
RNA, paired-end	350	101	18,144,387,277	181,045,356	52.41	99.25	97.34

De novo genome estimation and assembly

The raw genomic and transcriptomic *de novo* sequenced reads were verified using Q30 quality score (error rate of sequenced reads less than 0.1%) by FastQC v0.10.0 software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The ligated adapter sequences were trimmed using Trimmomatic (Bolger et al., 2014). The genome survey of *L. tanakae* was conducted using Jellyfish v1.1.11 software which assembled 350 bp and 550 bp insert-sized paired-end sequenced reads with 3 diff1ernt K-mer sizes (7, 21, 25bp) to estimate its genomic size (Marçais & Kingsford., 2011). According to the predicted 600Mb sized genome, *de novo* assembly was performed using SOAPdenovo2 (v2.04) (Luo et al., 2012) and Platanus v1.2.4 (Kajitani et al., 2014) with K-mer size parameter as variable and other parameters as default states, to obtain assembled genomic contigs from paired-end sequenced reads. Platanus v1.2.4 was applied for the scaffolding and gap-closing using mate pair sequenced reads with long insert-sizes, to generate consensus, scaffolded draft genomic sequences of *L. tanakae*. These scaffolded and gap-closed genomic sequences were trimmed out of short fragments whose length were less than 1,000bases, for increasing the quality of assemblage and genomic annotation. Finally, the draft genome was validated by searching core orthologous genes shared in actinopterygiians using BUSCO 2 (Simão et al., 2015) with respective databases (actinopterygii_odb9).

Transcriptomic analyses and genomic annotation

The filtered *de novo* sequenced transcriptomic reads were assembled into contigs using Trinity r20140717 software (Grabherr et al., 2011). These assembled contigs which representing hypothetical transcripts were then clustered to remove excessive redundant contigs using CD-HIT-EST v4.6 software (Li and Godzik, 2006) with default parameters. The TransDecoder v 3.0.1 (https://github.com/TransDecoder/TransDecoder/) was used to predict open reading frames (ORFs) from these clustered contigs with default parameters and minimal length threshold of 100 amino acids. The relative abundance of each predicted ORFs were calculated from RSEM algorithm (Li and Dewey, 2011) which is incorporated in the Bowtie v.1.1.2 softare (http://deweylab.github.io/RSEM/). In addition, hypothetical functions of these ORFs were also predicted by NCBI BLASTX local application (Cameron et al., 2004) and DIAMOND program (Buchfink et al., 2015) with default e-value threshold of 1.0E-5. The orthologous gene databases were used for the functional annotation of these hypothetical transcripts as following: Kyto Encyclopedia of Genes and Genomes (KEGG), NCBI nucleotide and non-redundant databases, Pfam, Gene ontology (GO), Uniprot and EggNOG.

The genomic annotation was conducted by combining *ab initio* prediction and two different types of intrinsic and homology-based, extrinsic evidences. To obtain intrinsic evidence of transcriptional sites (start, end, and splicing), Tophat v2.0.13 software (Trapnell et al., 2009) was used to perform the transcriptomic reads mapping against the genomic sequences. High-qualitied actinopterygiian proteins were obtained from NCBI Refseq Gene databases with following filter parameters “((actinopterygii[Organism]) AND "source genomic"[Properties]) AND "srcdb refseq reviewed"[Properties]”. These downloaded proteins were clustered to reduce redundancy by using CD-HIT PROTEIN

(Li and Godzik, 2006) and subjected as the verified extrinsic evidences. Finally, an automated *de novo* genome annotation pipeline, Seqping v0.1.33 (Chan et al., 2017) performed the genome annotation which incorporates *ab initio* repetitive element predictor of RepeatMasker (Tarailo-Graovac and Chen, 2009) and gene structure predicting softwares. MAKER2 v2.28 (Holt and Yandell, 2011) was used to perform initial gene prediction with collected intrinsic and extrinsic evidences to further train other 3 gene model prediction tools with default parameters. Then, GlimmerHMM v3.0.4 (Majoros et al., 2004), AUGUSTUS v3.2.2 (Stanke et al., 2006), and SNAP (2012/05/17) conducted independent *ab initio* gene model prediction according to the MAKER-resulted training parameters. MAKER2 was once more applied to merge these predicted gene models into the consensus gene sets with genome annotation information. Finally, the same orthologous gene databases used for the functional annotation of predicted transcripts were applied again to obtain functionally annotated final gene sets of *L. tanakae* with default e-value threshold 1E-05.

Basic comparative genomic analysis with model vertebrates

The reference proteomes of four thoroughly studied model vertebrates (*Danio rerio, Homo sapiens, Larimichthys crocea, Mus musculus*) were downloaded from the ftp service of the NCBI Refseq (ftp://ftp.ncbi.nlm.nih.gov/genomes/). The information of these reference vertebrate genomes are described in **Table 4**. The BLASTP all-to-all search (Delaney et al., 2000) was performed to find protein hits with high similarities of their amino acid sequences. Then OrthoMCL v2.0.9 (Fischer et al., 2011) was used to predict and cluster orthologous proteins from these 5 vertebrate proteomes. The resulted singletons and clustered orthologous proteins were visualized as a Venn diagram with OrthoMCL. In addition, the structural proteins belonging to the collagen gene families of 5 vertebrates are manually inspected based on these orthologue analysis results.

Table 4. The summary of downloaded 4 reference vertebrate genomes in this study.

Species	Assembly ID	RefSeq accession	No. of genes	Data sources
Danio rerio	GRCz11	GCF_000002035.6	39,988	RefSeq reference genomes
Homo sapiens	GRCh38.p12	GCF_000001405.38	59,026	RefSeq reference genomes
Larimichthys crocea	L_crocea_2.0	GCF_000972845.2	27,368	RefSeq reference genomes
Mus musculus	GRCm38.p6	GCF_000001635.26	50,865	RefSeq reference genomes

To analyze the phylogenetic relationships between these vertebrates, non-redundant orthologous genes that shared among all species of interest were collected from the orthologous gene results of BUSCO (with vertebrata_odb9 database) and OrthoMCL analyses. The amino acid sequences of these genes were aligned using MAFFT (Katoh et al., 2017) and poorly aligned regions were trimmed with trimAl (Capella-Gutiérrez et al., 2009). Each alignment of orthologous genes was fused to creat a supermatrix for

phylogenetic reconstruction using the precompiled bioinformatics script, BeforePhylo (https://github.com/qiyunzhu/BeforePhylo). Finally, RAxML 8.2.12 HPC (Stamatakis, 2014) performed phylogenetic reconstruction using maximal likelihood method.

1.1.3. Results

The genome size of *L. tanakae* was estimated to be approximately 598Mb by briefly assembling 111.1Gb paired-end genomic sequenced reads (**Figure 4**). The final assembly of *L. tanakae* was 499.08Mb sized genome which is composed of 27,879 scaffolds with N50 value of 375.22Kb with only ambiguous base contents of 6.11% (**Table 5A**). In addition, the total coverage depth of *de novo* sequenced genomic reads in this study was measured as more than 382 fold. From the *L. tanakae* draft genome, 3,837 genes (89.3%) amongst 4,584 actinopterygiian core orthologues were found with their sequences intact as the result of BUSCO assessment (**Figure 5**). Additionally, 381 genes (8.31%) were recovered with partial sequences, which leaves only 366 genes (7.98%) unrecovered from the *L. tanakae* genome.

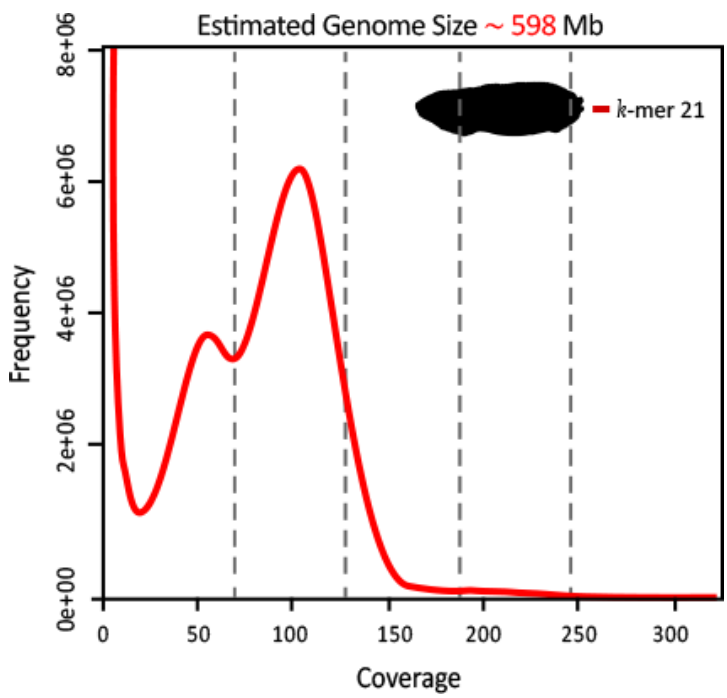

Figure 4. The estimated genome size of *L. tanakae*

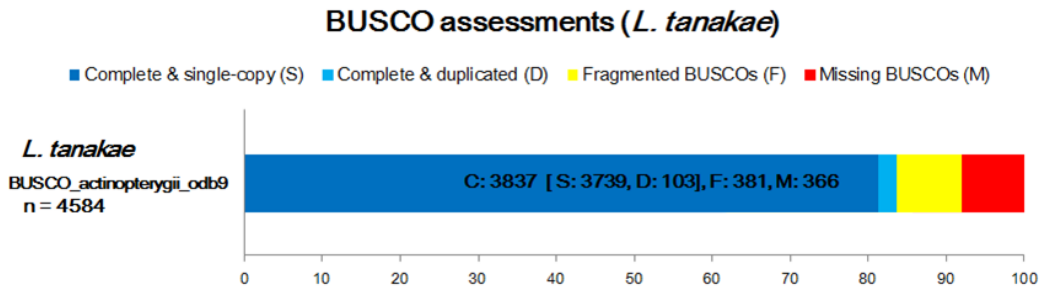

Figure 5. The genome assessment result using BUSCO

Table 5. The statistics of assembly and annotation of *L. tanakae*

A. Summary of statistics of the genome assembly	
Total bases (Mb)	499.08
No. of scaffolds	27,879
Average length (bp)	17,091
Maximum length (bp)	3,437,558
N50 (kb)	375.22
N's (%)	6.11
GC ratio (%)	42.20
B. Summary of statistics of the annotation	
Predicted gene models	68,356
Protein coding genes	28,882
Average transcript length (bp)	672
Average intron length (bp)	1,777
Average exons/gene	4.29
Average introns/gene	3.29
No. of tRNA	20
No. of rRNA	78

There were in total 68,356 predicted genes with their averge length of 2,449bp which were resulted from the Seqping gene annotation pipeline (**Table 5B**). After the quality curation by minimal length threshold of 100 amino acids and functional annotation process, 11,093 protein coding genes, 20 transfer RNA (tRNA) genes, and 78 ribosomal RNA (rRNA) genes were obtained (**Table 5B**). For functional categorization of these genes, 46.55% were recorded as no-hit against EggNOG database, which was followed by intracellular-or-extracelluar transportation function (11.75%), posttranslational

modification related function (8.03%), signal transduction pathways (7.05%), and general transcription (6.09%) as described in **Figure 6** and **Table 6**. The categorization using Gene Ontology database (GO) results showed that the majority of the of predicted *L. tanakae* genes clustered into the intracellular and intercellular processes, variety of metabolic pathways, and enzymatic functions (**Figure 7**).

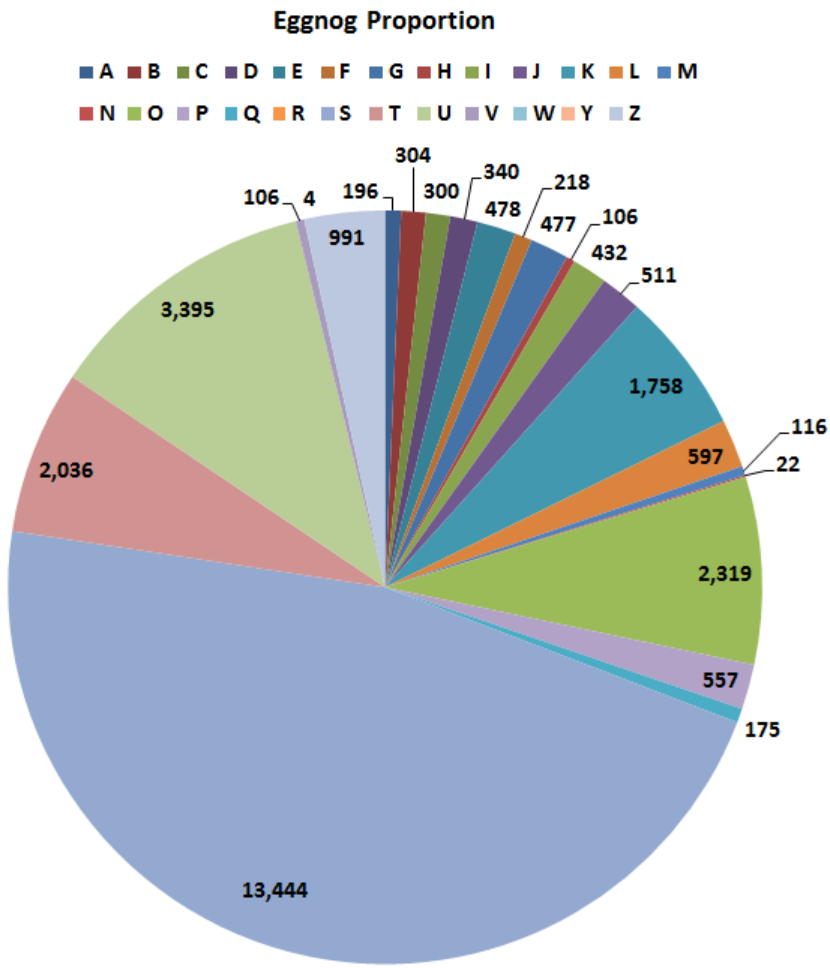

Figure 6. The categorization of predicted functions of *L. tanakae* genes by EggNOG

Table 6. The functional categories of predicted functions of *L. tanakae* genes in Figure 6

EggNOG category	Description	Count (ea)	Ratio (%)
A	RNA processing and modification	196	0.679
B	Chromatin structure and dynamics	304	1.053
C	Energy production and conversion	300	1.039
D	Cell devision and cycle factors	340	1.177
E	Amino acid transport and metabolism	478	1.655
F	Nucleotide transport and metabolism	218	0.755
G	Carbohydrate transport and metabolism	477	1.652
H	Coenzyme transport and metabolism	106	0.367
I	Lipid transport and metabolism	432	1.496
J	Translation and ribosmal biogenesis	511	1.769
K	Transcription	1,758	6.087
L	DNA replication, recombination, and repair	597	2.067
M	Cellular envelope biogenesis	116	0.402
N	Cell motillity	22	0.076
O	Posttranslational modifications	2,319	8.029
P	Inorganic transport and metabolism	557	1.929
Q	Secondary metabolites metabolism	175	0.606
R	General prediction only	0	0.000
S	Function unknown	13,444	46.548
T	Signal transduction pathways	2,036	7.049
U	Intracellular transportation	3,395	11.755
V	Defense mechanisms	106	0.367
W	Extracellular structures	4	0.014
Y	Nuclear structure	0	0.000
Z	Cytoskeleton	991	3.431
Total		**28,882**	**100.000**

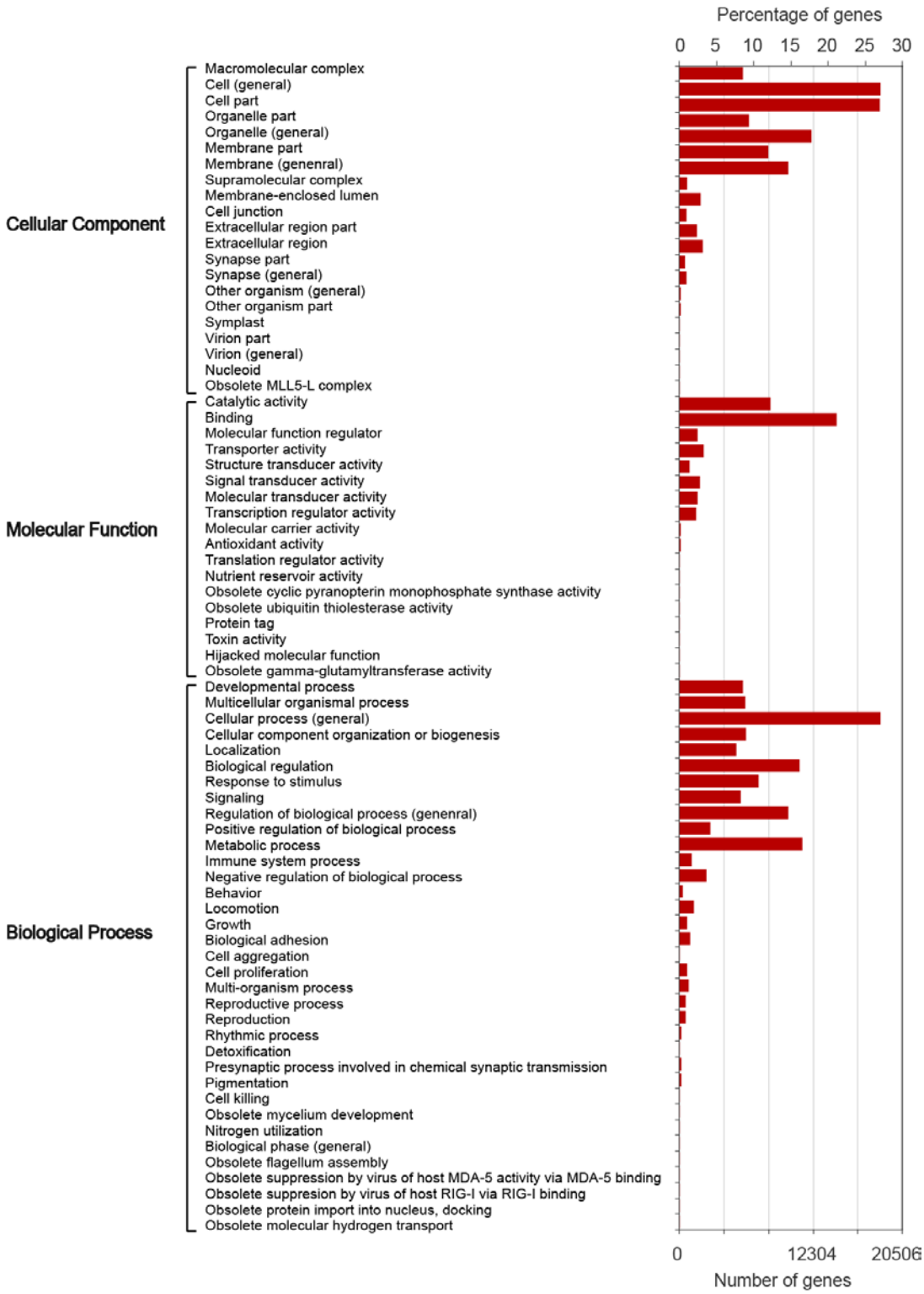

Figure 7. The categorization of predicted functions of *L. tanakae* genes against GO database

Among the proteomes of *L. tanakae* and other 4 vertebrates, in total 8,784 familes of shared orthologue were detected, and 884 genes were identified as singletons which were uniquely present only in *L. tanakae* genome (**Figure 8**). There were 785 non-redundant orthologues shared in these species after excluding the orthologous clusters containing at least one paralogous gene. In addition, BUSCO analysis on proteomes of these vertebrates with vertebrata_odb9 database found 209 non-redundant orthologues which were present in the database. The maximum likelihood phylogenetic reconstructions using the alignments of these two sets of orthologues with protein substitution parameter of "-m PROTGAMMAAUTO" were consistent with each other, and accorded to the well-known consensus relationship of these vertebrates (**Figure 9**).

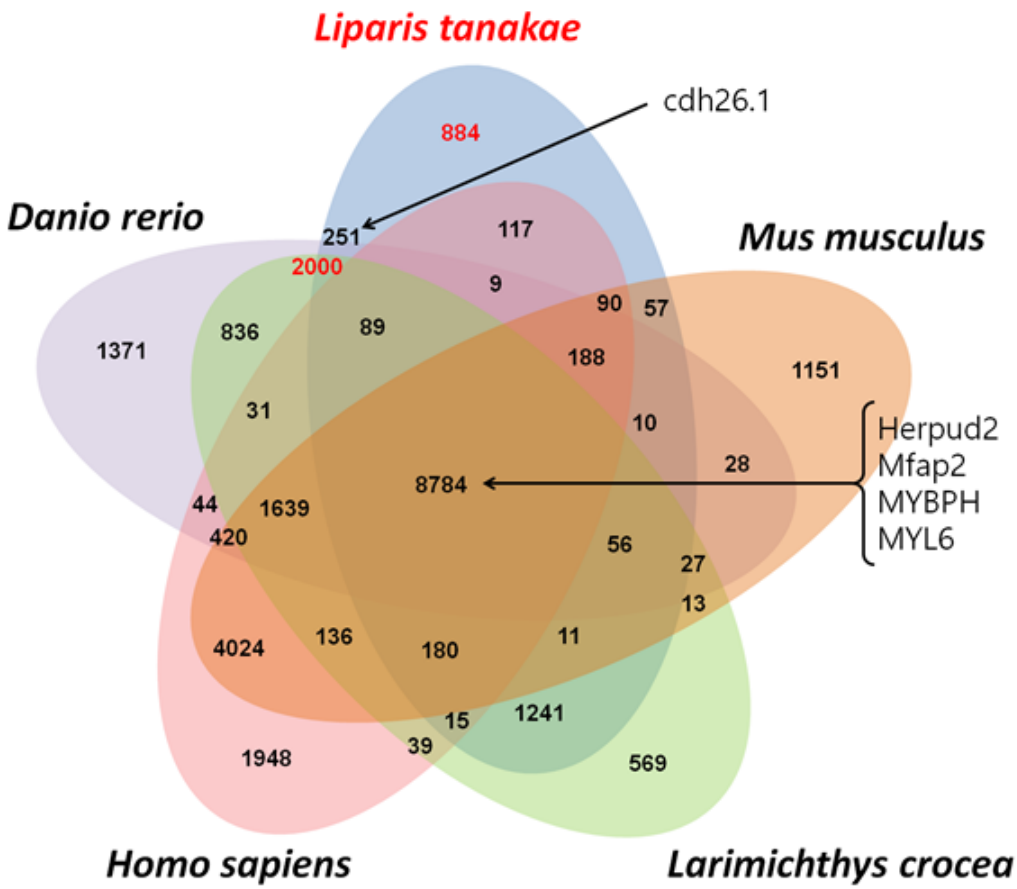

Figure 8. The Venn diagram of orthologous genes shared between five vertebrates.

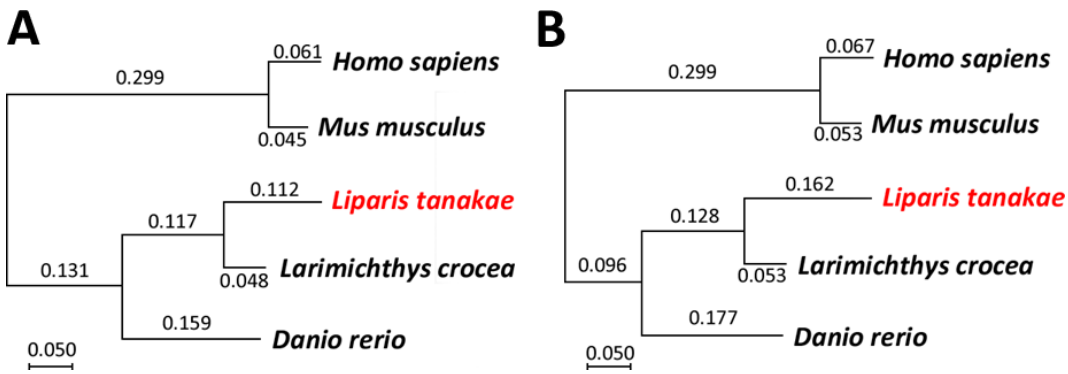

Figure 9. The unrooted phylogenetic trees reconstructed with maximum likelihood method and automatic estimation of protein substitution matrix of RAxML software. (**A).** The tree reconstructed using the aligned non-redundant OrthoMCL orthologues, (**B).** The tree reconstructed using the aligned non-redundant BUSCO orthologues.

1.1.4. Discussion

For the further validation of quality of assembly, the quality indicating parameters of the number of scaffold and its N50, the content of ambiguous bases, and ratio of complete BUSCO genes indicated that *L. tanakae* draft genome was nearly completed with little amounts of errors (**Figure 5**, **Table 5A**). When these values of quality indicating parameters were compared to those of initial versions of vertebrate reference genomes assembled only with short Illumina reads, *L. tanakae* genome was validated further with substantial support (Earl et al., 2011; Bradnam et al., 2013). For instance, there were 3 reference vertebrate genomes whose ratios of vertebrate-core orthologues were found to be lower than 80%, and whose scaffold N50 lower than 100kb (Bradnam et al., 2013). In addition, the longest scaffold in *L. tanakae* genomes was more than 3.43Mb long, which further verified the contiguity of the assembly.

Further investigation on structural genes belonging to the collagen families from each proteome of 5 studied vertebrates was conducted to find a potential trace of evolution of liparid specific morphological characteristics. The comparison of the numbers of

redundant collagen orthologues and present collagen families per each vertebrate species implied that *L. tanakae* is the most collagen gene-rich species, with 35 collagen genes from 26 families identified (**Figure 10**). On the other hand, there were 27 genes from 16 famiels in the *M. musculus*, 27 genes from 18 families in the *H. sapiens* which are exclusively terrestrial vertebrates in Class Mammalia. From *D. rerio* and *L. crocea* genomes, 29 collagen genes belonging 17 families and only 18 collagen genes belonging 13 families were found. The abundance of collagen genes in *L. tanakae* genomes was found to be well consistence with the previous interactive biochemical researches (Song et al., 2000; Song et al., 2002), with the intact matches of sequences of all 5 novel candidates reported by them (**Figure 8**). These novel candidate clones were matched within 884 *L. tanakae* specific signletones which further supported the novelty of these clones reported from these researches (Song et al., 2000; Song et al., 2002). When their hypothetical function and 3D structure predictions for these clones are considered, it is suggested that liparid genomes has experienced the expansion of collagen genes and aPRPs (acidic proline-rich-proteins) both of which can contribute to immune responses. Therefore, this study demonstrates the genomic context-understanding of the evolution of mucous rich tissues of liparid fishes, possibly related with the adaptation to increase immune activities. Neverthelss, it is required to conduct interactive further studies with the forward and reverse genetics to validate the hypothesis of this study thoroughly.

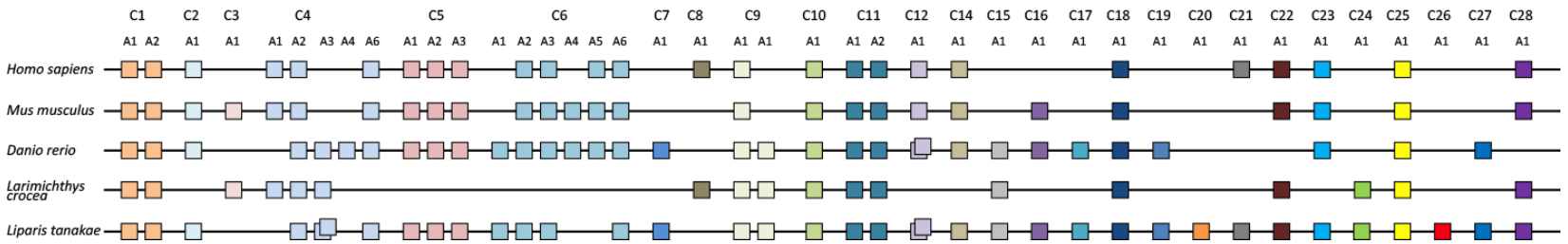

Figure 10. The comparison of the numbers of families and copies of annotated collagen genes found from the *L. tanakae* and 4 vertebrate reference proteomes

1.2. The *de novo* mitochondrial genome of *Chionoecetes opilio* : The manual curation of predicted genes and the phylogenomic analyses with large datasets

1.2.1. Introduction

The mitochondrial is an essential organelle performing the oxidized cellular respiration which exists in almost every eukaryote species. Even before a decade of the beginning of whole-genome era, mitochondrial genomes of various animal species, such as human, cow, *Xenopus laevis*, and a honeybee, *Apis mellifera*, had been sequenced (Anderson et al., 1981; Anderson et al., 1982; Roe et al., 1985). In subphylum Arthropoda, it was reported that an ordinary arthropod mitogenome (mitochondrial genome) is a closed circular molecule containing 15 to 20Kb nucleotides with 13 protein coding genes, 22 tRNA genes, and 2 rRNA genes (Pisani et al., 2013).

Snow crabs are famous food crab species belonging to the Genus *Chionoecetes* (Infraorder Brachyura: Superfamily Majoidea: Family Oregoniidae) which inhabit the cold, arboreal waters of the Northern Pacific and the Northwestern Atlantic regions (Alvsvåg et al., 2009; Ng et al., 2009). *Chionoecetes opilio* is the most important commercial species among the congeneric species due to its largest annual catches (FAO Fisheries and Aquaculture Department, 2019). Currently, there are about 100 sequenced mitogenomes for variety of brachyuran species according to the “NCBI Genome List” webpage (https://www.ncbi.nlm.nih.gov/genome/browse#!/, Latest update at 2020.06.03. Retrieved at 2020.06.03.). In Superclass Majoidea, at least three species, *Damithrax spinosissimus* (Márquez et al., 2014), *Maja crispata* and *M. squinado* (Basso et al., 2017) were reported with their completed mitochondrial genomes. When it is focused into

Family Oregoniidae, the mitogenome of *Chionoecetes japonicus*, a Japanese snow crab, was already sequenced and deposited at the NCBI (Accession number AB735678, data published in NCBI at 2013).

Therefore this study aims to provide *de novo* assembled complete mitogenome of a snow crab, *Chionoecetes opilio*. The predicted genes from the mitogenome assembly pipeline were further manually curated in this study which emphasized the importance of the curation process for accurate comparative genomic analyses, which was more thoroughly conducted at the Chapter 3. In addition, a phylogenomic analyses using 13 mitochondrial protein coding genes (PCGs) were conducted as the piliot studies with whole-genome scaled comparative analyses which in conducted at the Chapter 3. The *de novo* sequenced Illumina genomic reads used in this study were also produced more than 50 folds coverage depth (≥100Gb) which can be applied into genome survey and error correction of the whole-genome assembly described at the Chapter 2.

1.2.2. Materials and Methods

Sample collection and library preparation

An adult male *C. opilio* was collected from coastal water of at the offshore of Yeongdeok-gun (the East Sea, South Korea) on March 14th, 2019. The specimen was immediately brought to the laboratory with its body temperature kept low with ice cubes in order to prevent the degradation of its mitochondrial DNA molecules. To minimize possible contamination, the surface of specimen was rinsed with pure water, and then with 70% ethanol. The muscular tissues (approximately 5g) were isolated from the fouth pereopods pairs. These isolated tissues were immediately buffered with RNAlater™ (Thermo Fisher Scientific, MA, USA) to prevent the possible nucleic acid degradation. The whole genomic DNA was extracted with phenol-chloroform manualized extraction following the manufacture's instruction of RNAlater™ reagent. Approximately 3μg from in total 15μg of extracted high-molecular DNA was used to prepare the library for Illumina paired-end sequencing. The kits and reagents were the same as those used for the *Liparis tanakae* genomic paired-end sequencing as described in Chapter 1. Finally, the information of the specimen was deposited to the NCBI with following accession numbers (PRJNA602365, SAMN13893315).

De novo assembly and finalization of mitogenome

The HiSeq X Ten instrument (Illumina, CA, USA) was applied to sequence two copies of libraries with their insert-sizes 350bp. The same kits and reagents in the Chapter1.1 were applied to prepare *C. opilio* nucleic acids to be sequenced. The raw paired-end reads whose summed total bases were estimated as 101.90Gb underwent filtering and trimming to satisfy quality score of Q30 and free from adapter sequences. For these filtering and trimming steps, the same protocols described at the Chapter 1 applied. Lastly, to obtain mitochondrial genomic reads for assembling mitogenome, 10,000,000 filtered reads were randomly sampled. The detailed statistics of generated nucleic acid reads are described at the next chapter, Chapter 2.

The MitoZ software (Meng et al., 2019) was applied to conduct *de novo* assembly for *C. opilio* mitogenome with its default parameters. The reference proteins of decapod mitochondria were collected from the NCBI Refseq in order to provide “baits” or hints of conserved mitochondrial sequences scattered with those randomly sampled paired-end reads. After the assembly was confirmed to be circular closed molecule with desired size range (approximately 15-20Kb), the mitogenome was automatically annotated with the MITOS webserver (Bernt et al., 2013). The annotated mitochondrial coding genes resulted from MITOS webserver were then thoroughly verified by compared to the other sequenced brachyuran genomes which were downloaded from the NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/). The gaps, overlaps, and long misalignments were manually curated with alignments using the NCBI BLASTP (Delaney et al., 2000) and MAFFT (Katoh et al., 2017). Finally, the overall coding gene structures and phylogenetic tree reconstructions were analyzed with the other published majoidean mitogenomes (Márquez et al, 2016; Basso et al., 2017), and also with 6 non-majoidean

brachyurans (Shi et al., 2015; Cheng et al., 2016; Karagozlu et al., 2018; Lin et al., 2018; Kim et al., 2019) and an anomuran used as an outgroup taxon (Gan et al., 2016). The amino acid sequences of 13 PCGs from 12 decapod species in total were first aligned with MAFFT using JTT substitution model in MUSCLE algorithm (Edgar, 2014). Then these alignments were concatenated manually and analyzed by the maximum likelihood method and Bayesian inference of evolution using RAxML 8.2.12-HPC (Stamatakis, 2014) and MrBayes 3.2.7 (Ronquist et al., 2012), respectively. Due to the large size of the dataset and massive computerization hours required for these analyses, the CIPRES Science Gateway providing accelerated phylogenetic analyses with clustered-computing was used (Miller et al., 2010; Miller et al., 2011). The most probable consensus phylogenetic trees for each analysis methods were calculated by substantial rounds of pseudoreplication (1,000 independent bootstrap replication for RAxML, and 1,000,000 generations of pseudoreplication for MrBayes were applied). The final version of *C. opilio* mitogenome sequence with its curated coding gene annotations was deposited in to the NCBI Refseq database with its accession number, MT335860.

1.2.3. Results

The assembled mitogenome of *C. opilio* was a closed molecule consisted with 16,067bp circular nucleotides and 37 mitochondrial genes (13 PCGs, 22 tRNAs and 2 rRNAs) as in described in **Table 7** and **Figure 11**. The GC content for the whole mitogenome was 28.40%, and both AT and GC-skew were negative while GC-skew showed strongly negative value (-0.226) compared to that of AT-skew, -0.032 (**Table 7**). PCGs generally showed negative AT and GC-skew, and 4 NADH dehydrogenase subunit genes (nd1, nd4, nd4l, and nd5) located on the (-) strand showed positive GC-skew which were reported as the general features of arthropod mitogenomes (Pisani et al., 2013).

The interactive comparison between other majoidean mitogenomes indicated that *C. opilio* mitogenome has its unique characteristics. The mitogenome had 3 unusually long overlaps between the genes spanning up to 7bp amongst 6 totall overlaps (**Figure 11**). The unusual losses or additions of long amino acids at 5' or 3' ends were found in products of 3 PCGs belonging to the NADH dehydrogenase subunit family (ND4, ND4l, and ND1) as in **Figure 12A**. In detail, 5' amino acids deletion was found from ND4L (6aa long, 5' MMDLSF missing), while 3' addition was found from ND4 (10aa long, 3' SLIKMKCVKR). The 3' end replacement was detected from ND1 (LNLIFN to WI). Furtheremore, a putative D-loop region between *rrnS* and *trnI* was annotaetd as the same location those of other brachyuran mitogenomes, however its length is especially longer (1,216bp) when it is compared to the lengths of D-loop of other brachyurans (Basso et al., 2017; Karagozlu et al., 2018; Kim et al., 2019; Márquez et al., 2016; Shi et al., 2015). In general, *C. opilio* tRNAs had common cloverleaf shaped secondary structures, and all 22 tRNAs lacked variable arms (**Figure 13**). However, 5 tRNAs showed atypical secondary structures; TψC arm without the loop (*trnF* and *trnR*), 1bp mismatch at the acceptor or

anticodon stem (*trnK* and *trnW*, respectively). In addition, the DHU arm of *trnS1* is extremely reduced with short stem (1bp) and loop (3bp). Furthermore, the organizations of mitochondrial genes among majoidean mitogenomes were investigated. The majority of mitochondrial genomic regions showed generally conserved synteny patterns with almost the identical gene organizations starting from *cox1* and reaching to *trnE*. While, *C. opilio*, *C. japonicus* and *Damithrax spinosissimus* shared the almost identical synteny, there was an obvious gene order rearrangement observed in *Maja crispata* and *Maja squinado* mitogenomes (Basso et al., 2017), as the authors described in their article. These putative translocation patterns (*nd6-cytb-trns2* segment between *trnE* and *nd1*) uniquely observed in two *Maja* species were described in **Figure 12B**. The most probable consensus phylogenetic trees analyzed from the concatenated amino acid sequences of 13 PCGs strongly supports the monophyletic conditions of the following clades; Majoidea, Heterotremata, Thoracotreamata, Eubrachyura, and Raninoidea, with 100% bootstrap values and 1.00 posterior possibilities (**Figure 14**).

Table 7. The overall statistics of assembled *C. opilio* mitogenome

Assembled *C. opilio* mitogenome	
Total length (bases)	16,067 (completely closed)
Number of A's (bases)	5,567
Number of G's (bases)	1,767
Number of T's (bases)	5,937
Number of C's (bases)	2,796
Overall AT skew	-0.032
Overall GC skew	-0.226
AT bias (%)	71.60

Name	Start	Stop	Strand	Length	ovl/nc*	Codons	Ini/Ter
cox1	1	1534	+	1534	0	ATG/TT*	Met i/*Ter(-A)
trnL2(taa)	1535	1599	+	65	9		
cox2	1609	2296	+	688	0	ATG/TT*	Met i/*Ter(-A)
trnK(ttt)	2297	2363	+	67	0		
trnD(gtc)	2364	2428	+	65	0		
atp8	2429	2587	+	159	-7	ATG/TAG	Met i/*Ter
atp6	2581	3255	+	675	0	ATT/TAA	Ille i/*Ter
cox3	3255	4044	+	790	0	ATG/TT*	Met i/*Ter(-A)
trnG(tcc)	4045	4108	+	64	0		
nad3	4109	4462	+	354	2	ATT/TAA	Ille i/*Ter
trnA(tgc)	4465	4528	+	64	-1		
trnR(tcg)	4528	4588	+	61	0		
trnN(gtt)	4589	4655	+	67	4		
trnS1(tct)	4660	4727	+	68	2		
trnE(ttc)	4730	4797	+	68	22		
trnH(gtg)	4820	4883	-	64	0		
trnF(gaa)	4884	4947	-	64	3		
nad5	**4951**	**6678**	-	**1728**	3	ATG/TAA	Met i/*Ter
nad4	**6682**	**8049**	-	**1368**	-7	ATG/TAA	Met i/*Ter
nad4l	8043	8324	-	282	20	ATA/TAA	Met i/*Ter
trnT(tgt)	8345	8408	+	64	0		
trnP(tgg)	8409	8471	-	63	2		
nad6	8474	8980	+	507	-1	ATC/TAA	Ille i/*Ter
cob	8980	10116	+	1137	-2	ATG/TAG	Met i/*Ter
trnS2(tga)	10115	10181	+	67	25		
nad1	**10207**	**11163**	-	**957**	5	ATT/TAA	Ille i/*Ter
trnL1(tag)	11169	11234	-	66	0		
rrnL	11235	12546	-	1312	0		
trnV(tac)	12547	12619	-	73	0		
rrnS	12620	13434	-	815	**1216**		
trnI(gat)	14651	14720	+	70	-3		
trnQ(ttg)	14718	14789	-	72	6		
trnM(cat)	14796	14862	+	67	0		
nad2	14863	15868	+	1006	0	ATG/TT*	Met i/*Ter(-A)
trnW(tca)	15869	15936	+	68	4		
trnC(gca)	15941	16003	-	63	0		
trnY(gta)	16004	16067	-	64	0		

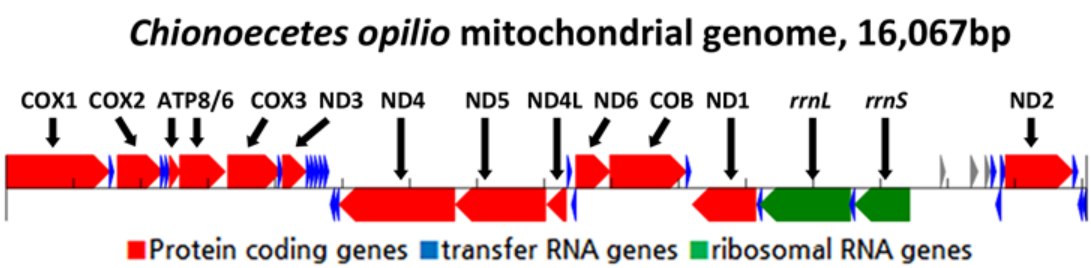

Figure 1. The structural information of annotated *C. opilio* mitogenome

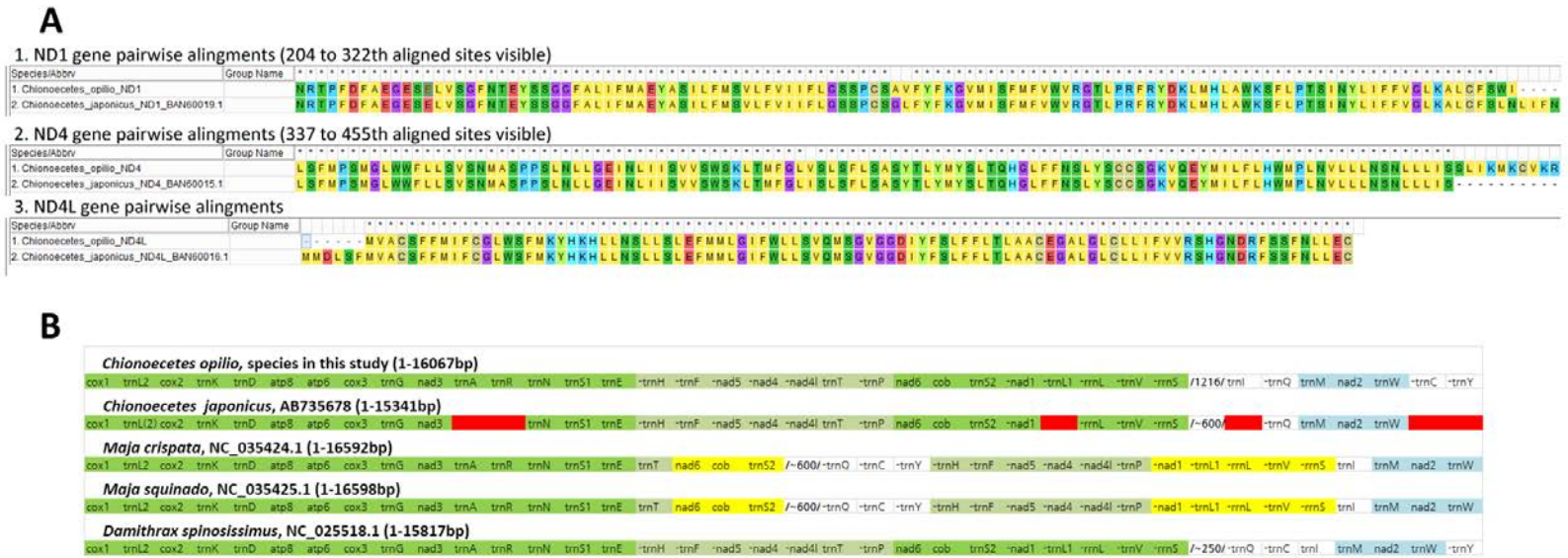

Figure 12. (**A**). The unusal amino acid deletion in 5' or 3' ends of three NADH dehydrogenase subunit genes visualized by pairwise sequence alignments with the identical genes from *Chionoecetes japonicus*. (**B**). The overall patterns of genetic synteny within 37 protein coding genes between majoidean mitogenomes indicated, Color index: **green**, highly conserved almost identical syntenies; **pale green**, less conserved syntenies which were disrupted with unique translocation in *Maja* species; **red**, 6 absent tRNA genes in *C. japonicus* mitogenome (*trnA*, *trnR*, *trnL1*, *trnI*, *trnC*, and *trnY*).

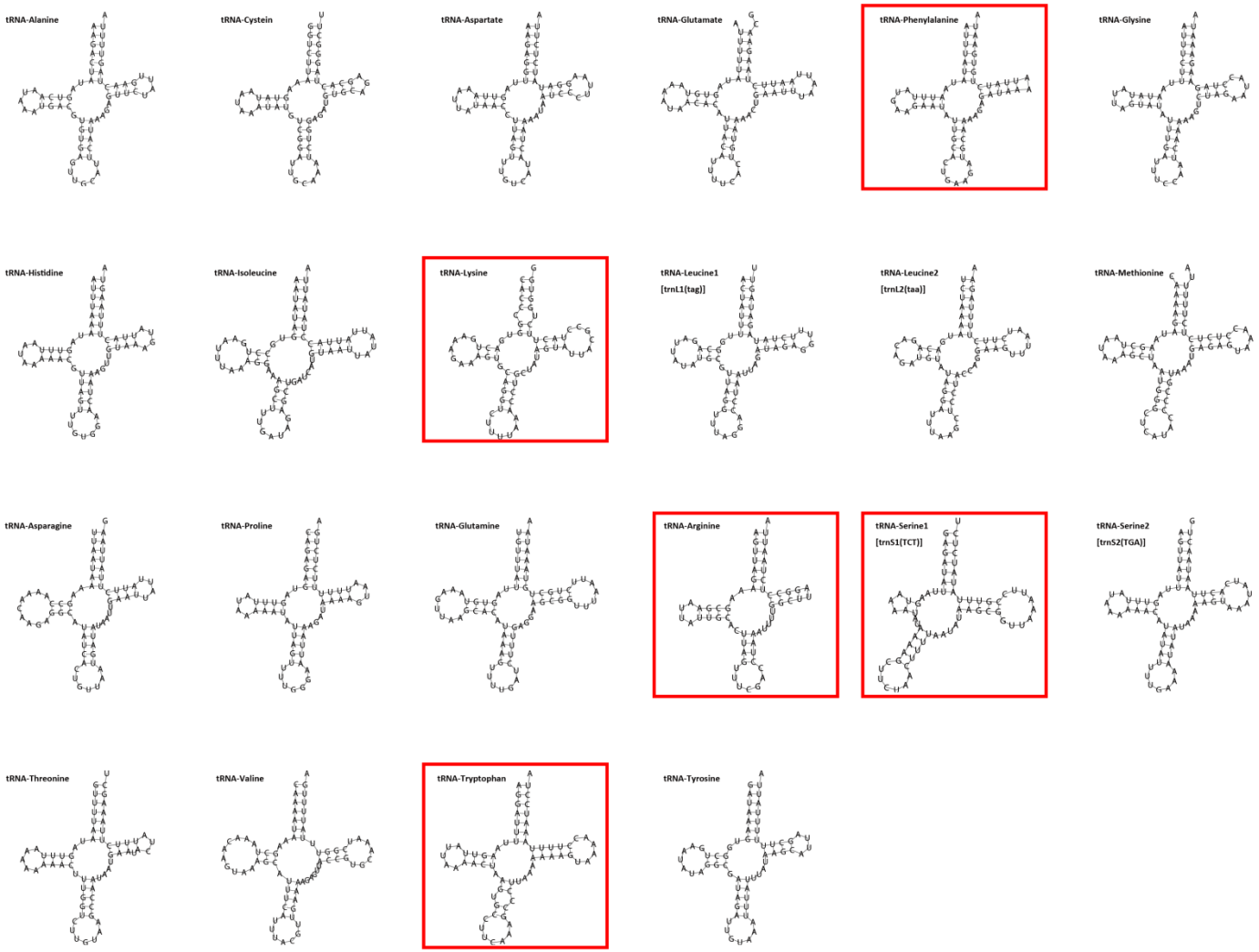

Figure 13. The overview of predicted secondary structures of tRNA gene transcripts; tRNA genes with unique characteristics described in the main text (*trnF*, *trnK*, *tRNR* and *trnY*) are indicated with the red-lined squares.

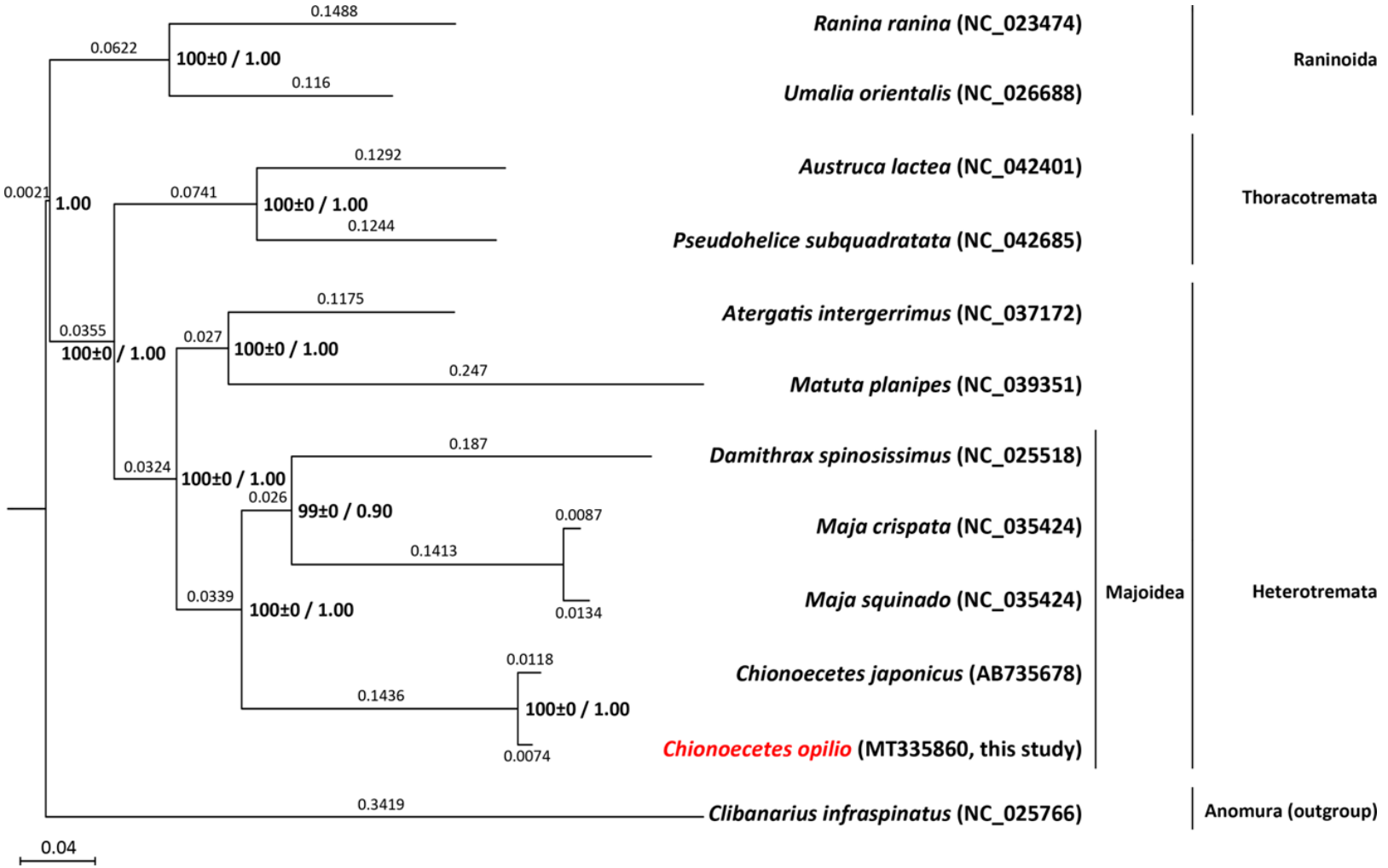

Figure 14. The phylogenetic tree showing relationships between *C. opilio* and 10 brachyurans with an outgroup taxon, *Clibanarius infraspinatus*

1.2.4. Discussion

The *C. opilio* mitogenome showed generally highly similar coding gene sequences with those of other majoidean mitogenomes, especially with those of congeneric species, *C. japonicus*. However, as in **Figure 12B**, there were significant differences between two *Chionoecetes* mitogenomes, the absences of 6 tRNA genes from *C. japonicus* and the D-loop of *C. opilio* which was almost the twice longer than those of *C. japonicus*. These differences were not likely probable considering the fact that the coding genes except 6 tRNA genes lack in *C. japonicus* showed more than 95% of amino acid sequence similarity and almost 100% of sequence coverage values. Therefore, the brief automatic gene annotation of *C. japonicus* was conducted with MITOS webserver (Bernt et al., 2013) with the same parameters previously used, in order to further investigate whether these *C. japonicus* mitogenomic features are genuine or artificial. The automatic annotation with MITOS successfully recovered 6 absent tRNA genes (*trnA*, *trnR*, *trnL1*, *trnI*, *trnC*, and *trnY*) in the NCBI-deposited mitogenomic sequence of *C. japonicus*, and their nucleotide sequences showed significantly high similarities with coverage values reaching almost 100%, when they were pairwisely aligned with the same 6 tRNA genes from *C. opilio* (**Figure 15A**). Furthermore, all these recovered *C. japonicus* tRNA genes were correctly located in its mitogenome with the same syntenic organization as those of *C. opilio* were (**Figure 15B**).

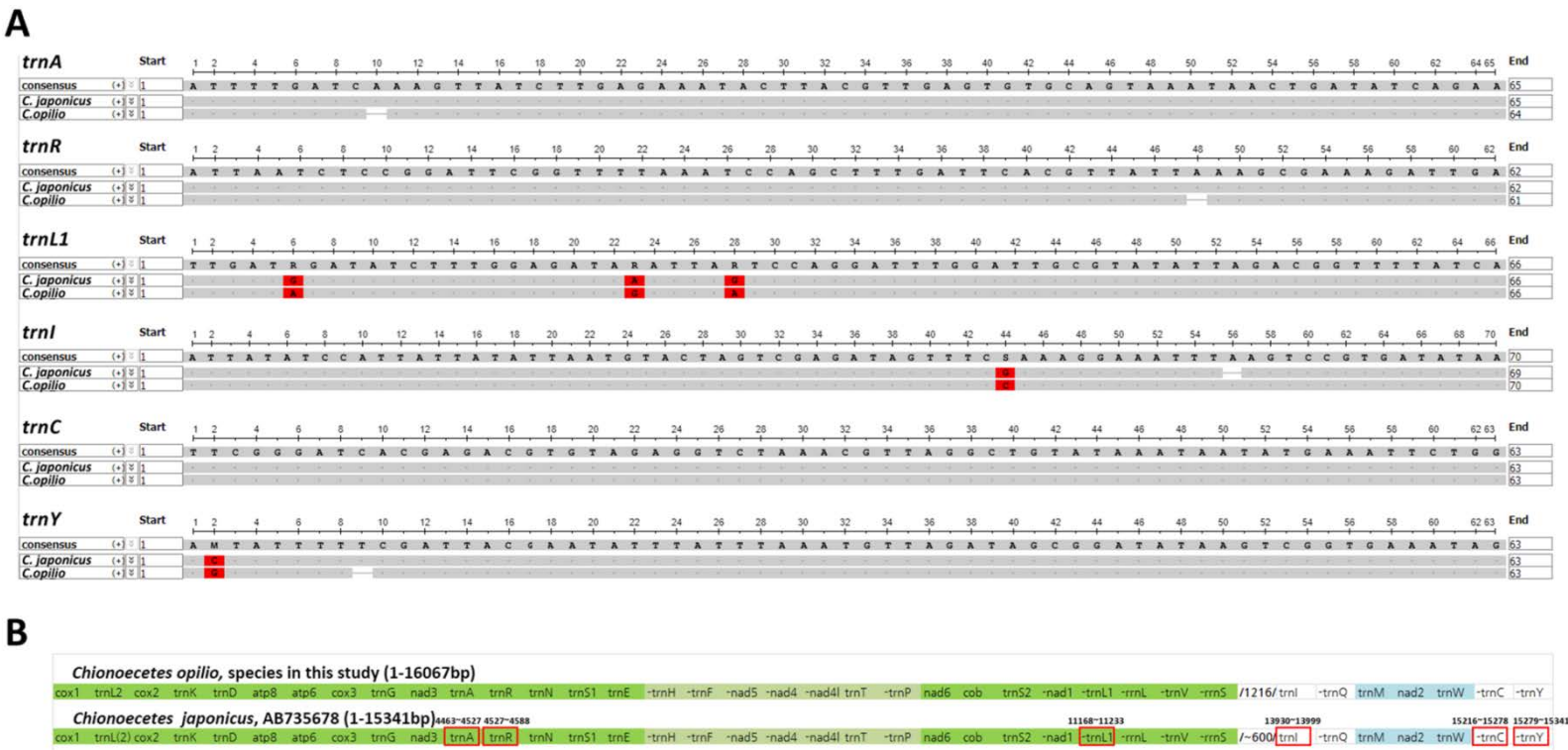

Figure 15. (**A**). The alignments between 6 transfer RNAs of *C. opilio* and MITOS-annotated *C. japonicus* mitogenomes, (**B**). The 6 tRNAs of MITOS-annotated *C. japonicus* mitogenomes were indicated with the red-lined boxes. There annotated genomic locations are further indicated above of each red-lined boxes.

This implies that it is necessary to continuously revise and renew the annotation of published genomic sequences, even if they were already reviewed and validated by the curators of the public biological data reservoir, such as the NCBI Refseq. In addition, the manual curation of *ab inito* genomic annotation is found to be essential for *de novo* assembled genomes, for instance in this study, three coding genes were revised significantly after the manual curation resulted in their transcriptome starting points and reading frames adjusted (COX2: -48bp, COX3: -42bp, and *rrnL*: +25bp) which resulted removal of all abnormally long overlaps between genes up to 47bp in the initial annotation results using the MITOS. The revised sequences of these manually curated genes were further aligned pairwisely with those of *C. japonicus* using NCBI Blast, and the pairwise alignments showed greatly improved similarities and removal of significant mismatches. In COX2 gene, the absence of 5' end MATWAYLGFQDASPL and the addition of 3' end of SPGDWKKVQVF were both removed after the manual curation, leaving only one amino acid substitution at the 50th site. Similiarly, in COX3 gene, the missing 5' end sequences (MTSSHSHHPYHLVD) and 19 amino acids long 3' end addition (WWGGYFFNMLVYLISNQKV) were removed by the manual curation, which resulted in only two positive substitutions of isoleucine in *C. japonicus* gene into valine in *C. opilio* gene at the 57th and 173th amino acid sites.

This study thus provides the significance of the manual curation, which can even revise the wrongly predicted open reading frames of *ab initio* annotated genes. In addition, both of the maximum-likelihood and Bayesian inference based phylogenetic analyses required the running-times than 3 hours for RAxML and 72 hours for MrBayes within an ordinary desktop with Windows operating system (a 3.00GHz processor with 8 threads, 6GB memories). The same analyses on the CIPRES Science Gateway, on the other hand,

required approximately 5 minutes for maximum-likelihood phylogenetic reconstructions with RAxML and 163 minutes for Bayesian inference analyses with MrBayes. Therefore, this study also indicated that the phylogenomic analysis of alignment matricies based on whole-genomic proteomes demands a computer-cluster based analytic server as in the CIPRES Science Gateway.

Chapter 2. THE *DE NOVO* GENOME ASSEMBLIES OF THREE MARINE ARTHROPODS

2.1. The first *de novo* assembled genome of *Portunus trituberculatus* indicating the bottlenecks in researching non-model marine arthropods

2.1.1. Introduction

As discussed in the backgrounds of this dissertation, Order Decapoda is significantly under-sampled crustacean clade despite of their economic and ecological impacts. The family Portunidae contains a number of famous edible crab species with wolrd-wide distribution. It is one of the most speciose families in Order Decapoda, with more than 410 species in 39 genera reported currently in the world (Ng et al., 2008). Portunid crabs inhabit a variety of marine environments such as muddy intertidal zone, pelagic water column, and deep water reaching 800 meters deep (Ng et al., 2008).

Portunus trituberculatus (Miers, 1876) is distributed primarily in the coast of East Asian countries. It is recorded as one of the most fished crab species since its annual amount of fishery occupies about a quarter of annual amount of worldwide commercial crab fishery (Liu et al., 2013). According to a 2016 report of FAO, 605,632 metric tons of *P. trituberculatus* were harvested in the year 2014 (FAO, 2016). Since its population is under continuous overexploitation, productivity of *P. trituberculatus* fishery has seriously decreased recently (Liu et al., 2013; Zhang et al., 2014). Yet, information about its whole-genomic affinity and resources is still limited. Although a *de novo* draft genome of *P. trituberculatus* has been reported recently (Lv et al., 2017), its datasets such as Illumina reads and assembly scaffolds still remain inaccessible. Its estimated genome size and reported genome size are 805.92 and 833.94 Mb that are relatively smaller compared to recent researches which estimated its genomic size using the flow cytometry (Li et al.,

2016). This study provides an alternative *de novo* assembled draft genome with its genomic annotations for *P. trituberculatus* for the first time in Korea. Although its assembly quality is not sufficient for satisfying a publication in peer-reviewed journals due to the bottlenecks of *de novo* genome sequencing and assembly processes in this study, the experimental trial and errors in this study contributes to the development and the optimization of workflows for marine arthropod *de novo* genome researches.

2.1.2. Materials and Methods

Sample collection and Whole-genome sequencing

Five adult male individuals of *P. trituberculatus* were collected from coastal water of Seosan, South Korea on December 10th, 2015 (36.615814°N, 125.242858°E). In order to prevent the degradation of genomic and transcriptomic nucleic acids, these specimens were brought to the laboratory alive immediately after the collection. The muscular tissues (approximately $1cm^3$) were isolated from each pair of the fouth pereopods for nucleic acid extraction to minimize the damage to the specimens. The genomic DNAs were extracted from these samples using the commercial DNA extraction kits suitable for Illumina Next-generation sequencing, QIAGEN Blood & Tissue Kit (Qiagen, Hilden, Germany). The transcriptomic RNA was extracted from muscular tissues of the other side of fouth pereopods of each individual with TRIzol® RNA Reagent (Thermo Fisher Scientific, MA, USA) following the manufacturer's instruction. The extracted DNA samples were validated with their quantities and qualities using agarose gel electrophoresis and instruments of NanoDrop 1000 spectrometer (Thermo Fisher Scientific, MA, USA) and 2100 Bioanalyzer (Agilent Technologies, CA, USA). The specimen coded as "Port_m005" was selected with its best qualitied DNA extract. The validated *P. trituberculatus* DNA and RNA extracts were about approximately 5μg, respectively. All the specimens collected in this study were deposited in to the Marine Arthropod Depository Bank of Korea (MADBK) in Seoul National University with voucher numbers (MADBK172910_021_001 ~ 005). Finally, the information of selected specimen coded as Potr_m005 (MADBK172910_021_005) was deposited to the NCBI with following accession numbers (PRJNA526559, SAMN11104290).

The same kits and reagents in the Chapter1.1 were applied to prepare *P. trituberculatus* nucleic acids to be sequenced. For constructing the DNA libraries with various insert sizes, TruSeq DNA Nano DNA Library Preparation Kit and Nextera Mate Pair Library Preparation Kit V2 (Illumina, CA, USA) were used (**Table 8**). TruSeq RNA library preparation kit v2 (Illumina, CA, USA) was applied to generate cDNA libraries from the transcriptomic extract. The nucleic acid extracts of *P. trituberculatus* underwent the same procedure of library preparation as described in the Chapter 1.1. The resulted sequencing libraries were validated by 2100 Bioanalyzer before the *de novo* genome sequencing. Finally, HiSeq X Ten insrument was applied to sequence genomic DNA libraries with HiSeq X Ten Reagent Kit v2.5 (Illumina, CA, USA). The transcriptomic library was sequenced with HiSeq 4000 instrument with HiSeq 4000 SBS Kit.

Table 8. The statistics of libraries and *de novo* sequenced reads of *Portunus trituberculatus* genome and transcriptome, after the quality control.

Library type	Insert-size (bp)	Read length (bp)	Total reads bases (bp)	No. of reads	GC (%)	Reads Q20 (%)	Reads Q30(%)
DNA, paired-end	350	151	81,969,028,146	542,841,246	41.81	92.00	85.92
DNA, paired-end	350	151	82,759,342,594	548,068,494	41.82	92.14	86.12
DNA, mate pair	3,000	151	8,655,838,195	280,809,228	43.08	92.66	85.68
DNA, mate pair	5,000	151	9,722,559,979	317,979,438	42.72	92.51	85.28
DNA, mate pair	8,000	151	5,549,331,548	185,422,654	42.48	91.83	84.28
DNA, mate pair	10,000	151	6,845,710,134	228,067,794	42.38	90.81	82.75
RNA, paired-end	350	101	10,121,527,769	101,233,994	47.41	99.14	96.94

De novo genome estimation and assembly

The raw genomic and transcriptomic *de novo* sequenced reads were verified and got removed with their adapters using FastQC v0.10.0 software and Trimmomatic, as described in the Chapter 1.1. The genomic size of *P. trituberculatus* was estimated by K-mer analysis using Jellyfish v1.1.11 software with 3 diff1ernt K-mer sizes (7, 21, 25bp) and flow cytometry approach. The additional two female *P. trituberculatus* were collected from coastal waters of Seosan, South Korea (36.615814°N, 125.242858°E). The hepatopancreas tissues were carefully separated from these specimens not to rupture their internal organs. Theses hepatopancreas tissues (approximately 2g each) underwent nuclei isolation by hydroshear homogenization. Then the obtained separate nuclei were stained with propidium iodide and estimated nuclear genomic DNA content per nucleus by flow cytometry analysis following a published protocol (Hare and Johnston, 2014). The C-value for *P. trituberculatus* was calculated by comparing its genomic DNA content per nucleus with to that of *Mus musculus* with its correleation to its genomic size. The contig-level *de novo* assembly was performed using SOAPdenovo2 (v2.04) (Luo et al., 2012) and Platanus v1.2.4 (Kajitani et al., 2014) with K-mer size parameter as variable and other parameters as default states, following the estimated genome sizes (approximately 1.5Gb) from K-mer analysis and flow cytometry. SOAPdenovo2 and Platanus v1.2.4 performed the scaffolding and gap-closing of *P. trituberculatus* genomic contigs using mate pair sequences to assemble scaffold level draft genome. These scaffolded and gap-closed genomic sequences were additionally validated with aspects of scaffold number, N50, contents of ambiguous bases, and BUSCO 2 (Simão et al., 2015) assessment with arthropoda_odb9 database. To reduce contents of ambiguous bases and incomplete BUSCO genes, Platanus was solely applied to re-conduct the scaffolding and gap-closing.

Transcriptomic analyses and genomic annotation

Trinity r20140717 (Grabherr et al., 2011) was used to assemble transcriptomic contigs from the filtered transcriptomic reads. These assembled contigs underwent clustering to reduce redundancy by CD-HIT-EST v4.6 software (Li and Godzik, 2006), and these clustered contigs were subjected to the ORF prediction using TransDecoder v 3.0.1 (https://github.com/TransDecoder/TransDecoder/) as previously described in Chapter 1.1. Following transcriptomic analyses of relative abundance calculation, mapping against the genomic sequences for obtaining hypothetical transcriptional sites were performed as the same processes and softwares used in the Chapter 1.1 (Trapnell et al., 2009; Li and Dewey, 2011). The functional annotation of these ORFs was also conducted following the softwares and biological databases as described in Chapter 1.1 (Cameron et al., 2004; Buchfink et al., 2015).

Seqping v0.1.33 (Chan et al., 2017) pipeline was applied to train and predict gene models based on the intrinsic and extrinsic evidences, and to merge independent predicted sets of gene models into consensus gene models. High-qualitied crustacean proteins were downloaded from NCBI Refseq Gene databases with following filter parameters “((Crustacea[Organism]) AND "source genomic"[Properties]) AND "srcdb refseq reviewed"[Properties]” and provided as the extrinsic evidences after CD-HIT clustering to reduce redundancy (Li and Godzik, 2006). The software RepeatMasker (Tarailo-Graovac and Chen, 2009) incorporated in the pipeline performed *ab initio* repetitive element prediction before gene model prediction. MAKER2 v2.28 (Holt and Yandell, 2011) was applied with default parameters to train GlimmerHMM v3.0.4 (Majoros et al., 2004), AUGUSTUS v3.2.2 (Stanke et al., 2006), and SNAP (2012/05/17) which conducted independent *ab initio* gene model prediction. MAKER2 analyzation was

repeated once more to merge these predicted gene models into the consensus gene sets to provide genomic annotations for *P. trituberculatus*. Finally, the same orthologous gene databases used for the functional annotation of predicted transcripts was used to perform functional annotation of final gene sets with default e-value threshold 1E-05.

Basic comparative genomic analysis

The reference proteomes of two thoroughly curated model arthropods (*Drosophila melanogaster*, *Daphina pulex*, and *Limulus polyphemus*, as in **Table 9**) were downloaded from the ftp service of the NCBI Refseq (ftp://ftp.ncbi.nlm.nih.gov/genomes/). The genomic sequences of two available decapod crustacean at the year 2016 (*Eriochier sinensis* and *Neocaridina denticulata*) were downloaded from the NCBI Refseq (ftp://ftp.ncbi.nlm.nih.gov/genomes/). OrthoMCL was used to perform orthologue analysis from these arthropod species with *P. trituberculatus*, except for *N. denticulata* whose proteome data was not accessible in public. Non-redundant orthologous genes that shared among all species resulted from both BUSCO and OrthoMCL analyses were selected and their amino acid sequences were aligned by MAFFT (Katoh et al., 2017), concatenated into supermatrix for RAxML 8.2.12 HPC (Stamatakis, 2014) analysis. The poorly aligned regions were trimmed using trimAl (Capella-Gutiérrez et al., 2009). A best-fit phylogenetic tree was reconstructed from RAxML 8.2.12 HPC using maximal likelihood method with parameter of "–m PROTGAMMAAUTO". In addition to the phylogenetic analysis, 3 more malacostracan arthropods with published genomes (*Exopalaemon carinicauda*, *Penaeus japonicus*, *Penaeus monodon*, and *Parhyale hawaiensis*) were investigated with comparison to *P. trituberculatus* genome for finding presence patterns of highly conserved developmental genes.

Table 9. The summary of downloaded 5 reference arthropod genomes in this study.

Species	Assembly ID	RefSeq accession	No. of genes	Data sources
Daphnia pulex	V1.0	GCA_000187875.1	30,907	Genbank reference genomes
Drosophila melanogaster	Release 6 plus ISO1 MT	GCF_000001215.4	30,559	RefSeq reference genomes
Eriochier sinensis	http://gigadb.org/dataset/100186	GCF_000972845.2	14,436	GigaScience database
Limulus polyphemus	Limulus_ polyphemus-2.1.2	GCF_000517525.1	38,676	RefSeq reference genomes

2.1.3. Results

The genomic reads of 164.73Gb (paired-end) and 30.77Gb (mate pair) were *de novo* sequenced from the male individual of *P. trituberculatus* as presented in **Table 8**. The estimated genome sizes from two independent measures, K-mer analysis and cyto flowmetry accorded with each other (approximately 1.3 to 1.5Gb, **Figure 16**). The initial version assembly of *P. trituberculatus* showed insufficient assembly quality which was indicated by its content of ambiguous bases occupying more than 52.34% of total length of the assembly (**Table 10**). In addition, the BUSCO assessment resulted only 224 complete genes (21.01%) from 1,066 arthropodan core orthologues with 605 missing genes which recorded as 56.75% (**Figure 17** and **Table 10**). Therefore, its genomic contigs were re-assembled without SOAPdenovo2 to reduce incorrect assemblies and excessively introduced ambiguous bases. The re-assembled *P. trituberculatus* genome showed almost identical genomic size (92.66%) to the initial assemblage, while its unambiguous base contents were dramatically decreased into 4.63% more than 10 folds (**Table 10**).

Table 10. The compared genomic statistics of the initial assemblage and the re-assembled genome of *P. trituberculatus*.

***P.trituberculatus* genome**	**Initial assembly (SOAPdenovo2+Platanus)**	**Final assembly (Platanus only)**
Total length (bp)	1,275,553,839	1,181,909,203
No. scaffold	1,423,367	2,675,465
Scaffold N50 (bp)	8,032	617
N's (%)	52.34	4.63
Transcripts Mapping ratio (%)	42.81	72.27
BUSCO		
complete (%)	21.01	61.82
partial (%)	19.61	14.35
missing (%)	56.75	23.83

Table 11. The statistics of finalized assembly and annotation of *P. trituberculatus*

A. Summary of statistics of the genome assembly	
Total bases (Mb)	1,181,909,203
No. of scaffolds	2,675,465
Average length (bp)	441
Maximum length (bp)	310,305
N50 (kb)	0.617
N's (%)	4.63
GC ratio (%)	42.20
B. Summary of statistics of the annotation	
Predicted gene models	87,564
Protein coding genes	34,536
Average transcript length (bp)	410
Average intron length (bp)	1,027
Average exons/gene	2.16
Average introns/gene	1.16
No. of tRNA	3,204
No. of rRNA	85

Although the revised assemblage was consisted with much more fragmented genomic sequences than the initial assemblage was (**Table 11A**), the BUSCO validation result was also greatly improved, with complete BUSCO genes recorded as 61.82% and less than about a half of missing BUSCO gene ratio (23.83% vs 56.75%, as in **Figure 17**). Furthermore, the transcriptome mapping ratio was also greatly improved from 42.81% of the initial version to that of 72.27% in the revised assembly. There were in total 87,564 predicted gene models with 34,536 of them were functionally annotated (**Table 11B**). The quality of the revised assembly then was further compared with those of other arthropod genomes which were available in year 2016 (**Table 12** and **Figure 18**).

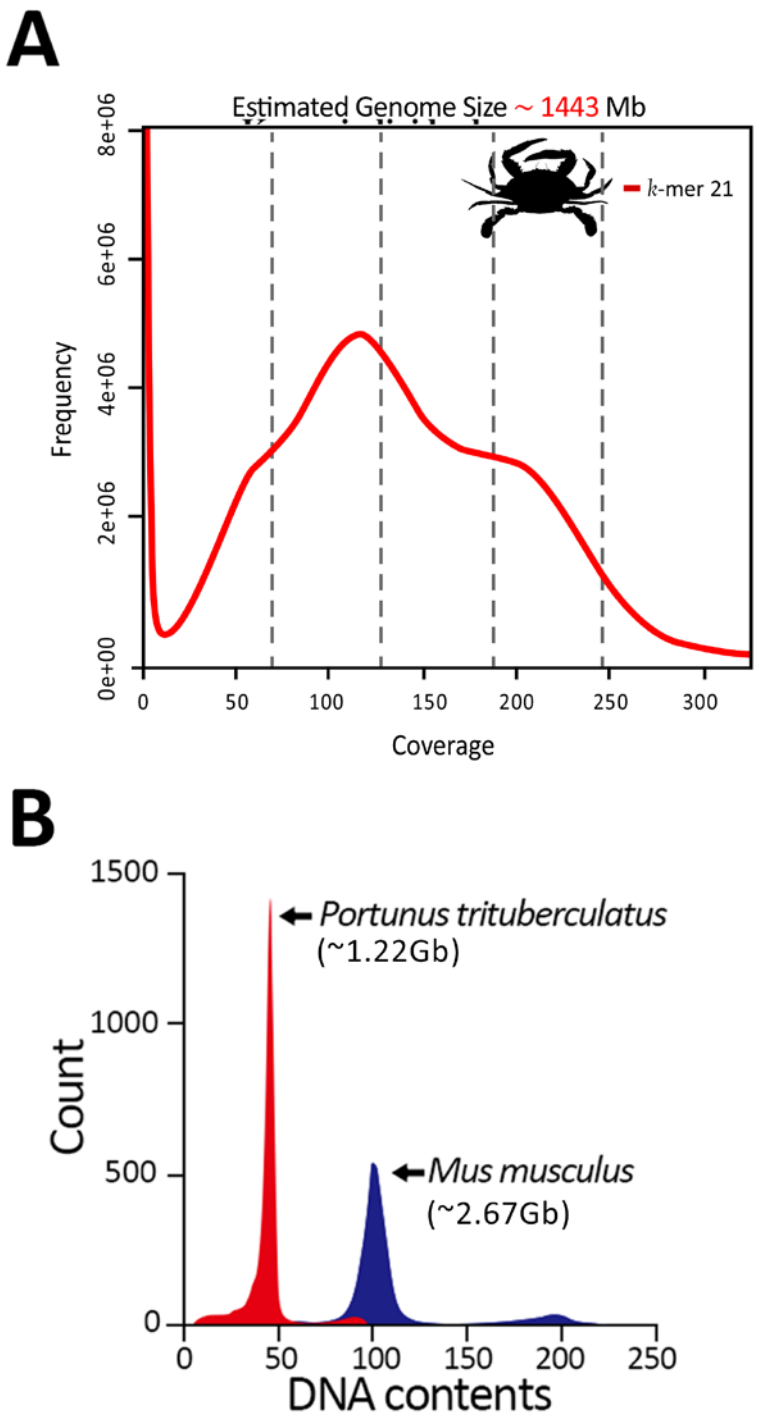

Figure 3. The well-according estimated genomic sizes between K-mer analysis (**A**) and flow cytometry approach (**B**) of *P. trituberculatus* genome.

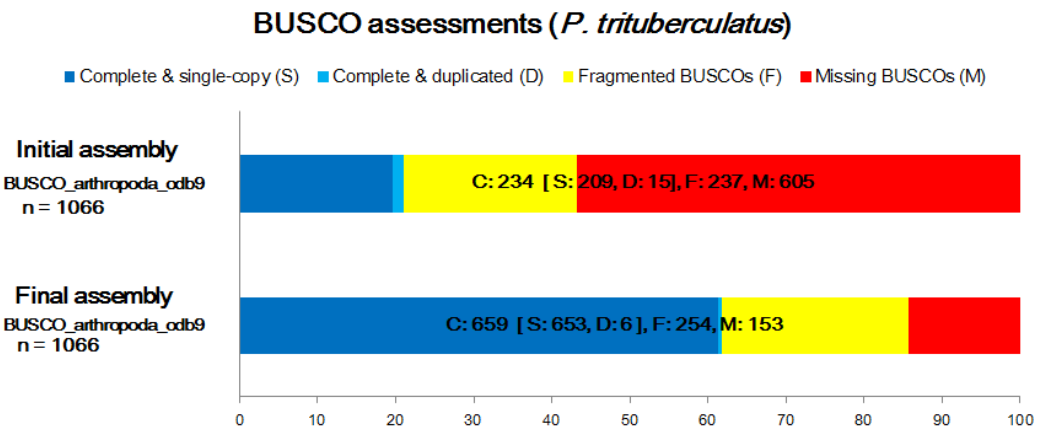

Figure 4. The comparison of BUSCO validation results between the initial assemblage and the revised, final assemblage of *P. trituberculatus* genome.

The core developmental genes conserved in bilaterian animals (Hugehes and Kufman, 2002), 10 *Hox* genes belonging to the *Hox* gene family were further investigated in *P. trituberculatus*, *N. denticulata*, *D. pulex*, and 4 additional malacostracan genomes (*Exopalaemon carinicauda*, *Parhyale hawaiensis*, *Penaeus japonicus*, and *Penaeus monodon*). The draft genome of *P. trituberculatus* was found to be intact with all 10 *Hox* genes with other 3 crustaceans, *D. pulex*, *N. denticulata*, and *Parhyale hawaiensis* (**Figure 19**). On the other hand, the *Hox* gene proboscipedia was absent in the draft genome of *Penaeus monodon*, Fushi tarazu, or *Hox7* orthologue, were not found in the *Exopalaemon carinicauda* and *Parhyale hawaiensis* genomes. Surprisingly, all core 10 *Hox* genes were identified in highly inaccurate and incomplete genome of *N. denticulata* (Kenny et al., 2014).

Table 12. The compared BUSCO validation statistics with 5 published genomes including marine species available in year 2016.

	Complete BUSCO, single-copied	Complete BUSCO, duplicated	Partial BUSCO	Missing BUSCO	Total BUSCO genes
***Portunus trituberculatus* (this study)**	653	6	153	254	1,066
Eriochier sinensis	525	19	54	468	1,066
Neocaridina denticulata	7	11	7	1,041	1,066
Daphina pulex	1,024	24	15	27	1,066
Drosophila melanogaster	540	526	0	0	1,066
Limulus polyphemus	534	479	41	12	1,066

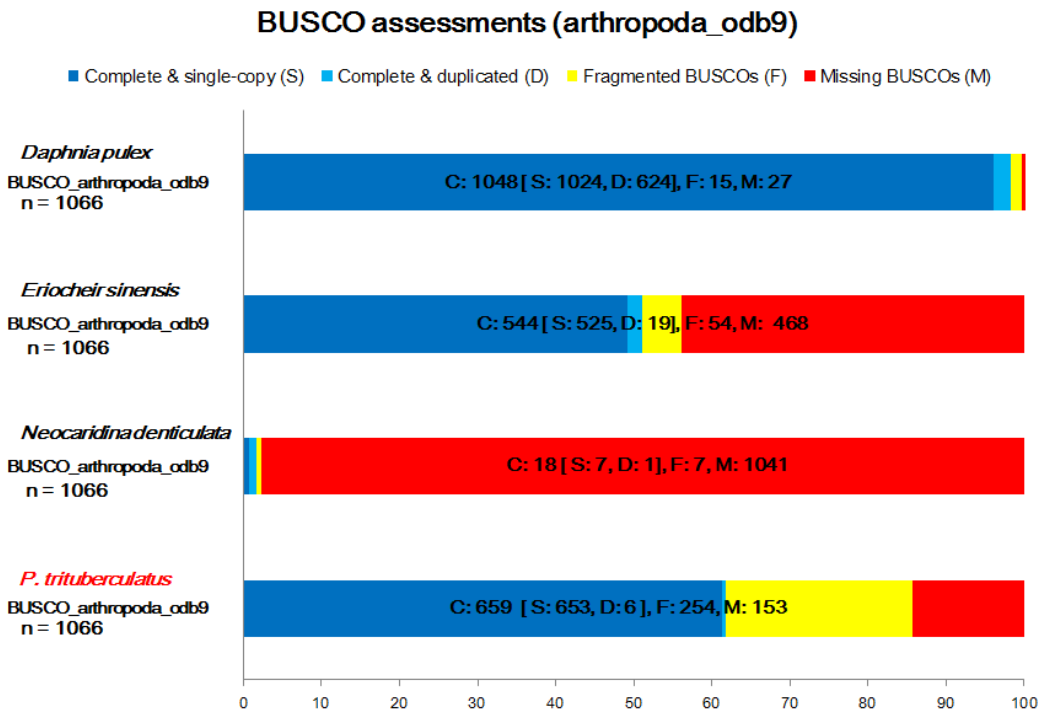

Figure 5. The visualized comparison of BUSCO validation statistics

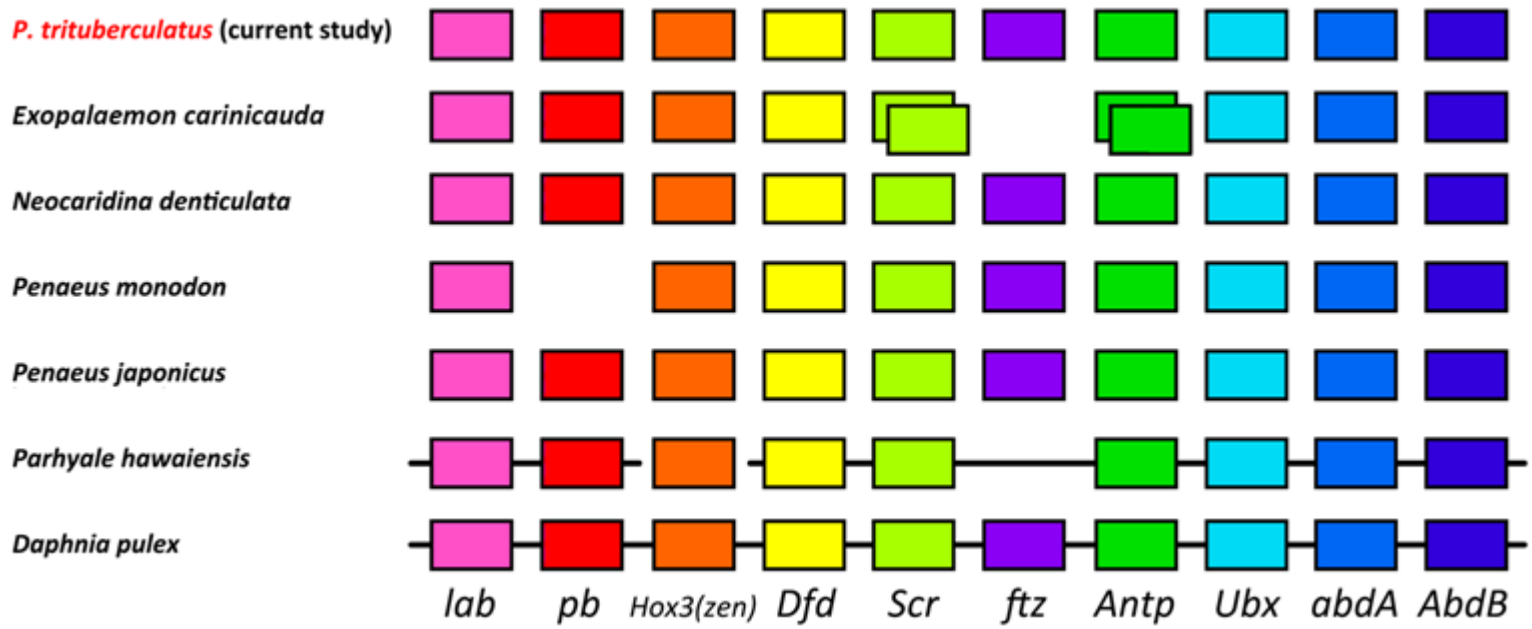

Figure 6. The comparison of *Hox* genes investigated with 5 available crustacean genomes in year 2016

2.1.4. Discussion

The *P. trituberculatus* draft genome was highly incomplete unlike that of *L. tanakae* in Chapter 1, which was further indicated by that more than a third of arthropodan core orthologues were missing or partial. Although the re-conducted assembly did improve the quality of *P. trituberculatus* draft genome, its highly fragmented genomic scaffolds negatively affected its genomic annotation. As the previous studies on non-model and short-read based genomes, these too short genomic sequences could result in the failure of predictions for some important genes, the overestimation of overall number of genes, and the incomplete amino acid sequences or improperly assembled haplotypic, polymorphic regions of these genes (Philliphy et al., 2008; Meader et al., 2010; Paszkiewicz and Studholme, 2010; Narzisi and Mishra, 2011). In addition, the highly fragmented scaffolds also affected the recovery of synteny between 10 *Hox* genes in all 4 decapods including *P. trituberculatus* in this study (**Figure 19**). As opposed to the two model arthropod genomes, those of *D. melanogaster* and *Parhyale hawaiensis*, with all or nearly all 10 *Hox* genes located in a single scaffold as a syntenic block, the *Hox* genes of 4 decapods genomes were found to be atomized since most of their genomic scaffolds were less than 1,000 bases long (**Table 2** from the Backgrounds section, and **Table 11**).

There are possible causes affecting such a low contiguity and completeness of the draft genome of *P. trituberculatus* although it was assembled by almost the same workflow to that of *L. tanakae* and with even larger sequencing coverage depth. Both species are well known to reproduce with the external fertilization, however, the relative strength of the dispersal of their eggs and larvae differs dramatically. In contrast to actinopterygiians like *L. tanakae* with their less dispersed egg and larvae, decapods including *P. trituberculatus* are reported to lay much smaller and easily dispersed egg and planktonic larvae, which

can greatly increase the genetic polymorphisms and heterozygosity ratios of their metapopulations (Domingues et al., 2010). It was reported frustratingly hard to sequence complex genomes with high ratios of polymorphism and heterozygosity, only with short read-length Next-generation sequencing technologies (Narzisi and Mishra, 2011; Bradnam et al., 2013). With further literature investigations, extracting the intact high molecular-weighted genomic DNA from decapod tissues are highly complicated, and if commercial extraction kits were carelessly applied to them, the microcolumns included on these kits were reported to increase fragmentation of extracted DNA (Bitencourt et al., 2007).

In summary, it is strongly recommended to extract high molecular genomic DNA following the manualized phenol-chloroform extraction protocol, and to sequence these high molecular-weighted DNA with the second or the third generations of Next-generation sequencing whose average read-length are much more elongated than that of Illumina paired-end sequencing. With these experiences of trial-and-error of this study, further *de novo* genome researches described in Chapter 2 were conducted with PacBio Single Molecule Real Time Sequencing (PacBio SMRT, Pacific Biosciences, CA, USA) which can produce genomic reads whose N50 longer than even 20Kb, with less than 1% of error rates (McCarthy, 2010; Rhoads and Au, 2015).

2.2. The high-qualitied marine arthropod assemblies : *De novo* assembled *Chionoecetes opilio* and *Nymphon striatum* genomes and their characteristics

2.2.1. Introduction

Backgrounds of *Nymphon striatum*

The Pycnogonida, or sea spiders, are essential arthropod taxa for understanding the early evolutionary history of arthropods and relationships between primary clades of this phylum. Studies based on both molecular and morphological data suggest that sea spiders have archaic origins as old as the early Cambrian. However, their early fossil records are incomplete and their extant members diversified relatively recently, implying that they are basal arthropods with a deep split origin (Dunlop and Selden, 2009; Rota-Stabelli et al., 2013; Sabroux et al., 2019). Multiple studies had reported that their morphologies such as reduced trunks, and 8 to 12 walking legs containing part of their digestive and reproductive organs are highly diverged and very unusual among arthropods (Sabroux et al., 2019). Their peculiar developmental morphologies such as the lack of a labrum and the presence of a terminal mouth instead of a ventrally opened mouth as in most arthropods have led the Cormogonida hypothesis, which places Pycnogonida as a sister taxon to all other arthropods, as opposed to the Chelicerata/Mandibulata hypothesis, which places them in a basalmost position nested in the monophyletic Chelicerata (Giribet, 2003; Machner and Scholtz, 2010). Despite numerous morphological and molecular studies on pycnogonid phylogeny, there has been no congruent settlement between these two conflicting phylogenetic hypotheses (Brennis et al., 2013; Giribet and Edgecombe, 2013).

The recent advances of the Next-generation sequencing technologies have enabled several phylogenomic studies based on widely sampled expressed sequence tags or transcriptomes, which have repeatedly designated sea spiders as basalmost chelicerates (Meusemann et al., 2010; Reiger et al., 2010; Rehm et al., 2011). However, the placement of Xiphosura has been controversial among recent phylogenomic studies incorporating *de novo* sequenced chelicerate genomes and transcriptomes, with the suggestion that Xiphosura partly (Sharma et al., 2014) or strongly (Ballesteros and Sharma, 2019) nested in the paraphyletic Arachnida, while another study found a conventionally accepted sister clade relationship between Xiphosura and Arachnida (Lozano-Fernandez et al., 2019). Although these phylogenomic studies thoroughly investigated *de novo* sequenced genomes representing almost all major arachnid clades and xiphosuran, however, several chelicerate taxa, including sea spiders lacked datasets based on *de novo* whole-genome. Therefore, high-qualitied genome assemblies of species representing these taxa are required to improve the resolution and reliability of the chelicerate phylogeny (Garb et al., 2018). To the best of current knowledge, no pycnogonid genome has been assembled or sequenced to date.

Backgrounds of *Chionoecetes opilio*

Among one of the most commercially important crustacean taxa, the Decapoda, the Infraorder Brachyura, or brachyuran (true) crab, is the most diverse decapod infraorder consisted of more than 6,500 extant species in 93 vailid families (De Grave et al., 2009; Ng et al., 2009). The Genus *Chionoecetes* contains seven species which are famous edible crabs from the waters of the North Pacific and the Northwestern Atlantic regions (Alvsvåg et al., 2009; Hardy et al., 2011; Ng et al., 2009). As previously referred, *Chionoecetes opilio* is recorded as the most commercially important species among its congeneric species, whose global annual catches had been exceeding 100,000 metric tons during the year 2007 to 2016 (FAO Fisheries and Aquaculture Department, 2019). The economic importance of this genus has led a number of researches on the various fields of *Chionoecetes* biology, including their physiology (Chung et al., 2015; Demian et al., 2013; Rahman et al., 2011), pathology (Mullowney et al., 2011; Ryazanova et al., 2016), population structures and phylogenies (Albrecht et al., 2014; Azuma et al., 2011; Kang et al., 2013; Johnson, 2019), and the hybridization between congeneric species by molecular methods (Kim et al., 2012). However, their whole-genome and transcriptome resources are required to understand further details of their biology, nevertheless, these resources are not currently available (Rotllant et al., 2018).

The decapod whole-genomes were reported to be highly complex due to their extremely large number of chromosomes, and large c-values (Lécher et al., 1995; Niiyama, 1966; Zhu et al., 2005) and their genomic complexity was suggested as major barriers against assembling high-qualitied genomes (Nguyen et al., 2018; Yuan et al., 2017). Although several decapod *de novo* whole-genomes were published recently, their species of interest were mostly limited to commercial shrimps (Kenny et al., 2014; Yuan

et al., 2017; Yuan et al., 2018) and their genomic sequences remained heavily fragmented. The genomic resources of brachyurans are much more limited than these non-brachyuran decapods both in the number of species and the data accessibility in public. Until the year 2019, only two draft genomes of true crabs have been reported yet their genomic resources remained insufficiently informative, due to lack of the reliable gene annotation and their fragmented assemblies with high contents of ambiguous bases (Song et al., 2016; Lv et al., 2017). To the best of current knowledge, there are only three cases of decapod genomes whose qualities are adequate to be reference genomes; the white legged shrimp, *Penaeus vannamei* (Zhang et al., 2019), *Eriochier sinensis* (Tang et al., 2020a) and *Portunus trituberculatus* (Tang et al., 2020b). These studies, nevertheless, have following limitations; the *Penaeus vannamei* genome was not primarily assembled with PacBio long-read sequences, the latter two high-qualitied crab genomes could not properly referred to understand the *C. opilio* biology, and lastly there is no currently NCBI verified annotated proteome of *Eriochier sinensis* genome.

Objectives of this study

Here, this study aims to provide the first cases of high-qualitied *de novo* assembled genomes of a common sea spider in the Korean waters, *Nymphon striatum*, and a deep cold-water living commercial crab, *Chionoecetes opilio*. These genome assemblies are the first *de novo* assembled genomes with reliable genomic annotations that represents the Class Pycnogonida and the Genus *Chionoecetes* by applicating high coverages of PacBio long-read sequencing. In addition, the *N. striatum* and *C. opilio* genomes assembled in this study will further provide proteomic resources required for the preliminary comparative genomic analysis on the evolution of deep branched arthropod clades which will be described in the Chapter 3.

2.2.2. Materials and Methods

Sample collections and preparations

The 40 individuals of *N. striatum* were collected by a collegue in the same laboratory, Damin Lee, at the location of Sacheon-hang, (37.82609°N, 128.93379°E, at a depth of 32 m, on 2018.07.12., NCBI BioSample accession ID: SAMN13567730) via SCUBA diving. These 40 sea spiders were brought to the laboratory alive, and then pooled together to compensate for the small size of the organisms and to ensure that the amount and quality of extracted DNA are acceptable for PacBio sequencing. All 40 sea spider individuals were collected from a single population at the same location to minimize the heterozygosity of the sequenced genomic reads. These pooled sea spiders were then buffered with RNAlater™ (Thermo Fisher Scientific, MA, USA) and lysed using QIAzol Reagent (Qiagen, MA, USA) according to the manufacturer's protocols. To isolate the DNA, the lysate was centrifuged according to the QIAzol Reagent protocol. The 15μg of genomic DNA was then extracted from the interphase of the lysate using the MG Genomic DNA Purification kit (Macrogen Inc, Seoul, Korea). The transcriptomic RNA was extracted from the same pooled organismal lysate using TRIzol® RNA Reagent (Thermo Fisher Scientific, MA, USA) following the manufacturer's instruction. The extracted nucleic acid samples were quantified by NanoDrop 1000 spectrometer (Qiagen, MA, USA) and qualified using a 2100 Bioanalyzer (Agilent Technologies, CA, USA).

The same adult male specimen of *C. opilio* and its genomic DNA extract were subjected in this study, which were described in detail in the Chapter 1.3. The specimen was collected from coastal water of at the offshore of Yeongdeok-gun (the East Sea, South Korea, 2019.03.14., NCBI accession number PRJNA602365, SAMN13893315). To minimize possible microbial contamination, firstly the surface of the specimen was rinsed with pure water, and then with 70% ethanol, and tools used for its dissection were also sterilized. From the specimen, four different tissues were isolated; the digestive gland tissues, the heart muscles, the muscular tissues, and the testicular tissues. The muscular tissues (approximately 5g) were isolated from the fouth pereopods pairs. To reveal the internal organs, the carapace was cut along its lateral edges. The epidermis underlying the carapace was carefully removed in order to prevent the disintegration of its organs. The digestive gland, testis, and heart were carefully isolated to avoid the possible contamination from collapsing irrelevant organs such as stomachs, guts, and gills. These isolated tissues were immediately buffered with RNAlater^{TM} (Thermo Fisher Scientific, MA, USA) to prevent the possible nucleic acid degradation. The whole genomic DNA samples were extracted with phenol-chloroform manualized extraction as following the same protocols described in the Chapter 1.3, which resulted in approximately 15μg of extracts per each type of tissues. The transcriptomic RNA samples were extracted from these tissues using TRIzol® RNA Reagent (Thermo Fisher Scientific, MA, USA) following the manufacturer's instruction. The 12μg of extracted high-molecular DNA from the muscular tissues was used to prepare the library for PacBio long-read sequencing and Illumina mate pair sequencing. In addition, four independent cDNA libraries were constructed from transcriptomic RNA extracts of *C. opilio* tissues with TruSeq RNA library preparation kit v2 (Illumina, CA, USA).

Whole-genome sequencings

To sequence the whole genomic and transcriptomic nucleotides of *N. striatum* and *C. opilio*, the PacBio Single Molecule Real Time (PacBio SMRT, Pacific Biosciences, CA, USA) and the Illumina sequencing technologies were applied. The PacBio long-read libraries were constructed from approximately 10μg (*N. striatum*) and 12μg (*C. opilio*) of genomic DNA extracts. A hydorshead system (Digilab, MA, USA) was applied to shear these DNA molecules into 8-12kb sized fragments. The PacBio SMRT library was constructed with C4 chemistry on a PacBio Sequel II platform (Pacific Biosciences, CA, USA). Two copies of 350bp insert-sized paired-end libraries and mate pair libraries with different insert-sizes were constructed (**Table 13 and Table 14**). Additionally, the cDNA libraries for the *N. striatum* and the four sampled tissues of *C. opilio* transcriptomes were also constructed (**Table 13 and Table 15**).

Table 13. The statistics of *N. striatum de novo* sequenced reads

Library type	**Insert-size (bp)**	**Read length (bp)**	**Total subreads bases (bp)**	**No. of subreads**	**GC (%)**	**Subread N50 (bp)**	**Average length (bp)**
PacBio SMRT	20,000	~20,000	84,833,283,304	5,480,059	35.27	20,750	15,480
Library type	**Insert-size (bp)**	**Read length (bp)**	**Total reads bases (bp)**	**No. of reads**	**GC (%)**	**Reads Q20 (%)**	**Reads Q30(%)**
DNA, paired-end	350 (2 copy)	151	98,760,925,162	654,045,862	35.22	99.79	98.61
DNA, mate pair	550	151	42,850,119,408	424,258,608	36.21	98.25	90.21
DNA, mate pair	3,000	151	40,912,863,052	405,077,852	36.22	98.5	92.46
DNA, mate pair	5,000	151	59,403,563,090	588,154,090	35.68	98.44	93.34
DNA, mate pair	8,000	151	17,360,189,161	171,883,061	35.30	99.12	93.54
DNA, mate pair	10,000	151	38,893,293,312	385,082,112	35.17	98.30	89.73
RNA, paired-end	350	101	13,074,260,928	129,448,128	52.06	98.37	95.36

Table 14. The statistics of *C.opilio de novo* sequenced genomic reads

Library type	Insert-size (bp)	Read length (bp)	Total subreads bases (bp)	No. of subreads	GC (%)	Subread N50 (bp)	Average length (bp)
PacBio SMRT	20,000	~20,000	201,361,187,452	23,504,401	41.30	13,535	8,556
Library type	**Insert-size (bp)**	**Read length (bp)**	**Total reads bases (bp)**	**No. of reads**	**GC (%)**	**Reads Q20 (%)**	**Reads Q30(%)**
DNA, paired-end	350 (2 copy)	151	105,604,752,180	704,510,174	41.32	98.32	96.68
DNA, mate pair	2,000	151	13,323,586,687	114,718,488	44.35	84.86	93.71
DNA, mate pair	5,000	151	13,776,748,112	115,628,464	43.27	84.94	93.66
DNA, mate pair	8,000	151	28,181,064,061	230,823,386	45.77	83.19	92.31
DNA, mate pair	10,000	151	49,285,131,114	375,149,202	48.16	84.01	92.59

Table 15. The statistics of *C.opilio de novo* sequenced transcriptomic reads

Tissue type	Insert-size (bp)	Read length (bp)	Total reads bases (bp)	No. of reads	GC (%)	Reads Q20 (%)	Reads Q30(%)
Digestive gland	2,000	151	13,323,586,687	114,718,488	44.35	84.86	93.71
Heart	5,000	151	13,776,748,112	115,628,464	43.27	84.94	93.66
Muscle	8,000	151	28,181,064,061	230,823,386	45.77	83.19	92.31
Testes	10,000	151	49,285,131,114	375,149,202	48.16	84.01	92.59

De novo Whole-genome assembly and its improvement processes

De novo sequenced genomic Illumina reads were assessed using FastQC v0.11.7 (Marçais and Kingsford, 2011) and then underwent adapter trimming and filtering (Q>30) as following the criteria described in the Chapter 1. These filtered genomic PE reads were subjected to the genome survey by Jellyfish v2.2.10 using its configurations of count step (-C -c 3 -s 100000000), merge step (default parameter), histo step (-h 10000000000), and k-mer sizes (17 bp, 21 bp, 25 bp).

To conduct contig-level *de novo* genome assembly of *N. striatum*, the HGAP4 software (Chin et al., 2013) was applied to assemble PacBio subreads with its default operating options for alignment, assembly, consensus, and polishing using the Arrow application. On the other hand, three different *De novo* genome assembly strategies were used to assemble *C. opilio* sequenced genomic PacBio reads, by comparing the performances of HGAP4, Wtdbg2 (Ruan and Li, 2019), and FALCON-Integrate with their respective default operating parameters and the genome size option as 2Gb. The FALCON-Integrate assembly further underwent FALCON-Unzip to merge heterozygous haplotypic contigs and increase contig N50. The assembled contigs of these genome assemblies were error-corrected by mapping filtered genomic PE sequences using default parameters of Pilon v1.21, followed by additional polishing by mapping PacBio reads using SMRT Link (v6.0.0.47841) to obtain consensus genomic contig sequences. To verify if there is any negative effect on the genomic assembly possibly caused by pooling 40 wildtype *N. striatum* individuals, the error-corrected contig-level of assembly was initially assessed using BUSCO v2 with eukaryota_odb9 and arthropoda_odb9. In addition, these three intermediate versions of *C. opilio* genomic assemblies were compared with their

respective BUSCO assessment results (database= eukaryota_odb9 and arthropoda_odb9) in order to validate the best qualited genome assembly for post-assembly analyses.

The Purge Haplotigs software (Roach et al., 2018) was applied to reorganize the initial contig-level *N. striatum* assembly into the revised contig-level assembly by removing the detected genomic reads redundancy. The Purge Haplotigs analysis was conducted by curating and merging haplotypic contigs by mapping PacBio subreads into the initial contig-level genome with its default parameters. The BUSCO assessment with the same databases used for the initial contig-level assembly was also conducted for these revised contig-level assembly. In addition, the K-mer analysis toolkit (Mapleson et al., 2017) was used to validate these two versions of contig-level assemblies before and after purging haplotypic contigs using its default parameters.

The Scaffolding Pre-assembled Contigs after Extension (SSPACE, Boetzer et al., 2010) program was used to scaffold the contigs of haplotig-purged *N. striatum* genome and the best qualited *C. opilio* genome with their repective mate pair reads. The gaps between genomic scaffolds were closed using PBJelly (English et al., 2012) and GMcloser (Kosugi et al., 2015). After gap closing, the scaffolds were polished once more using the SMRT Link to finalize the scaffold-level of the draft genomes of *N. striatum* and *C. opilio*. In order to assess the final versioned draft genomes, the same BUSCO assessment was performed.

Transcriptomic analyses and genomic annotation

The filtered transcriptomic reads of *N. striatum* and *C. opilio* were assembled into transcriptomic contigs by Trinity r20140717 (Grabherr et al., 2011). These assembled contigs underwent clustering to reduce redundancy by CD-HIT-EST v4.6 software (Li and Godzik, 2006), and these clustered contigs were subjected to the ORF prediction using TransDecoder v 3.0.1 In addition, the transcriptomic reads were mapped against the genomic scaffolds for obtaining hypothetical transcriptional sites were performed as the same processes and softwares used in the Chapter 1.1 (Trapnell et al., 2009). The functional annotation of these ORFs was also conducted following the softwares and biological databases as described in Chapter 1.1 (Cameron et al., 2004; Buchfink et al., 2015).

To reduce possible errors during *ab initio* gene model prediction, repetitive sequences of *N. striatum* and *C. opilio* genomes were predicted, annotated, and then masked with RepeatMasker (Tarailo-Graovac and Chen, 2009) v4.0.6 with custom sorted repeat library (based on RepBase 24.03). The *ab initio* and trained genome annotation was conducted by Seqping v0.1.33 (Chan et al., 2017) pipeline. To obtain extrinsic evidence for annotating *N. striatum* genome, high-qualitied chelicerate proteins were obtained from NCBI Refseq Gene databases with following filter parameters “((chelicerata[Organism]) AND "source genomic"[Properties]) AND "srcdb refseq reviewed"[Properties]”. For annotating *C. opilio* genome, the reference malacostracan proteins were downloaded from the NCBI Refseq with following filter parameters “((malacostraca[Organism]) AND "source genomic"[Properties]) AND "srcdb refseq reviewed"[Properties]”. To obtain extrinsic evidence inputs for *ab initio* gene prediction pipeline, these reference chelicerate

and malacostracan proteins were further clustered using CD-HIT PROTEIN to remove redundancy (Li and Godzik, 2006).

After the *N. striatum* transcriptomic reads were detected with excessive microbial contaminations, the transcriptome based intrinsic evidence for gene model training was abandoned. Instead of the contaminated transcriptome data, a homology-based software, GeMoMa (Keilwagen et al., 2018) was applied to cluster, analyze, and predict ORFs of *N. striatum* genome with high-qualitied chelicerate transcriptomic SRA datasets which were also downloaded from the NCBI.

Seqping v0.1.33 (Chan et al., 2017) performed the genome annotation of repeat masked genomic scaffolds with MAKER2 v2.28 (Holt and Yandell, 2011) initial gene model training, and then GlimmerHMM v3.0.4 (Majoros et al., 2004), AUGUSTUS v3.2.2 (Stanke et al., 2006) driven independent *ab initio* gene model prediction according to these training parameters. MAKER2 performed consensusing these predicted gene models into the finalized gene models with their genome annotation information, by discarding the gene models without any supports from the submitted evidences (eAED value=1). Finally, the following orthologous gene databases; Kyto Encyclopedia of Genes and Genomes (KEGG), NCBI nucleotide and non-redundant databases, Pfam, Gene ontology (GO), Uniprot and EggNOG, was used to perform functinaol annotation of these final gene sets of *N. striatum* and *C. opilio* with default e-value threshold 1E-05.

Basic comparative genomic and phylogenomic analyses

For orthologous gene analysis of *N. striatum*, Six ecdysozoan reference genomes were selected and downloaded (**Table 16**) from the NCBI. To conduct orthologous gene analysis for *C. opilio*, Seven ecdysozoan reference genomes were selected and downloaded (**Table 17**). BLASTP all-to-all search (Delaney et al., 2000) was conducted for these reference proteomes with two proteomes in this study. These BLASTP results were submitted as the input of OrthoMCL v2.0.9 (Fischer et al., 2011), which conducted the orthologue searching and clustering analyses using its default parameters. Datasets of non-chelicerate and non-mandibulate species were excluded for *N. striatum* and *C. opilio* respectively, in order to increase the visual legibility of the Venn diagrams of analyzed orthologues. To construct phylogenetic tree, non-redundant orthologous genes of *N. striatum* (106 genes) and *C. opilio* (160 genes) which were shared with all analyzed proteomes were curated for the supermatrix construction. The aminoacid sequences of these orthologues were aligned by MAFFT (Katoh et al., 2017) and then these alignments were concatenated by BeforePhylo (https://github.com/qiyunzhu/BeforePhylo) to obtain a supermatrix for phylogenomic analyses. The RAxML 8.2.12 HPC (Stamatakis, 2014) and MrBayes 3.2.7 (Ronquist et al., 2012) were used to conduct phylogenetic reconstruction of these analyzed ecdysozoans with 1,000 psudoeplicated maximum-likelihood and 3,000,000 pseudoreplicated Bayesian inference estimation of phylogenetic relationships, respectively. The amino acids substitution models used for these analyses were invariable Gamma distribution and the mixed, undefined substitution matrix (-m PROTGAMMAI for RAxML and –lset coding=variable Rates=invgamma, -prset aamodelpr=mixed for MrBayes).

Table 16. The summary of proteomes used in the orthologue analysis of *N. striatum*

Species	No. of genes	No. of clusters	No. of singletons	Data sources
Caenorhabditis elegans	28,416	7,628	5,069	Ensembl Metazoa Release 46
Centruroides sculpturatus	35,229	10,805	4,066	Ensembl Metazoa Release 46
Daphnia magna	26,646	8,289	9,539	Ensembl Metazoa Release 46
Drosophila melanogaster	30,559	8,858	2,259	Ensembl Metazoa Release 46
Limulus polyphemus	38,676	10,433	4,548	Ensembl Metazoa Release 46
Nymphon striatum	28,539	7,597	3,888	Current study
Parasteatoda tepidariorum	27,515	10,132	3,428	Ensembl Metazoa Release 46

Table 17. The summary of proteomes used in the orthologue analysis of *C. opilio*

Species	No. of genes	No. of clusters	No. of singletons	Data sources
Caenorhabditis elegans	28,416	7,718	5,076	Ensembl Metazoa Release 46
Chionoecetes opilio	22,659	7,305	3,075	This study
Daphnia magna	26,646	8,369	9,436	Ensembl Metazoa Release 46
Drosophila melanogaster	30,559	8,980	2,231	Ensembl Metazoa Release 46
Limulus Polyphemus	38,676	10,147	4,375	Ensembl Metazoa Release 46
Tigriopus californicus	15,577	7,347	5,128	Ensembl Metazoa Release 46
Parasteatoda tepidariorum	27,515	9,614	3,288	Ensembl Metazoa Release 46
Penaeus vannamei	33,273	10,495	6,574	Ensembl Metazoa Release 46

2.2.3. Results

The *Nymphon straitum de novo* assembled genome and its characteristics

The genome size of *N. striatum* was estimated at approximately 607 Mb while both the prominent heterozygosity peaks and relatively high (~1.9%) heterozygosity ratios from the genome survey (**Figure 20**). The initial contig-level genomic assembly was sized approximately 1.26 Gb long, of which the total base length exceeded twice those of the genome survey results (**Table 18A**). Furthermore, the BUSCO assessment on the initial contig-level assembly showed 53.47% of BUSCO genes from the initial assembly were duplicated which further indicates the presence of numerous redundant haplotypic contigs (**Figure 21** and **Table 18B**).

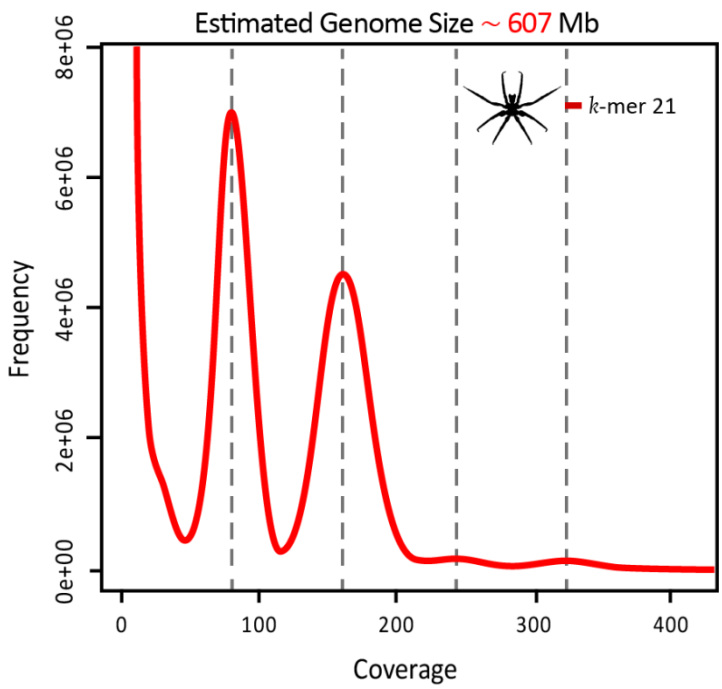

Figure 20. The estimated *N. striatum* genomic size indicating its significant heterozygosity ratio

Table 18. The summary of statistics of the initial and revised contig-level and finalized scaffold-level *N. striatum* genome assemblies

A. Summary of statistics of the genome assemblies			
	Initial assembly	Revised assembly	Final assembly
Total bases (bases)	1,260,501,127	732,914,915	744,788,989
No. of contigs	8,733	2,946	2,946
Average contig length (bases)	144,337	248,783	248,783
Maximum contig length (bases)	2,477,793	2,479,102	2,479,102
Contig N50 (bp)	221,141	360,904	360,904
No. of scaffolds	-	-	1,638
Average scaffold length (bases)	-	-	454,694
Maximum scaffold length (bases)	-	-	3,927,965
Scaffold N50 (bases)	-	-	701,800
N's (%)	0.00	0.00	0.04
GC ratio (%)	35.37	35.37	35.37
B. BUSCO validations of genome assemblies (arthropoda_odb9)			
	Initial assembly	Revised assembly	Final assembly
Complete BUSCOs (C=S+D) (%)	96.53	95.22	96.53
Complete & single-copy (S) (%)	43.06	78.42	78.80
Complete & duplicated (D) (%)	53.47	16.79	17.73
Fragmented BUSCOs (%)	0.94	1.69	0.94
Missing BUSCOs (%)	2.53	3.10	2.53
C. Brief statistics of genomic annotations			
Total bases, repeat elements (bases)			52,434,830
No. of repeat elements (hits)			564,918
Genome coverage of repeats (%)			7.14
No. of predicted genes			28,539
Genome coverage of gene regions (%)			55.06
D. Gene annotations			
Blast hits			27,086
No hits			1,453
Average gene length (bases)			2,130
Average intron length (bases)			1,311
Average exons/gene			10.33
Average introns/gene			9.33
No. of transfer RNAs			14,247
No. of ribosomal RNAs			308

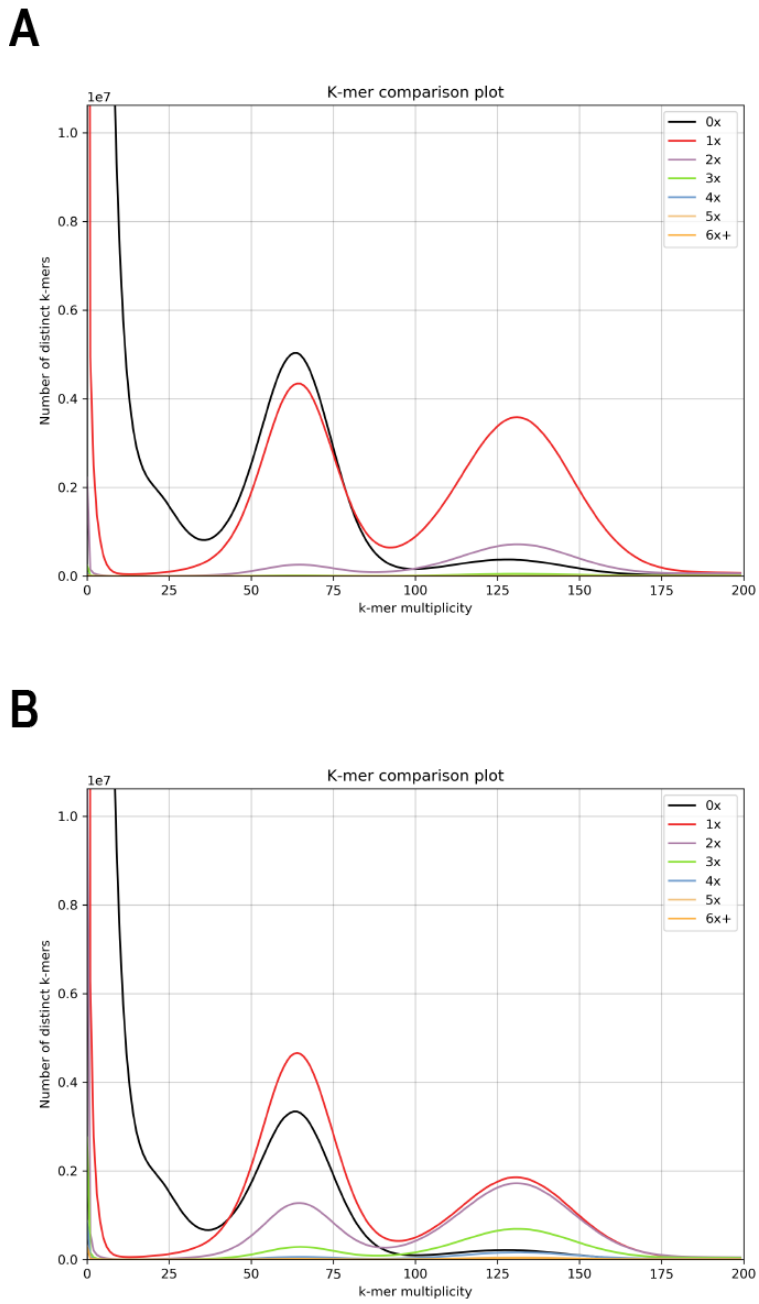

Figure 21. The comparison of *k*-mer distribution plots before and after the curation of purging haplotypic contigs. (**A**). The *k*-mer distribution plot of the initial assembly before the curation, (**B**). The *k*-mer distribution plot of the gap-closed final version of genome which underwent purging haplotypic contigs.

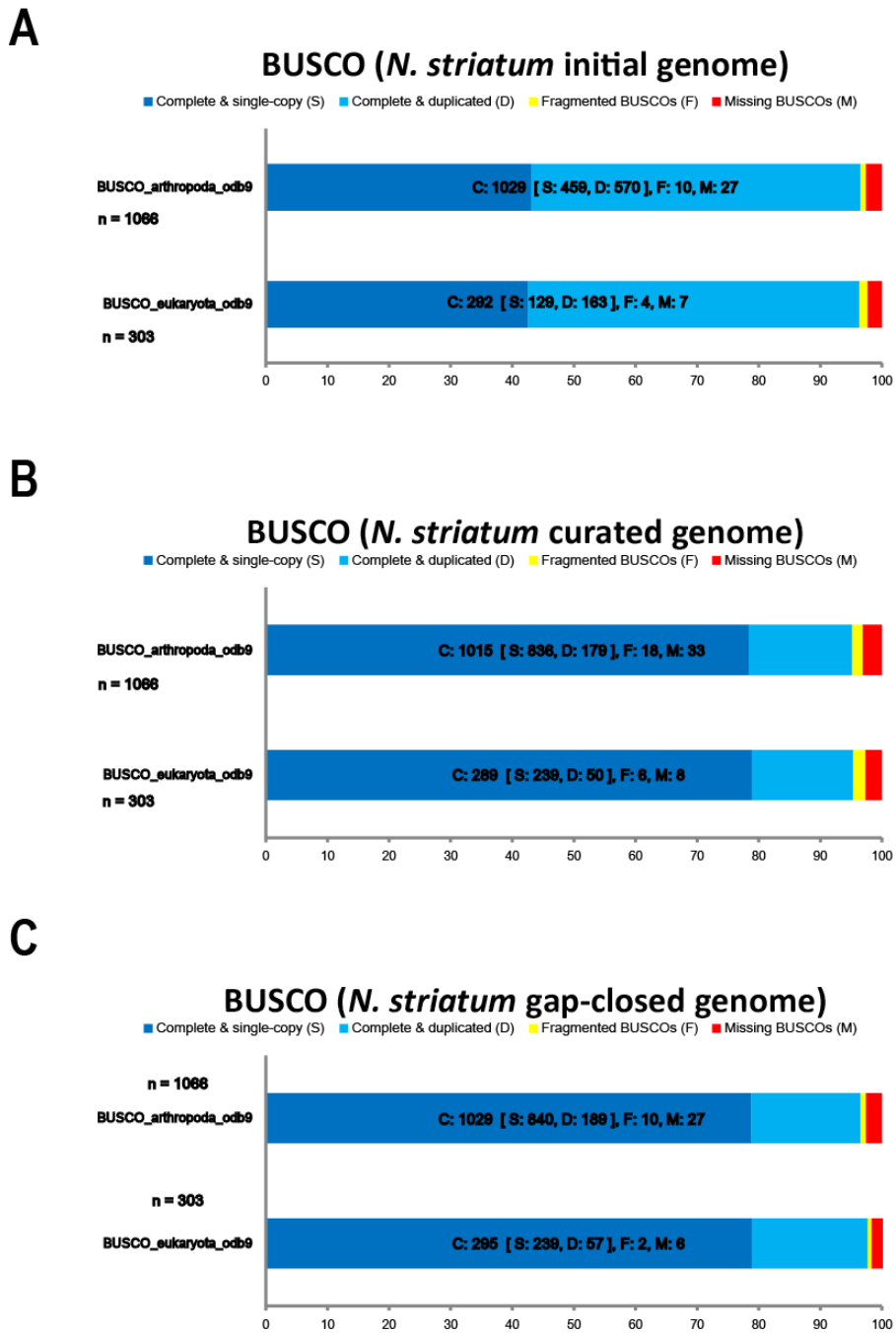

Figure 22. The results of BUSCO analysis of *N. striatum* genome assemblies. (**A**). The BUSCO result of HGAP4 initial assemblage, (**B**). The BUSCO result of curated assemblage by Purge_haplotigs, (**C**). The BUSCO result of gap-closed final assembly. **Color indexes**. deep blue: complete and single-copy genes; light blue: complete and duplicated genes; yellow: fragmented genes; red: missing genes

With these resulted indicating a highly heterozygous *N. striatum* genome, the Purge Haplotigs software (Roach et al., 2018) was applied to merge haplotypic contigs to reduce the redundancy of draft genome caused by its high heterozygosity ratio. The revised contig-level draft genome with purged haplotypic contigs showed great improvements in its quality of genomic assembly. While its total length was reduced to 732.9 Mb as the 121.78% of the estimated genome size, and contig N50 was almost doubled to 360.90 Kb, with the number of contigs reduced to 2,946 (**Table 18A**). The BUSCO assessment of the revised contig-level assembly found that the percentage of complete, but duplicated BUSCO genes were reduced to 16.79%, which is a substantial improvement from the initial BUSCO result (**Table 18B** and **Figure 22**). The K-mer distribution analyses using the K-mer analysis toolkit (Mapleson et al., 2017), further found that the ratios of heterozygous haplotigs were readily reduced as comparable as those of homozygous primary contigs after running Purge Haplotigs (**Figure 21**). The finalized scaffolded and gap-closied draft genome was composed of only 1,638 scaffolds with scaffold N50 increased to 701.80 Kb and a minute fraction (0.04%) of ambiguous bases (**Table 18A**). In addition, 31.9% of the genomic scaffolds were longer than 500 Kb, and among them, 159 scaffolds were over 1000 Kb. In particular, only seven scaffolds were shorter than 10 Kb, which indicated that *N.striatum* draft genome was highly completed (**Figure 23A**).

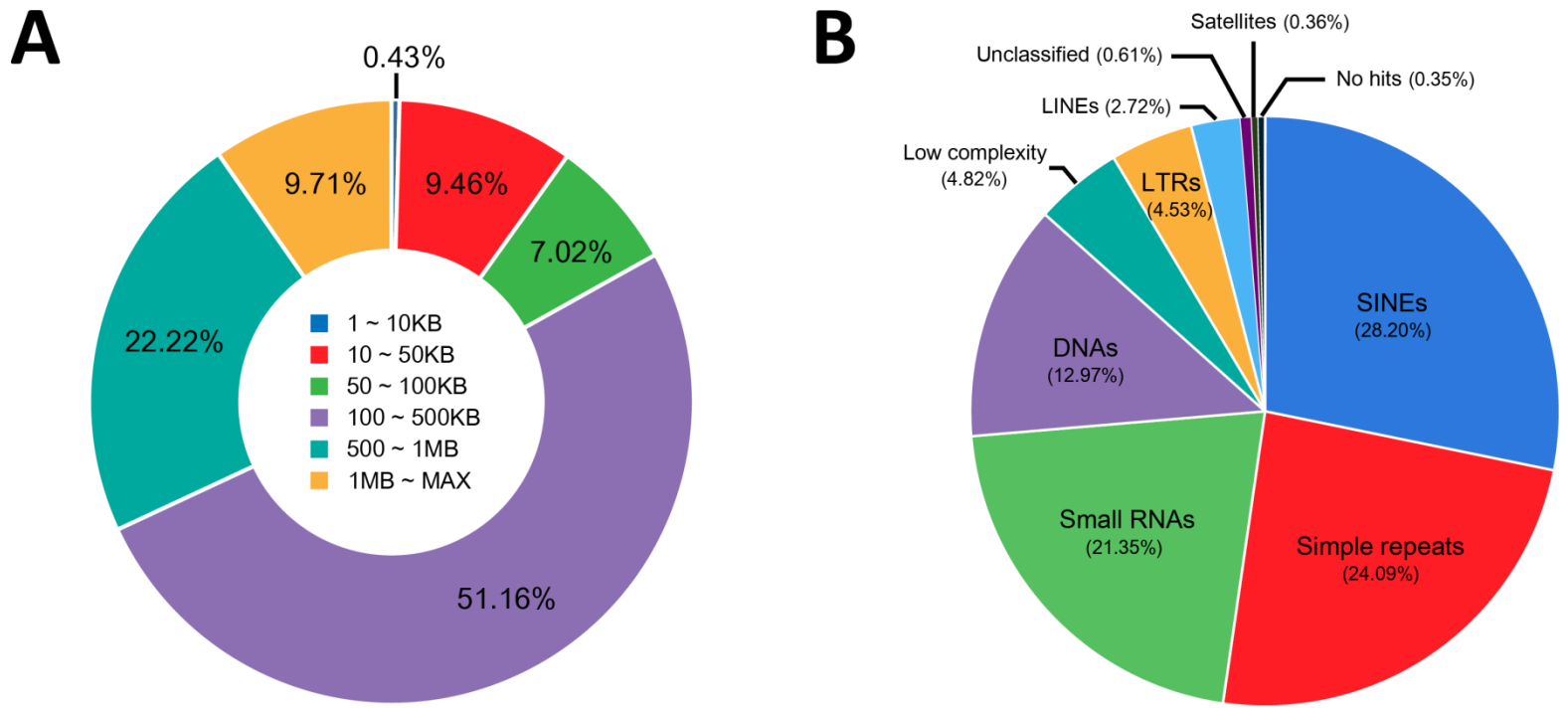

Figure 23. Characteristics of the *N. striatum* genome assembly. (**A**) the length distributions of the gap−closed scaffolds; (**B**) *ab initio* predicted repetitive elements and their subclass distributions.

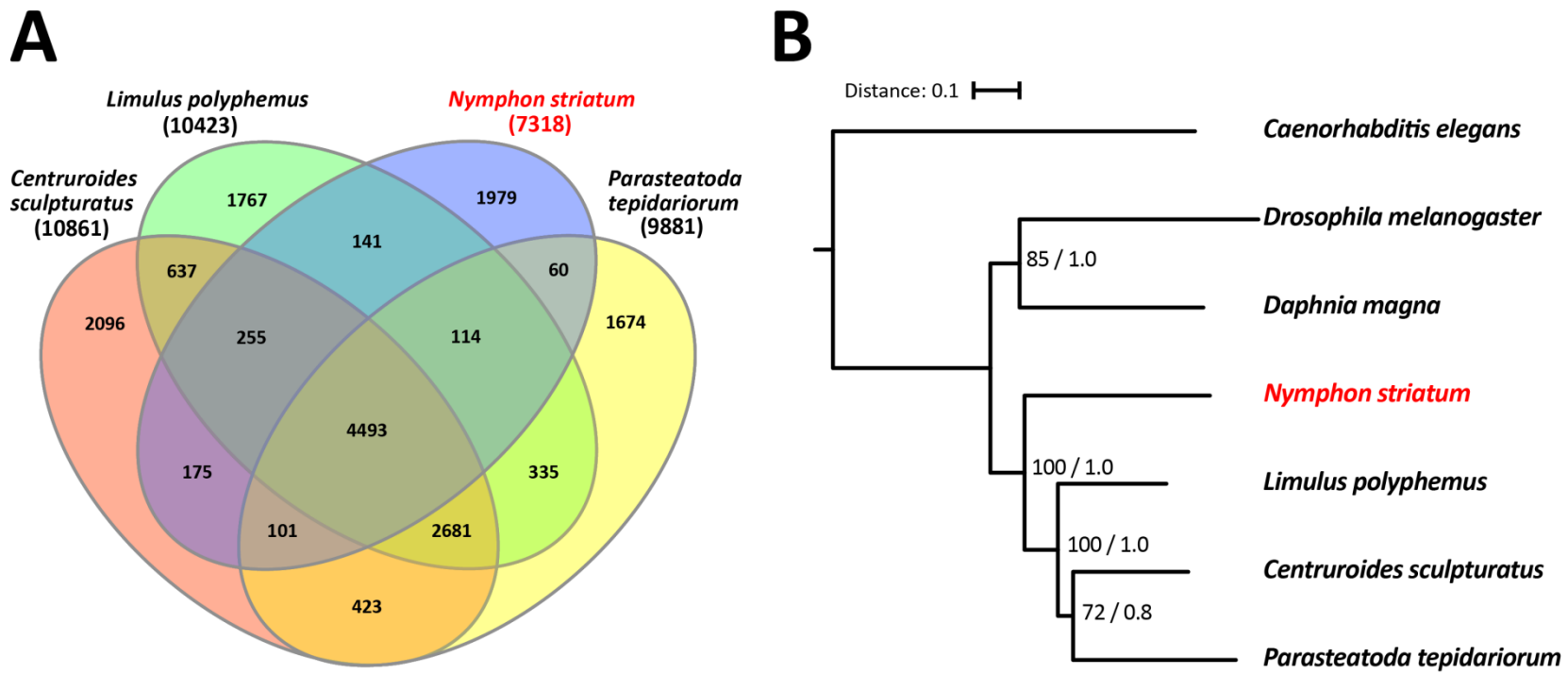

Figure 24. Comparative genomic analyses of *N. striatum* genome assembly. (**A**) a Venn diagram of the orthologous clusters among four chelicerate species; (**B**) the phylogenetic relationship of *N. striatum* with other six ecdysozoan species; *Caenorhabditis elegans* was selected as an outgroup taxon for the analysis. For each node, its bootstrap support value and the posterior probability are indicated at the base of the node.

An *ab initio* repeat element prediction resulted in a total of 7.14% of the genomic sequences being annotated as repetitive sequences (**Table 18C**). Among these repeat elements, short interspersed nuclear elements (SINEs) accounted for 28.20% of the total length of the annotated repeat sequences, and they were recorded as the most enriched repeats from the *N. striatum* genome. The SINEs were followed by simple repeats (24.09%), small RNAs (21.35%), and non-SINE interspersed elements (20.82%), while satellites (0.36%) occurred the least among the categorized repeat elements (**Figure 23B**).

The validation of *de novo* sequenced transcriptome of *N. striatum* was eventually concluded as the failure, which was indicated by the their genomic mapping ratio lower than 7% and the contents of predicted microbial orgined reads, such as *Vibrio* genera, exceeding 91%. Therefore, CD-HIT-EST clustered reference chelicerate transcriptomes were applied to the GeMoMa (Keilwagen et al., 2018) holmology-based ORF prediction, resulting in 362,016 hypothetical *N. striatum* ORFs. These homology-based modeled ORFs then underwent filtering out of incomplete transcripts, to obtain 220,011 predicted transcripts which were submitted as the intrinsic evidences for Seqping pipeline.

The final *N. striatum* genome consisted of 28,539 genes which spanned 56.01% of the total genomic length. Additionally, 14,247 transfer RNA and 308 ribosomal RNA genes were annotated (**Table 18D**). The orthologue analysis conducted using 6 ecdysozoan genomes (**Table 16**) resulted in 7,597 orthologous clusters shared within all 7 species and 3,888 singletons which were found to be unique for *N. striatum*. There were 4,493 orthologous clusters shared within four chelicerate species (**Figure 24A**). Phylogenetic tree reconstruction strongly supported the basalmost position of *N. striatum* nested in the monophyletic Chelicerata and a sister clade relationship between *Limulus polyphemus* with two arachnids (**Figure 24B**) with maximum support values.

The *Chionoecetes opilio de novo* assembled genome and its characteristics

A *C. opilio* genome was estimated to be approximately 1.89Gb in size with relatively high (~1.47%) heterozygosity ratios implied from the prominent heterozygosity peak (**Figure 25**). The initial trial of *de novo* assembly with HGAP4 was turned out into a failure due to the memory overflow. The other two assembler, FALCON-Integrate and Wtdbg2 were performed the contig-level of intermediate genomic assemblies (**Table 19A**). The quality of these intermediate assemblies were compared and verified by the assembly statistics and BUSCO assessments in terms of their contiquity, correctness, and completeness (**Table 19A, Table 19B** and **Figure 26**). All cases of the BUSCO statistics and assembly statistics such as total length, contig number and N50 indicated that the Wtdbg2 resulted assembly was the best-qualitied contig-level of *C. opilio* assembly.

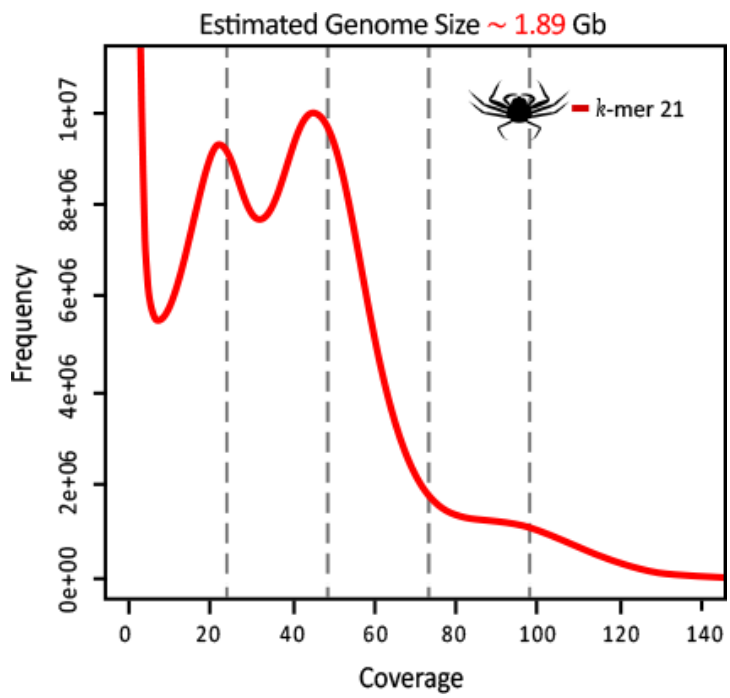

Figure 25. The estimated *C.opilio* genomic size indicating its significant heterozygosity ratio

Table 19. The summary of statistics of the initial and revised contig-level and finalized scaffold-level *C. opilio* genome assemblies

A. Summary of statistics of the genome assemblies			
	FALCON	Wtdbg2	Final assembly
Total bases (bases)	1,558,661,392	1,988,549,646	2,002,919,378
No. of contigs	22,381	45,098	45,098
Average contig length (bases)	69,642	44,093	44,093
Maximum contig length (bases)	1,433,041	2,094,150	2,094,150
Contig N50 (bp)	91,303	112,239	112,239
No. of scaffolds	–	–	26,514
Average scaffold length (bases)	–	–	75,491
Maximum scaffold length (bases)	–	–	2,536,572
Scaffold N50 (bases)	–	–	208,145
N's (%)	0.00	0.00	8.49
GC ratio (%)	41.31	41.31	41.31
B. BUSCO validations of genome assemblies (arthropoda_odb9)			
	FALCON	Wtdbg2	Final assembly
Complete BUSCOs (C=S+D) (%)	79.46	92.87	93.34
Complete & single-copy (S) (%)	54.78	91.18	91.46
Complete & duplicated (D) (%)	24.67	1.69	1.88
Fragmented BUSCOs (%)	4.22	2.53	2.16
Missing BUSCOs (%)	16.32	4.60	4.50
C. Brief statistics of genomic annotations			
Total bases, repeat elements (bases)			428,465,429
No. of repeat elements (hits)			3,467,483
Genome coverage of repeats (%)			21.68
No. of predicted genes			22,659
Genome coverage of gene regions (%)			0.06
D. Gene annotations			
Blast hits			22,659
No hits			4,401
Average gene length (bases)			6680.248
Average intron length (bases)			5705.681
Average exons/gene			4.147
Average introns/gene			3.147
No. of transfer RNAs			33,258
No. of ribosomal RNAs			274

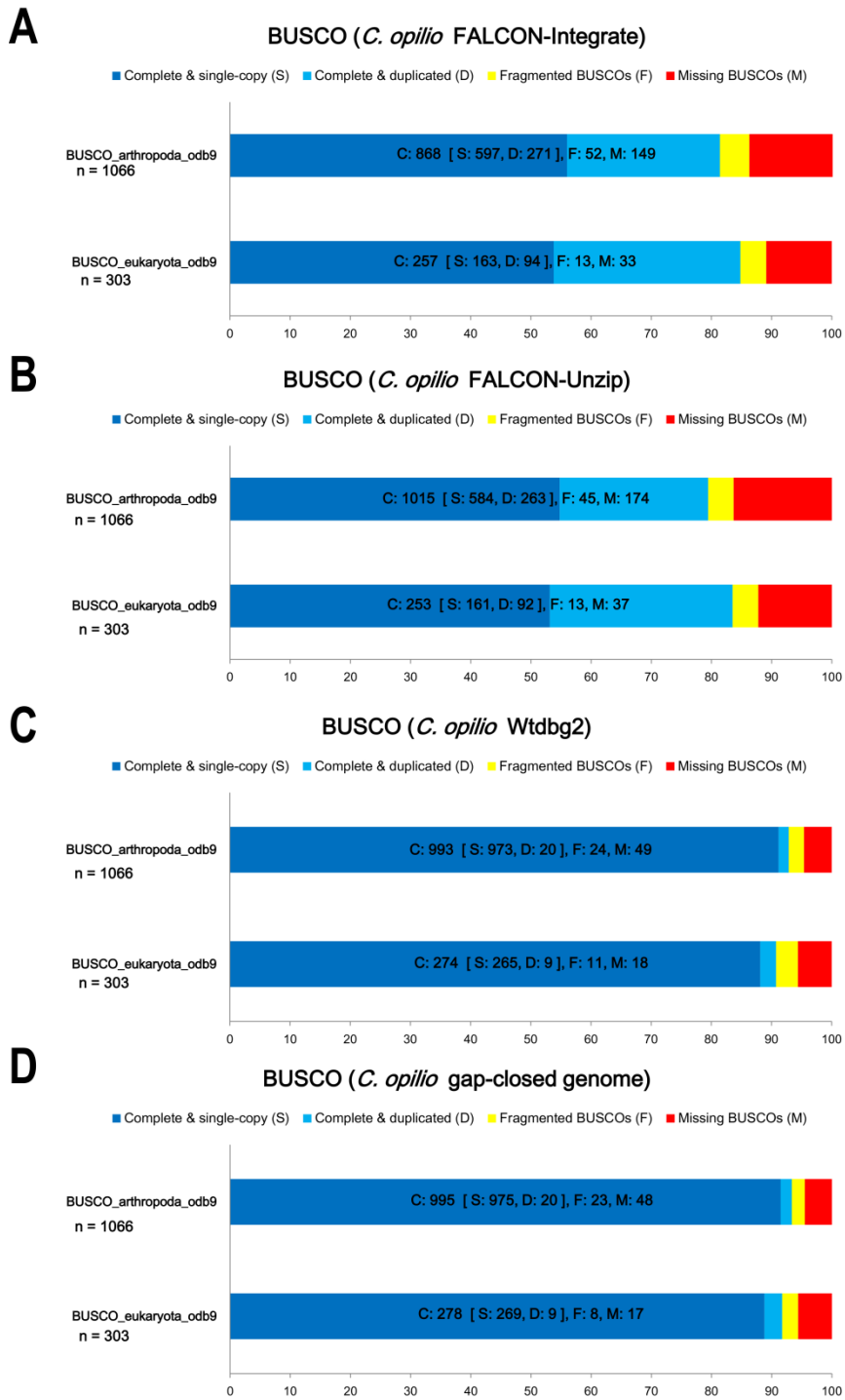

Figure 26. The results of BUSCO analysis of *C. opilio* genome assemblies; The BUSCO results of FALCON-Integrate assembly, (**A**); FALCON assembly after the FALCON-Unzip phasing, (**B**); Wtdbg2 assembly, (**C**); and the gap-closed final assembly, (**D**). **Color indexes**. deep blue: complete and single-copy genes; light blue: complete and duplicated genes; yellow: fragmented genes; red: missing genes

The final version of draft genome was consisted of 26,514 scaffolds with scaffold N50 208.145Kb and and ambiguous base content of 8.49% (**Table 19A**). In contrast to those of *N. striatum* final assembly, the occupancy of scaffolds shorter than 10Kb was recorded as 20.96% which indicated that the *C. opilio* final assembly shows insufficient genomic contiguity. There were only 521 scaffolds whose lengths were longer than 500Kb (1.96%) from the *C. opilio* draft genome (**Figure 27A**).

An *ab initio* repeat element prediction resulted in a total of 21.68% of the genomic sequences being annotated as repetitive sequences (**Table 19C**). Among these repeat elements, simple repeats accounted for the most abundant category of repetitive elements occupying 54.13% of the total length of the annotated repeat sequences. The simple repeats were followed by DNAs (14.08%), LINEs (10.92%), and low complexity repeats (6.35%), while unclassified repeats (1.35%) occurred the least among the categorized repeat elements (**Figure 27B**).

The final *C.opilio* genome consisted of 22,659 genes which occupied only 0.061% of the total genomic length, which was contrasted to those of repeat elements. Additionally, 33,258 transfer and 274 ribosomal RNA genes were annotated (**Table 19D**). The orthologue analysis conducted using 7 ecdysozoan genomes (**Table 17**) resulted in 2,459 orthologous clusters shared within all 8 species and *C. opilio* unique 4,075 singletons from 771 clusters. In sum, 3,250 orthologous clusters were found to be shared within five pancrustacean species (**Figure 28A**). The reconstructed consensus phylogenetic tree strongly supported the widely accepted relationships between these species, monophyletic clades of Arthropoda, Chelicerata, Decapoda (**Figure 28B**) with maximum support values. On the other hand, the trichotomic relationship was found between *Daphina pulex*, *Drosophila melanogaster*, and *Tigriopus californicus* (**Figure 28B**).

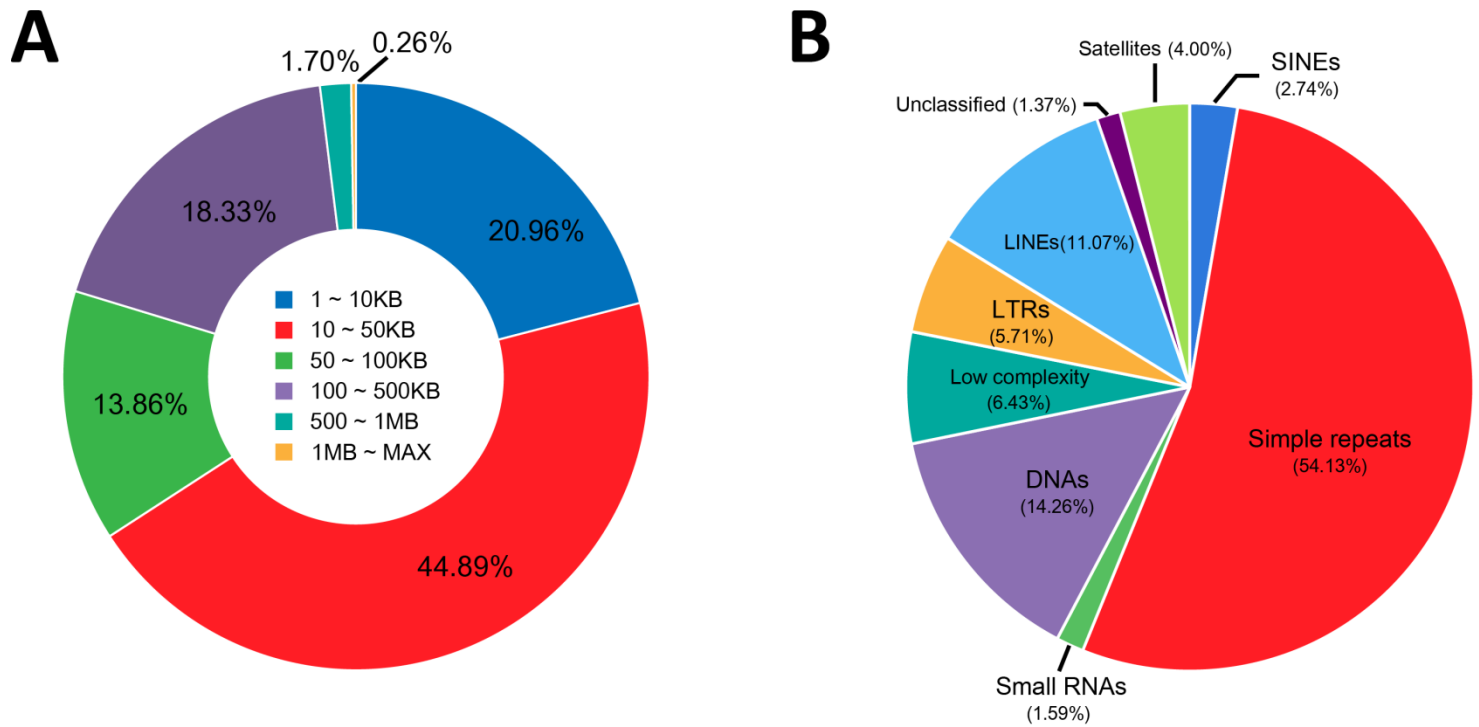

Figure 27. Characteristics of the *C. opilio* genome assembly. (**A**) the length distributions of the gap−closed scaffolds; (**B**) *ab initio* predicted repetitive elements and their subclass distributions.

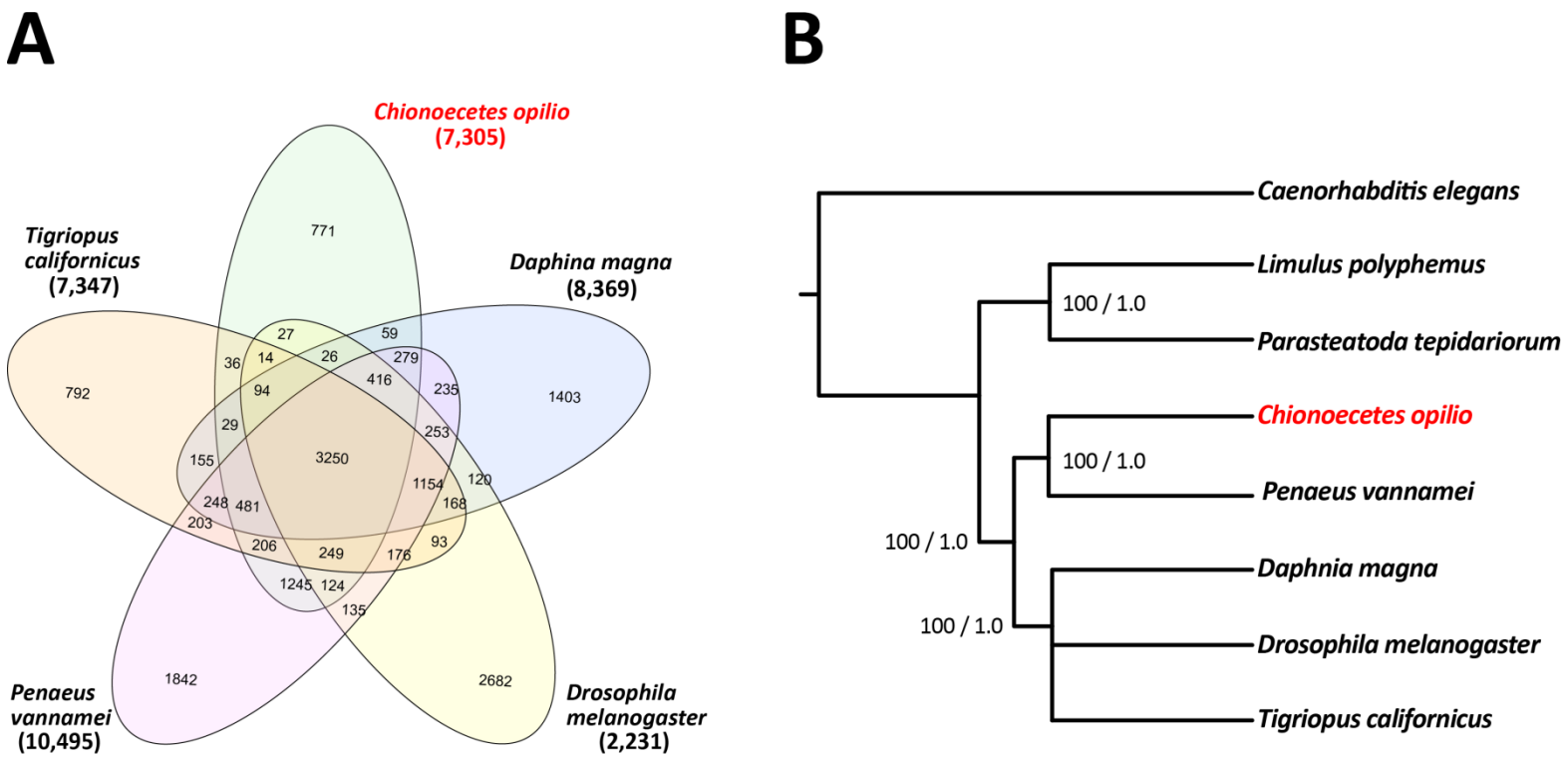

Figure 28. Comparative genomic analyses of *C. opili*o genome assembly. (**A**) a Venn diagram of the orthologous clusters among five pancrustacean species; (**B**) the phylogenetic relationship of *C. opilio* with other seven ecdysozoan species; *Caenorhabditis elegans* was selected as an outgroup taxon for the analysis. For each node, its bootstrap support value and the posterior probability are indicated at the base of the node.

2.3. General discussion

2.3.1. The *ab initio* prediction and annotation of marine arthropod *Hox* genes

The *Hox* genes are evolutionarily conserved transcription factors containing homeodomain motifs. They play critical roles in the patternization of the embryonic segments throughout the anterior-posterior axis. The hypothetical ancestral arthropod genome is suggested to contain 10 *Hox* genes (*labial*, *proboscipedia*, *Hox3*, *Deformed*, *Sex combs reduced*, *fushi tarazu*, *Antennapedia*, *Ultrabithorax*, *abdominal-A*, and *Abdominal-B*) as a well conserved cluster (Akam et al., 1994; Pace et al., 2016). In addition, these 10 *Hox* genes are usually found to be organized in a cluster as the same order of their domains of expression throughout the embryonic anterior-posterior axis, which is typically described as spatial collinerarity (Pace et al., 2016). These *Hox* genes have been intensly studied within diverse clade in the Phylum Arthropoda which supported 10 *Hox* genes generally conserved (Cook et al., 2001; Hughes and Kaufman, 2002). On the other hand, the conservation of their genomic arrangements remains unclear throughout the phylum, due to the majority of researches focusing on to the subphylum Hexapoda (Pace et al., 2016). Therefore, most of the cases from non-hexapod arthropod *Hox* gene studies have been conducted as gene-based surveys, and some early studies could not recover several *Hox* genes from decapod species (Mouchel-Vielh et al., 1998; Abzhanov and Kaufman, 2000; Deutsch and Mouchel-Vielh, 2003). Among pycnogonids, the study on their *Hox* genes also was not able to recover some *Hox* genes, such as *abdominal-A* (Manuel et al., 2006).

Nevertheless, considering that few examples of *Hox* gene losses were reported witin arthropods, the loss of several core *Hox* genes from decapods and pycnogonids are not plausible (Pace et al., 2016). In recent genomic studies, however, entire 10 *Hox* genes

were present in *de novo* assembled decapod shrimp genomes, such as *Neocaridina denticulata* (Kenny et al., 2014), *Penaeus japonicus* (Yuan et al., 2018) and *P. vannamei* (Zhang et al., 2019). In addition, all 10 *Hox* genes were recovered from the *Portunus trituberculatus* and *Chionoecetes opilio* genomes in this study. On the other hand, *fushi tarazu* was absent in *E. carinicauda* genome (Yuan et al., 2017) and *proboscipedia* was not present in *P. monodon* genome (Yuan et al., 2018), which was unnatural considering that their closely related species show fully recovered 10 *Hox* genes (**Figure 19**). Since both of *E. carinicauda* and *P. monodon* demonstrate highly uniform "decapod shrimp" morphology, their lack of *fushi tarazu* and *proboscipedia* is possibly resulted from incomplete genome assembly or gene prediction. Except *P. vannamei*, all referred decapod shrimp genomes were *de novo* assembled solely from the short read-lengthed Illumina paired-end sequenced reads, thus their genomic assemblies consist with more than a million scaffolds with N50 value less than 1,000 bases long (**Table 2**).

As previously discussed in Chapter 2.1, these highly fragmented genome assemblies result in the overestimated numbers of fragmented genes by splitting the exons of a single gene into different fragmented genomic scaffolds. Furthermore, *Hox* genes, whose expressions are highly limited in the embryonic development, are more vulnerable being not detected when the sufficient clues from embryonic transcriptomes or the sequences of homologous genes from closely related species cannot be provided. The presence of all 10 *Hox* genes from the *Portunus trituberculatus* and *Chionoecetes opilio* genomes therefore further validate the quality of their assemblies and also suggest that the decapod ancestor contained all these *Hox* genes as more ancient, arthropod ancestor did.

In contrast to two brachyuran crab genomes in this study, there were only 9 recognizable *Hox* genes present from the *Nymphon striatum* genome. The *abdominal-A*

gene was lost from its genome, which accords well with preceeding studies on pycnogonid *Hox* gene expressions (Manuel et al., 2006; Pace et al., 2016). These studies suggested that the parallel cases of *abdominal-A* lost in barnacles and chelicerates (pycnogonids and mites) correlated with their morphologies of highly reduced abdomen. In addition, some *Hox* genes from *Chionoecetes opilio* and *Nymphon striatum* showed unique characteristics from their genomic arrangement. In *Chionoecetes opilio* genome, 5 *Hox* genes (*deformed*, *sex combs reduced*, *Antennapedia*, *Ultrabithorax*, and *abdominal-A*) are located in the minus strand of a single genomic sequence, scaffold5763. However, *abdominal-A* is translocated between *sex combs reduced* and *Antennapedia*, which disturbs the genomic collinearity well known in hexapods. Furthermore, rest of 5 *Hox* genes are scattered into different scaffolds (*labial* and *proboscipedia*: scaffold3496, *Hox3*: scaffold18071, and *fushi tarazu*: scaffold25914), which is uncommon case among arthropod, while these atomized pattern of *Hox* genes are reported from mollusk species (Albertin et al., 2015; Kwak, 2017).

More surprisingly, Hox genes of *Nymphon striatum* are also atomized (*labial*: scaffold409, *proboscipedia*: scaffold386, *Hox3*: scaffold434, *deformed*: scaffold170, *sex combs reduced* and *fushi tarazu*: scaffold379, *Antennapedia*: scaffold973, *Ultrabithorax* and *Abdominal-B*: scaffold229) despite of its far longer genomic scaffolds compared to those of *Chionoecetes opilio*. Furthermore, putative duplicated *Hox* genes were detected from both of *Chionoecetes opilio* and *Nymphon striatum* genomes. A partially duplicated *Ultrabithorax* was present between *Antennapedia* and orthologous *Ultrabithorax* in *Chionoecetes opilio* genome. On the other hand, fragments of *labial* (scaffold98) and *proboscipedia* (scaffold20) were found in *Nymphon striatum* genome. Lastly, the sequence of *Nymphon striatum Abdominal-B* was greatly truncated when it was aligned

with those of other chelicerates. This finding also accords with the previous diagnosis of pycnogonid *Abdominal-B* (Manuel et al., 2006) which further supports the correlation suggested by them between pycnogonid reduced abdomen and its truncated posterior *Hox* genes.

Nevertheless, to test the hypothesis on the abdomen reduction of decapods and pycnogonids in the context of *Hox* genes evolution, further studies are required. Firstly, *Hox* genes from various taxa with reduced abdomens and their close relatives with elongated abdomens must be compared each other comprehensively. In addition, Architeuthis dux, a cephalopod, whose genome showed its all core *Hox* genes located on a single genomic scaffold with especially long intervals, which opposed to the previous cases of atomized *Hox* genes in mollusk genomes (da Fonsca et al., 2020). This further stress the importance of increase of genomic contiguity up to sub-chromosomal level, therefore the genome assemblies of *Chionoecetes opilio*, *Nymphon striatum*, and *Portunus trituberculatus* require improved scaffolding analyses. Lastly, the *Hox* genes from three arthropod genome assemblies in this study need to be further curated manually. Since these *Hox* genes were predicted with the automated Seqping pipeline alone, it is possible to some of them are erroneously predicted. Therefore, manual curations such as comparing predicted transcript structures, transcriptomic clues and these genes must be conducted appropriately. These further studies are currently in progress, their completed results could not be included in this dissertation.

2.3.2. The optimizied workflow of *de novo* whole-genome researches of marine arthropods

The final draft genome of both *C. opilio* and *N. striatum* were greatly improved when they were compared to those of *L. tanakae* and mostly fragmented *P. trituberculatus* genomes. Nevertheless another limitations and further efforts to improve the assembly quality were also found in this study, for instance, the *C. opilio* genome with less genomic contiguity compared to the *N. striatum* genome. The *C. opilio* final assembly showed generally less scaffold N50 value and the efficiency of scaffolding process when it was compared to that of *N. striatum* final assembly (**Table 18A**, **Table 19A**, **Figure 23A**, and **Figure 27A**). This resulted insufficient contiguity of *C. opilio* assembly has not been expected at the genome survey stages of *de novo* genome researches of these two marine arthropods. While the specimen for *C. opilio* was much larger, single individual, the specimens of *N. striatum* were multiple wildtype individuals which resulted in much higher genomic heterozygosity estimation than that of *C. opilio* (~1.47% vs ~1.9%). The initial HGAP4 assembly of *N. striatum* required collapsing haplotypic contigs in order to reduce excessive redundancy which indicated that almost entire bodies of haplotypic contigs were not phased into the their main, homozygotic contigs. This was further supported by the pseudo-tetraploid status with almost the same coverages of 1X and 2X peaks (**Figure 21A**) and the content of duplicated BUSCO genes of 53.47% (**Figure 22A**). On the other hand, the *C. opilio* genome did not show the excessive redundancy pattern observed in *N. striatum* genome, as indicated with its Wtdbg2 assembly with less than 2% of duplicated BUSCO gene ratio (**Figure 26C**).

Thus, the reason why *C. opilio* genomic assembly was resulted in much lower genomic contiguity than that of *N. striatum* assembly needs to be discussed with other factors, such

as the degradation of genomic DNA extracts, or their genomic size differences. The main obstacle which hindered assembling *N. striatum* genome in this study was its tiny organismal size resulting less than 1μg of DNA extract per individual, therefore the sampling of *N. striatum* were conducted multiple times to gather sufficient number of individuals in the sample population (40 individuals, ultimately). However, for *C. opilio*, the major limitation was the repeated failures of quality control of genomic DNA extracts sufficient for constructing PacBio and long insert-sized (8kb, 10kb) mate pair libraries. While the PacBio sequenced subreads of *N. striatum* showed 15,480bp average length and 20,750bp N50 vaule, those of *C. opilio* were much lower, 8,556bp and 13,535bp (**Table 20**). In addition, the coverage depths of long insert-sized (8kb and 10kb) mate pair sequenced reads were significantly different between these two species, 75.53 folds coverage for *N. striatum* and 38.68 folds coverage for *C. opilio*. These diffence between mate pair sequence reads contrasts to those of PacBio subreads, which were generated with similar genomic coverage values between two species (*N. striatum*: 113.90 folds, *C. opilio*: 100.53 folds).

Table 20. The compared statistics of *de novo* sequenced long reads in this study

Species	Library type	Insert-size (bp)	Total subreads bases (bp)	No. of subreads	Subread N50 (bp)	Average length (bp)
N. striatum	PacBio SMRT	20,000	84,833,283,304	5,480,059	20,750	15,480
	Illumina	8,000	17,360,189,161	171,883,061	N/A	N/A
	mate pair	10,000	38,893,293,312	385,082,112	N/A	N/A
C. opilio	PacBio SMRT	20,000	201,361,187,452	23,504,401	13,535	8,556
	Illumina	8,000	28,181,064,061	230,823,386	N/A	N/A
	mate pair	10,000	49,285,131,114	375,149,202	N/A	N/A

Considering multiple failures occured during the 8kb, 10kb insert-sized mate pair library construction for *C. opilio*, therefore it can be inferred that the quality of the DNA extracts may affected negatively both of the PacBio subread lengths and the coverages of mate pair sequences with 8kb, 10kb insert-sizes. Although the *C. opilio* tissues were carefully treated as I discussed its necessity in the Chapter 1.2, another obastacles were identified during these processes. Initially, as for *L. tanakae* in Chapter 1.1, approximately 2μg of tissues from each of four *C. opilio* organs subjected to the liquid nitrogen homogenization. However, these frozen and homogenized tissues were extremely slimy so that their tissue samples were not actually powdered, but clotted with each other. The DNA extracts from these liquid nitrogen homogenized *C. opilio* tissues were validated as containing large contents of fragmented DNAs indicated by the smear pattern of the electrophoresis results (**Figure 29A**). On the other hand, when these tissues were buffered into RNALater reagent rather than homogenized with the liquid nitrogen, the resulted DNA extracts were passed the verficiation with minimized DNA degradation (**Figure 29B**). Although the reagent buffered DNA extracts showed improved verification results, the higher GC contents of *C. opilio de novo* sequenced reads (41~48%) than those of *N. striatum* (35~36%) (**Table 13** and **Table 14**) was also inferred that negative factor for the sequencing read-length (Shin et al., 2013).

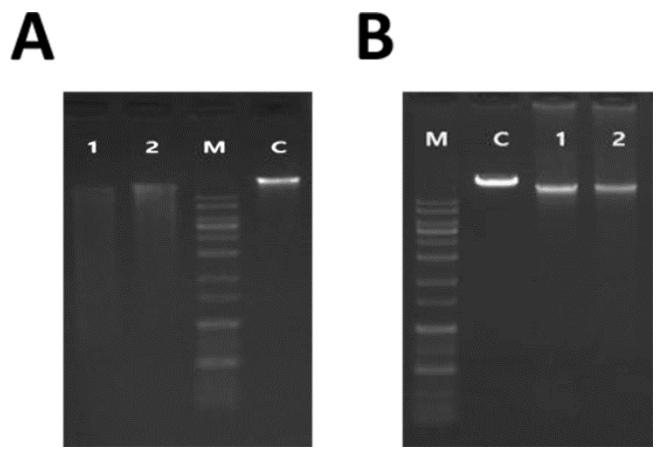

Figure 29. The agarose gel electrophoresis validations of the DNA extracts from *C. opilio* tissues; the lane number 1 indicating the muscular DNA extracts, the lane number 2 indicating the digestive glandar DNA extracts. In this figure, other two types of tissues were omitted. (**A**). The DNA extracts from the liquid nitrogen-homogenized tissues, (**B**). The DNA extracts from the RNALater reagent-buffered tissues

The transcriptomic *de novo* sequenced reads of *N. striatum* also showed high GC content, but even higher (52.06%) than those obeserved from *C. opilio*, which was abnormally high for animal transcriptome (**Table 14**). Considering that RNA molecule is much more vulnerable than DNA molecule is, the main cause of the bacterial contamination of *de novo N. striatum* transcriptome can be inferred as pooling step for compensating the small expected nucleic acid extract per an individual. Therefore, these limititaions observed form this study strongly suggest that the necessity of more sophisticated nucleic acid preparation for *de novo* genome and transcriptome sequencing.

There are possible solutions for overcoming these limitiation as follows. To obtain minimally degraded genomic DNA extract from polysaccharide or secondary metabolite rich, mucuous tissues as in *C. opilio* in this study, the CTAB (cetyltrimethylammonium bromide) containing detergent needs to be applied to the nucleic acid extraction buffers (Arseneau et al., 2016; Kono and Arakawa, 2019; Lienhard and Schäffer, 2019; Chakraborty et al., 2020). The minute expected quantity of genomic DNA extract, such as

in *N. striatum* in this study, can be compensated with the Oxford Nanopore Sequencing technologies. Especially, the MinION (Oxford Nanopore Technologies, Oxford, UK) is specialized to generate more than 10 to 20X coverage of long reads from less than a 1μg of high-molecular genomic DNA extract, without amplification process (Lu et al., 2016; Jain et al., 2018; Joshua and Loman, 2019). In addition, the agarose molded plug nuclei isolation method was reported to yield extremely high molecular weight genomic DNA, and suggested as the most optimized DNA extraction protocol for the Oxford Nanopore sequencing (Brown and Coleman, 2018; Joshua and Loman, 2019).

With considering that minimizing genomic DNA degradation greatly improve the read-length yields (Kono and Arakawa, 2019; Joshua and Loman, 2019), it is concluded that future *de novo* genome research for marine arthropods needs to be conducted using Oxford Nanopore sequencing with high molecular genomic DNA extracted by CTAB buffering or the agarose molded plug nuclei isolation methods. In addition, these more sophisticated DNA extraction protocols also can be applied to near-chromosomal genome scaffolding methods, such as by Bionano genome scaffolding (Bionano Genomics, CA, USA), 10X Genomics sequencing (10X Genomics, CA, USA), and Illumina Hi-C sequencing (Illumina, CA, USA) which resulted in highly completed arthropod genomes recently published (Wallberg et al., 2019; Tang et al., 2020a; Tang et al., 2020b). As the result of this discussion and the studies in Chapter 2, an optimized workflow for *de novo* genome researches on the non-model marine arthropods in the laboratory without sufficient bioinformatics background is described (**Figure 30**).

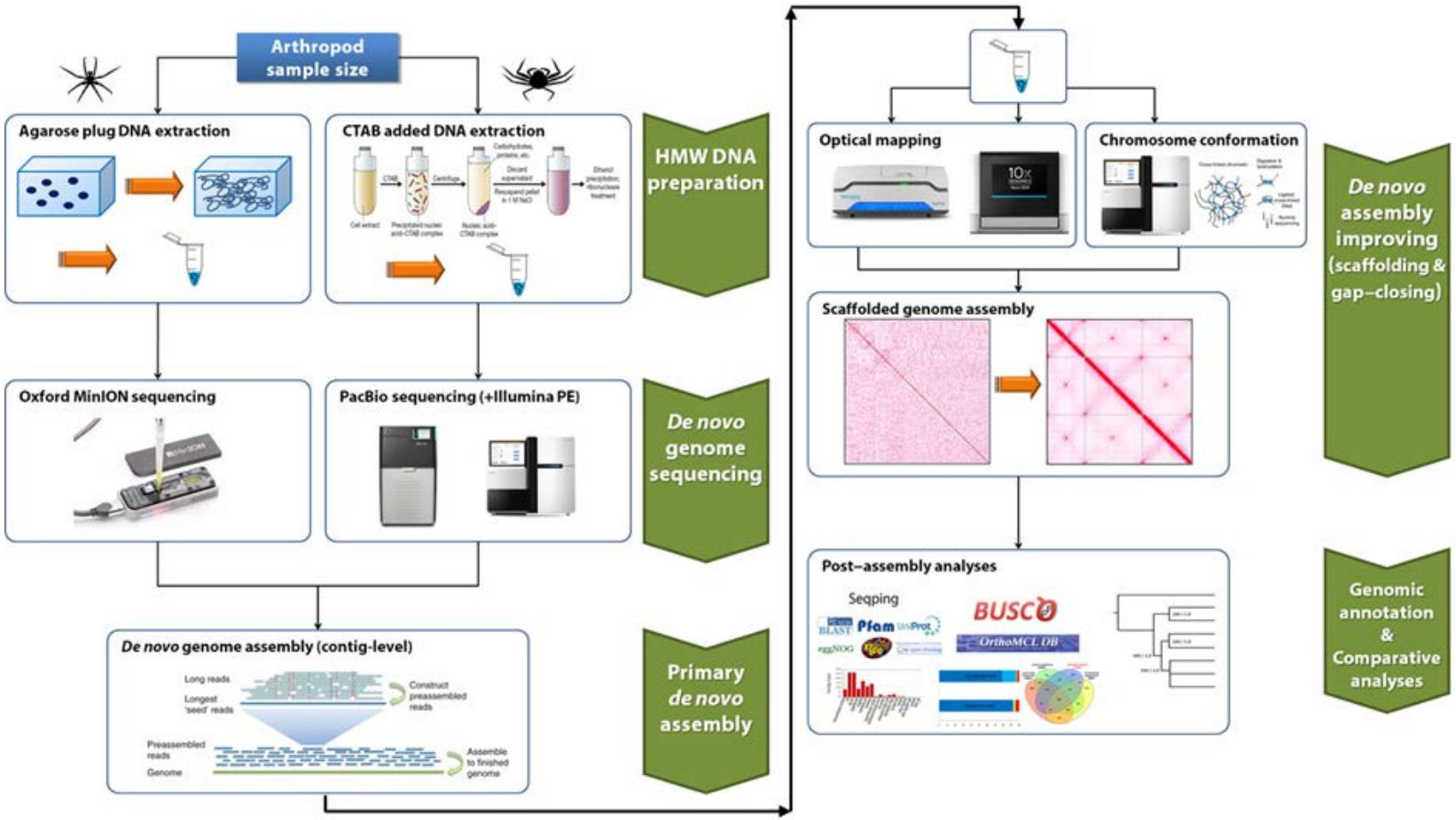

Figure 30. The optimized workflow for *de novo* genome research on non-model marine arthropods which incorporates improved DNA extraction and genomic scaffolding technologies for future studies

Chapter 3. THE CASE STUDY OF THE ARTHROPOD EVOLUTION THROUGH THE COMPARATIVE WHOLE-GENOME ANALYSES

3-1. The preliminary chelicerate phylogenomic analyses incorporating under-sampled taxa

3.1-1. Introduction

Arthropod morphology diversities in the genomic context

The arthropods adapted almost every ecological systems existing in the Earth, thanks to their morphological characteristics of segmented body, paired jointed appendages and moulting, which are frequently used to define them (Minelli et al., 2013; Stork et al., 2018). In addition, interactive studies on their embryogenesis suggested that their segmented nature of bodyplans can be altered with only a slight shift of the *Hox* gene expression patterns which result various combinations of the functions and repeated numbers of each segments and appendages (Carroll, 1995; Averof, 1997; Grenier et al., 1997; Angelini and Kaufman, 2005). The essential role of *Hox* genes in early stages of arthropod development, or morphogenesis, was firstly reported from *Drosophila melanogaster* (Nüsslein-Volhard and Eric, 1980). The deep-homology of the *Hox* genes dated back to the even before the last common ancestor of all bilaterians was reported from the following studies (Carroll et al., 1995; Finnerty and Martindale, 1999; Hueber et al., 2013).

There are 8 core *Hox* genes reported in arthropods with strongly conserved function (Cook et al., 2001; Hughes and Kaufman, 2002), which are *labial* (*lab*), *proboscipedia* (*pb*), *Deformed* (*Dfd*), *Sex combs reduced* (*Scr*), *Antennapedia* (*Antp*), *Ultrabithorax* (*Ubx*), *abdominal-A* (*abd-A*), and *Abdominal-B* (*Abd-B*). These 8 core *Hox* genes regulate the specification of segmental identities of the embryonic segments, and in *Drosophila melanogaster*, they are aligned serially in order of the anterior-most Hox (*lab*) to

posterior-most *Hox* (*Abd-B*) on a single chromosome (Hueber et al., 2010; Hueber et al., 2013). On the other hand, there are two *Hox* genes with less constrainted function, one is *zerknüllt* (*zen*), a Hox3 homologue and the other as *fushi tarazu* (*ftz*). The comparative studies on the functional expressions of these two genes revealed that the chelicerate and myriapods showed conserved ancestral *Hox* gene functions of *zen* and *ftz*, whereas in crustaceans and insects demonstrated more divergent functions such as *ftz* regulating the central neural system development or segmentation (Damen, 2002), and *zen* for patterning anterior-posterior axis and dorsal embryonic structures (Stauber et al., 2002). In addition, these studies showed that the evolution speed of *zen* and *ftz* are more accelerated in crustaceans and insects than in chelicerates and myriapods, as their functional constraints more reduced in crustaceans and insects (Damen, 2002; Stauber et al., 2002).

Arthropod phylogenomics

The Arthropoda is the most diverse animal phylum with more than 1.2 million extant species reported, thus it is recorded as occupying more than 80% of all currently known living animals. Arthropods also demonstrate extremely various body plans throughout their major members (Oakley et al., 2012), with differentiated combinations of segmentation patterns and appendage morphologies (Deutsch and Mouchel-Vielh, 2003; Grimaldi, 2009). Furthermore, the arthropod origin was dated back more than 540 million years, which is well supported by numerous early arthropod fossils discovered from the early Cambrian rocks with enormously diverse morphologies of these fossil taxa that some of them even cannot be classified into extant groups (Marshall, 2006).

These intense arthropod diversity has triggered numerous phylogenetic studies focusing to the relationships of arthropod with other ecdysozoan, as well as those of within its four subphyla; Crustacea, Chelicerata, Hexapoda, and Myriapoda. The earliest studies argued that the Arthropoda is a polyphyletic group (Anderson, 1973; Manton, 1977; Manton and Anderson, 1979; Fryer, 1998). The later studies rejected this concept of arthropod polyphyly which conducted with cladistics approach, molecular marker based analyses, or even increased sampled taxa including newly discovered fossil species (Kristensen, 1975; Cutler, 1980; Wheeler et al., 1993; Chen et al., 1994; Boore et al., 1995). After the monophyly of arthropod was accepted, the focus of phylogenetic arguments has been moved into the hypotheses for relationships between its subphyla. There were arguments between Tracheata hypothesis, which suggested a sister group relationship between hexapod and myriapod (Beall et al., 2000; Haas et al., 2003; Bäcker et al., 2008), and Mandibulata hypothesis, which argued a monophyletic clade consisted with crustacean, hexapod, and myriapod (Boor et al., 1995; Wägele et al., 1995; Shultz et

al., 2000; Cook et al., 2005; Rota-Stabelli and Telford., 2008). The introduction of phylogenomics approach using more than 100 molecular marker genes finally brought the end of these debates. A number of researches with early approaches for phylogenomics strongly supported monophyletic Mandibulata and Pancrustacea and rejected Tracheata hypothesis (Meusemann et al., 2010; Reiger et al., 2010; Rehm et al., 2011).

However, unlike those of between arthropod subphyla, the relationships between the major groups of Chelicerate are still interactively debated. For instance, the placement of Xiphosura has been controversial among recent phylogenomic studies; some studies suggested that Xiphosura is nested in the paraphyletic Arachnida (Sharma et al., 2014; Ballesteros et al., 2019; Ballesteros and Sharma, 2019; Nolan et al., 2020; Nong et al., 2020), while others found a conventionally accepted Xiphosura-Arachnida sister group relationship (Lozano-Fernandez et al., 2016; Lozano-Fernandez et al., 2019). These phylogenomic studies included chelicerate *de novo* sequenced genomes representing almost all major clades, nevetheless, some taxa such as opiliones, pseudoscorpions, and pycnogonids are remained limited in their genomic. Therefore, these studies included *de novo* assembled transcriptomes of these poor sampled taxa to compensate the lack of accessible genomic assemblies. To further improve the credibility and resolution, a phylogenomic study which includes representative *de novo* genome assemblies for all major chelicerate taxa are required (Garb et al., 2018; Giribet, 2018).

Approaches and limitations of this study

The main goal of this study is to provide a case of preliminary phylogenomic analyses using the datasets constructed from the *de novo* assembled arthropod genomes including two under-sampled taxa in preceeding studies, the Class Pycnogonida and the Infraorder Brachyura from Order Decapoda. In order to achieve this main goal, *de novo* genomes of a pycnogonid, *Nymphon striatum* and brachyuran crabs, *Chionoecectes opilio* and *Portunus trituberculatus* aseembled and analyzed at the previous Chapters, were included to construct data matrix for phylogenetic analyses. In addition, the proteomes based on *de novo* genome assmeblies of 11 chelicerate and 5 pancrustacean species were also incorporated in the data matrix analyzed in this study. With this interactively sampled 19 species of arthropod genome based proteomes, major clades were represented with at least two sampled species as following; Chelicerata including Arachnida, Acariformes, Mesostigmata, and Pancrustacea including Hexapoda and Multicrustacea. The details of sampled species which were subjected in this study are further described at **Table 21**. To the current best knowledge, this is the first case of phylogenomic study which incorporates whole-genome scaled proteomes of pycnogonid and brachyuran decapod in its analyzed datamatrix.

The approaches of this study, nevertheless, have limitations of analyzing huge datasets which were constructed from *de novo* arthropod whole-genomes. As discussed in the previous two chapters, the phylogenomic analyses based on the 13 PCGs (protein-coding genes) and much limited number (up to 8 species) of whole-genome based proteomes required significantly long analytic times. Although submitting these analyses on the CIPRES Science Gateway (Miller et al., 2010; Miller et al., 2011) greatly reduced the required time, the analytic approaches in this chapter were found to be abnormally

terminated with the memory overflow and the processes exceeding allowed running times. The further details of these abnormally terminated analyses are decribed in **Table 22** and dicussed in Results and Discussion section. The investigated literatures of phylogenomic studies on the large sized data matrices were perfomed by cooperations with the academic or commercial computing servers, which was unfortunately, not possible in this study.

Therefore, to reduce the required analytic time acceptable at current environment, some methodologies and models were scaled down. First, arthropod species with available *de novo* genomes, but not with proteomes were excluded from the taxa sampling of this study. *De novo* genomes of such arthropods were deposited to the NCBI oftenly without the submission of their genomic annotations by authors since the genomic annotation is not necessary requisite for validating the submitted genomic assemblies at the NCBI. This could be compensated with the homology-based gene prediction and then following manually performed gene curation. The homology-based gene prediction of multiple genomes can be performed with the latest version of AUGUSTUS, which requires more than 20 folds of CPU core days per the summed total length of analyzed genomes (Nachtweide and Stanke, 2019). Unfortunately, the computering resources for this study were heavily limited, as the UNIX system installed in the laboratory with 20 available physical CPU cores and 24GB sized memory. Another operating system for analyses of this study is the CIPRES Scientific Gate web based analytic server, which only strictly provide the bioinformatics softwares for inferring phylogenetic relationships.

The other scaling down was applied to the level of sequence alignments which is analyzed by the phylogenetic inferring softwares of RAxML 8.2.12 HPC (Stamatakis, 2014) and MrBayes 3.2.7 (Ronquist et al., 2012). In this study, the amino acid-level of orthologous genes alignments and substitution models were selected, instead of codon-

level models with matrices of degenerated third codons used in referred researches (Ballesteros et al., 2019; Lozano-Fernandez et al., 2019; Noah et al., 2020). The application of amino acid-level of alignments can greatly reduce the required analytic times for orthologous search and the phylogenetic tree reconstructions as following. In the ortholgous gene search step, BLAST based pairwide comparison for entire genes are the first bottleneck in terms of the speed and time. Apparently, this all-to-all similarity search can be performed much faster and simpler for amin acid-level datasets than for codon-level of nucleotide datasets, since the latter require BLASTX which is known to be slow in speed to analyze large datasets such as proteomes based on the *de novo* genomes (Buchfink et al., 2015). In contrast, BLASTP (Delaney et al., 2000) can be used to perform relatively faster all-to-all similarity searches of proteomes based on the *de novo* genomes, since its algorithm does not include intermediate prediction of nucleotide sequences from the amino acid sequences or translation of nucleotide sequences back to amino acid sequences. In addition, the total length of the supermatrix used for the phylogenetic analyses can be reduced more than 3 folds by using amino acid-level of sequence alignments rather than codon-level of nucelotide alignments.

3.1.2. Materials and Methods

Taxa sampling of representative arthropods with whole-genome based proteomes

The "NCBI Genome List" was investigated (Latest update at 2020.05.03., Retrieved at 2020.05.03.) to sample representative chelicerates and their serial outgroups whose *de novo* whole-genome based proteomes were available for the downstream analyses. At least 2 species were sampled which represent the major chelicerate taxa, except for xiphosuran and pycnogonid whose respective case of available proteome was only one. There were 3 sampled species representing the clade Arachnopulmonata, and 7 species representing 3 superorders of Subclass Acari. The proteomes of *Limulus polyphemus* and *Nymphon striatum* were selected as representatives for xiphosuran and pycnogonida, respectively. As outgroup to Chelicerata, 7 pancrustacean species were included in the sampled taxa, which also included two brachyuran crab proteomes of *Chionoecectes opilio* and *Portunus trituberculatus*. The sampled representative species of the clade Multicrustacea were comprised with 3 decapods and one copepod. There were two sampled species as representative taxa of Hexapoda, and *Daphnia magna*, a water flea, was selected to represent the Class Branchiopoda. Finally, *Caenorhabditis elegans* was selected as an outgroup species against the Phylum Arthropoda. The detailed information about these 19 sampled arthropod taxa and a nematode are indicated in **Table 21**.

Table 21. The summarized information of 20 selected species with *de novo* sequenced whole-genome based proteomes in this study

Species	Group	Representing taxa Level1	Representing taxa Level 2	No. of genes	Data sources
Caenorhabditis elegans	Outgroup (Nematoda)	Outgroup	Outgroup	28,416	Ensembl Metazoa Release 46
Nymphon striatum	Chelicerata	Pycnogonida	Pantopoda	28,539	Current study
Limulus polyphemus	Chelicerata	Merostomata	Xiphosura	38,676	Ensembl Metazoa Release 46
Ixodes scapularis	Chelicerata	Arachnida	Acari-Ixodida	32,572	Ensembl Metazoa Release 46
Varroa destructor	Chelicerata	Arachnida	Acari-Mesostigmata	24,430	Ensembl Metazoa Release 46
Galendromus occidentalis	Chelicerata	Arachnida	Acari-Mesostigmata	11,923	Ensembl Metazoa Release 46
Dermatophagoides pteronyssinus	Chelicerata	Arachnida	Acari-Acariformes	12,824	Ensembl Metazoa Release 46
Sarcoptes scabiei	Chelicerata	Arachnida	Acari-Acariformes	10,473	Ensembl Metazoa Release 46
Tetranychus urticae	Chelicerata	Arachnida	Acari-Acariformes	15,671	Ensembl Metazoa Release 46
Dinothrombium tinctorium	Chelicerata	Arachnida	Acari-Acariformes	19,024	Ensembl Metazoa Release 46
Centruroides sculpturatus	Chelicerata	Arachnida	Arachnopulmonata	35,229	Ensembl Metazoa Release 46
Parasteatoda tepidariorum	Chelicerata	Arachnida	Arachnopulmonata	27,515	Ensembl Metazoa Release 46
Stegodyphus mimosarum	Chelicerata	Arachnida	Arachnopulmonata	27,515	Ensembl Metazoa Release 46
Daphnia magna	Pancrustacea	Branchiopoda	Cladocera	26,646	Ensembl Metazoa Release 46

Table 21. Continued from the previous page

Drosophila melanogaster	Pancrustacea	Hexapoda	Insecta	30,559	Ensembl Metazoa Release 46
Folsomia candida	Pancrustacea	Hexapoda	Entognatha	25,774	Ensembl Metazoa Release 46
Tigriopus californicus	Pancrustacea	Multicrustacea	Copopoda	15,577	Ensembl Metazoa Release 46
Chionoecetes opilio	Pancrustacea	Multicrustacea	Decapoda	22,659	Ensembl Metazoa Release 46
Penaeus vannamei	Pancrustacea	Multicrustacea	Decapoda	33,273	Ensembl Metazoa Release 46
Portunus trituberculatus	Pancrustacea	Multicrustacea	Decapoda	34,536	Ensembl Metazoa Release 46

Construction of phylogenetic data matrix

To obtain universally shared orthologous genes from these 20 selected proteomes, OrthoMCL pipeline was applied. A local BLAST database was constructed from all protein sequences from 20 proteomes, and then BLASTP (Delaney et al., 2000) performed all-to-all sequence similarity search for OrthoMCL pipeline. The OrthoMCL v2.0.9 (Fischer et al., 2011) was used to conduct homology-searching and clustering steps against the BLASTP-resulted all-to-all similarity files with its default parameters. To obtain the data matrix with sufficient numbered universal orthologoues, the orthologous clusters containing at least one paralogous genes were further analyzed instead of non-redundant clusters. Co-orthologous genes from these clusters were verified and then selected according to their similarity support values with increased weight for the length of genes, in order to minimize the content of fragmentary genes included in these clusters.

Finally, one orthologous cluster containing synthetic proteins was filtered out to finalize 1,189 clustered ortholgous genes shared universally in 20 studied species.

The amino acid sequences of each orthologous genes were merged into a multiple sequenced FASTA formatted file which contains 20 genes from each species in the same order. These 1,189 FASTA files for each orthologous genes were aligned pairwisely using MAFFT (Katoh et al., 2017) with its operating options as following (--maxiterate 1000 -geneafpair). The MAFFT-aligned 1,189 aligned FASTA files underwent trimming out of bad qualitied alignments or gaps using trimAl (Capella-Gutiérrez et al., 2009) with its following parameters (-gappyout –automated1). These trimmed aligned FASTA files were concatenated using BeforePhylo (https://github.com/qiyunzhu/BeforePhylo) to obtain the finalized supermatrix for phylogenetic reconstruction analyses.

Phylogenetic reconstruction analyses of genome based data matrix

In order to convert large sized supermatrix into PHYLIP format (for RAxML) and NEXUS format (for MrBayes) precisely, trimAl was used (-phylip for PHYLIP formatted file and –nexus for NEXUS formatted file). To reconstruct the phylogenetic relationships of 19 arthropods and a nematode, RAxML 8.2.12 HPC (Stamatakis, 2014) and MrBayes 3.2.7 (Ronquist et al., 2012) were applied both on the UNIX system of the laboratory and the CIPRES Science Gateway (Miller et al., 2010; Miller et al., 2011). *Caenorhabditis elegans* was used as the outgroup species for both analyses. The maximum-likelihood approach of phylogenetic analysis was performed by RAxML 8.2.12 HPC with its amino acid substitution model of Gamma distribution of variable sites, estimation of the most probable substitution ratios from unconstrainted initial matrix (-m PROTGAMMAAUTO). In addition, the default random seed value was provided to infer

parsimony during analysis (-p 12345) and the accelerated fast bootstrap analysis with 1,000 replications (-f a -N 1000 -x 12345). The Bayesian inferred phylogenetic reconstruction was conducted using MrBayes 3.2.7 with its operating parameters following; the Gamma distributed substitution model for variable sites, (lset coding=variable Nucmodel=protein Rates=invgamma Covarion=Yes) and the analytical replications by the Markov chain Monte Carlo (MCMC) methods, (mcmcp ngen=1000000 samplefreq=1000). The initial 25% of replicated resulted were discarded as burn-in by using (burnin=10 relburnin=Yes burinfrac=0.25) options. To reduce analytical time by maximizing parallelized analysis, multi-threading was activated on the laboratory UNIX system with following commands (-T 16 for RAxML, and --use-hwthread-cpus -np 16 mb, nchains=4 nruns=4 for MrBayes). On the other hand, the maximum parallelized threads allowed for RAxML (-T 24) and MrBayes (24 threads, nchains=4 nruns=6) were submitted. In addition, the maximum allowed analytic hours for RAxML (48hours) and MrBayes (168hours) on the CIPRES Science Gateway were applied.

3.1.3. Results and Discussion

The phylogenetic analyses of Subphylum Chelicerata based on the curated whole-genome datasets

To reconstruct phylogenetic relationships of 19 selected arthropods with an outgroup taxon, *C. elegans*, three replicative analyses per each phylogenetic programs were performed both in the laboratory UNIX system and the CIPRES Scientific Gate service. These runs could not be executed parallelly, since their CPU and memory occupancy were estimated as almost reaching the maximally allowed limits for both operating systems. As the result of these analyses, unfortunately, almost every runs were abnormally terminated except only one RAxML analytical run was successfully ended (**Table 22**).

Although there was only one successfully completed analysis, the RAxML resulted on the CIPRES Scientific Gate, its consensus tree clearly supported monophyletic status of all its inferred clades (**Figure 31**). The Phylum Arthropoda was resolved but not subjected into the bootstrapping procedure due to the constraint parameter between *C. elegans* and ingroup, or Arthropoda. On the other hand, the consensus tree topology supported the monophyl of following nodes; Clade Pancrustacea, Clade Altocrustacea, Subphylum Chelicerata, Class Arachnida , Superfamily Acariformes, and Clade Arachnopulmonata including a xiphosuran, *Limulus polyphemus* (**Figure 31**). In addition, a sister group relationship between two mites superfamilies, the Ixodida and Merostigmata, was supported with the maximum bootstrap support value.

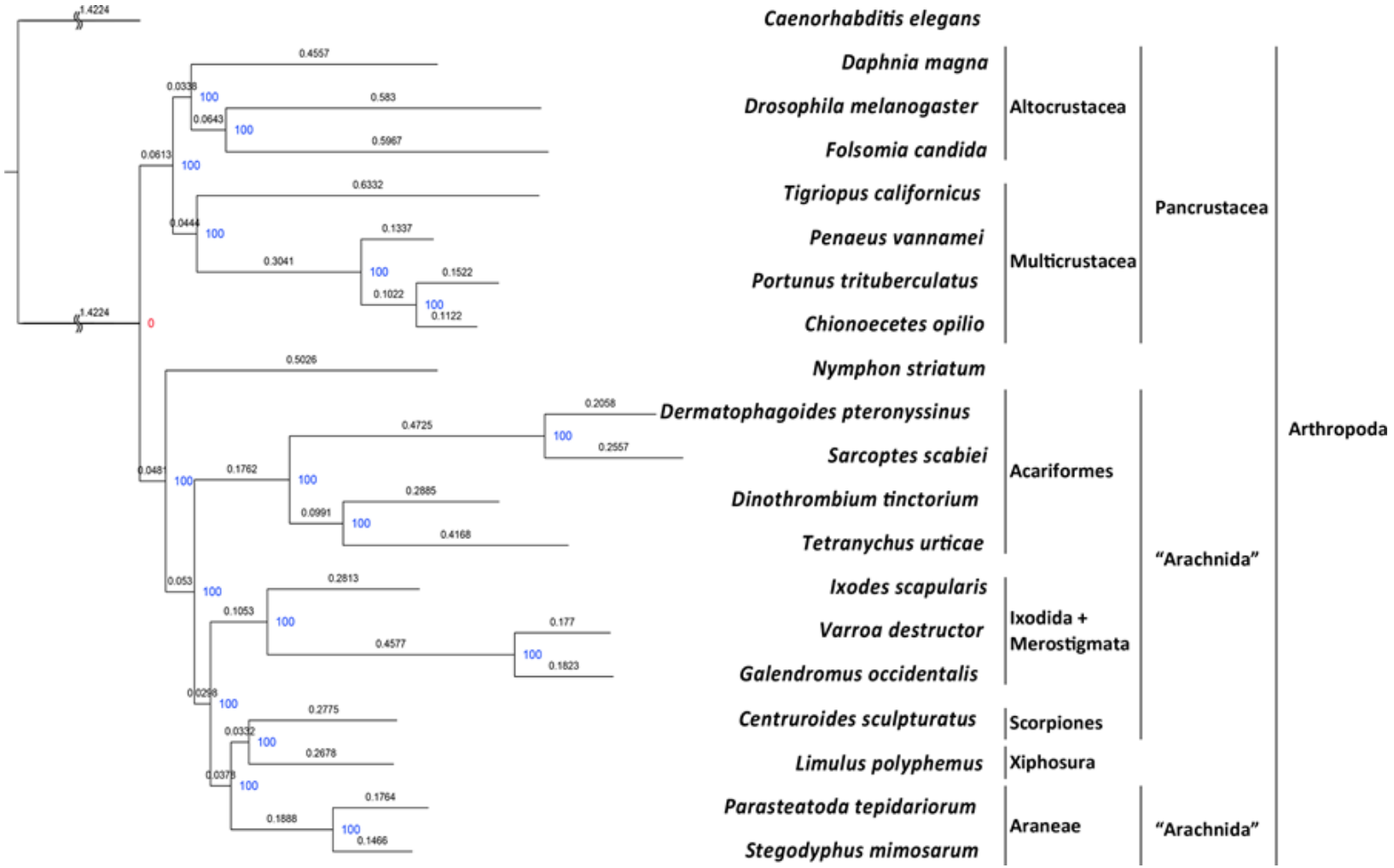

Figure 31. The most probable consensus tree reconstructed from the supermatrix with 1,189 ortholgous genes from the 20 species in this study

Surprisingly, *Limulus polyphemus* was placed in the monophyletic Arachnopulmonata as the sister taxon of a scorpion, *Centruroides sculpturatus*, both with 100% bootstrap support values. This result accords with the two most recently conducted chelicerate phylogenomic researches (Nolan et al., 2020; Nong et al., 2020) which also strongly supported the monophyly of Arachnopulmonata with Xiposura nested in it. In addition, the Acariformes and the monophyletic group composed of the Ixodida and Mesostigmata (**Figure 31**) were recovered as the serial outgroups against the monophyletic Arachnopulmonata, which further rejected the monophyly of Acari. This result was also well accorded recent phylogenomic researches on chelicerate which also argued the polyphyletic status of Acari (Pepato and Klimov, 2015; Van Dam et al., 2018; Li et al., 2019; Lozano-Fernandez et al., 2019). This study, therefore, provide the first result from the whole-genome based data matrix, which supports the paraphyly of Arachnida and Xiphosura nested in the monophyletic Arachnopulmonata, a clade with most derived arachnids having book lungs as their repiratory systems. This finding also implies the plausibility of the hypothesis suggesting Xiphosura as a secondary marine arachnid, and two independent land invasions of arachnids, which was suggested by recent research incorporating both molecular phylogenomic and fossil record datasets (Noah et al., 2020).

Despite of the very stable tree topology observed from the consensus tree in this study, there were major limitations caused by abnormal terminations of rest of 11 analytical trials. In detail, the first trial of RAxML analysis on the laboratory UNIX system was ended with memory overflow error, thus any unnecessary processes were

terminated before starting the second trial. The second, and following last RAxML analysis were however, terminated due to a process error of which cause was not clearly designated by the UNIX system. On the other hand, both of the first and second RAxML analyses on the CIPRES Scientific Gate were terminated due to exceeding maximum time allowed in the system (**Table 22A**). These five terminated RAxML analyses produced incomplete bootstrapping result files which could not be subjected for consensusing the most probable tree, since RAxML could not understand the incompletely terminated boostrapping files. In the cases of MrBayes analyses, all 3 replicated runs in the UNIX system and the first replicate in the CIPRES Scientific Gate were abnormally terminated due to undefined process error. Furthermore, the MrBayes analyses on the UNIX system could not finish its very first process of reading NEXUS file inputs, without any MCMC (Markov chain Monte Carlo) replications conducted (**Table 22B**). On the other hand, the second MrBayes run on the CIPRES Scientific Gate performed 5,000 MCMC replications, nevertheless it reached 168 hours of maximally allowed in the system. The last CIPRES submitted MrBayes analysis therefore underwent the reduction of required parallel chains from 24 (nchains=4 nruns=6) into 8 (nchains=2 nruns=4). Even its analytic parameters were arranged for reducing required time, unfortunately, the final analysis was terminated incompletely by reaching the limited 168 hours, but with slight increase of its conducted MCMC replication as 48,000.

Table 22. The statistics of each 3 copies of trial of phylogenetic analyses using RAxML and MrBayes on the laboratory UNIX system and the CIPRES Scientific Gate

A. RAxML analysis

Operating System	Computering resources	Analysis hours	Replications done	Job status
UNIX, laboratory	Available CPUs: 16/20 Available RAM: 24GB	0.5	0/1,000	Terminated due to memory overflow
UNIX, laboratory	Available CPUs: 16/20 Available RAM: 24GB	116.2	439/1,000	Terminated due to an process error
UNIX, laboratory	Available CPUs: 16/20 Available RAM: 24GB	131.7	481/1,000	Terminated due to an process error
CIPRES	Available CPUs: 24 Available RAM: >20GB	48.0	492/1,000	Terminated due to exceed allowed time
CIPRES	Available CPUs: 24 Available RAM: >20GB	48.0	533/1,000	Terminated due to exceed allowed time
CIPRES	Available CPUs: 24 Available RAM: >20GB	43.1	1,000/1,000	Completed successfully

B. MrBayes analysis

Operating System	Computering resources	Analysis hours	Replications done	Job status
UNIX, laboratory	Available CPUs: 16/20 Available RAM: 24GB	2.5	0/1,000,000	Terminated due to an process error
UNIX, laboratory	Available CPUs: 16/20 Available RAM: 24GB	2.3	0/1,000,000	Terminated due to an process error
UNIX, laboratory	Available CPUs: 16/20 Available RAM: 24GB	2.6	0/1,000,000	Terminated due to an process error
CIPRES	Available CPUs: 24 Available RAM: >20GB	1.7	0/1,000,000	Terminated due to an process error
CIPRES	Available CPUs: 24 Available RAM: >20GB	168	5,000 /1,000,000	Terminated due to exceed allowed time
CIPRES	Available CPUs: 8* Available RAM: >20GB	168	48,000 /1,000,000	Terminated due to exceed allowed time

Therefore, further analyses are necessary to be followed in order to reinforce the insight provided from this study. An accessible cloud system providing clustered computering is the most significant requisite in order to stably conduct RAxML and MrBayes driven phylogenetic analyses, with even larger data matrices containing more number of sampled taxa. Second, it is also necessary to incorporat arthropod species with availabble *de novo* genomes, but not proteomes as sampled taxa for future studies. This can be accomplished with the homology-based gene prediction and following manual curation processes, but as in enlarged, whole-genome scale of analyses. The increase of sampled species representing under-sampled taxa in this study such as Order Scorpiones and Xiphosura will contribute the improvement of phylogenetic resolution between these taxa greatly. Lastly, future phylogenetic inference analyses need to be performed in the codon-scale nucleotides, with incorporating the substitution matrices of first two codon sites and degenerated third codon site, instead of in amino acid sequences.

CONCLUSION

CONCLUSION

This study has described five cases of *de novo* assemblies, their genomic annotations and characteristics which were newly conducted. In addition, this study has discussed three levels of bottlenecks in *de novo* whole genome analysis for non-model arthropod species and the possible solutions for these bottlenecks. This study also has described the genomic level of phylogenetic analyses focusing on the Subphylum Chelicerata with inclusion of the first case of pycnogonid representative with *de novo* genome based data.

The first chapter of this dissertation describes the assembly quality, genomic annotations and features of a marine fish, *Liparis tanakae*. In addition, *L. tanakae de novo* genome containing 35 copies of various collagen genes provided putative genomic contexted explanation of its mucous rich skin and muscle tissues. The first chapter also describes the unique characteristics of *de novo* assembled *Chionoecetes opilio* mitochondrial genome and the phylogenetic relationship of 12 decapod species using 13 protein-coding genes.

The Chapter 2 of this dissertation provides the genomic characteristics and discussions of three *de novo* assembled marine arthropod genomes. The *Portunus trituberculatus* genome was *de novo* assembled basically as the same workflow with *L. tanakae* described in the previous chapter, significantly low assembly quality was yielded from the *P. trituberculatus*. The difference of reproductive ecology and genomic complexity were discussed as the probable causes of much lower genomic contiguity and completeness of *P. trituberculatus* genome. On the other hand, the high coverages of *PacBio* de novo genome sequencing has enabled the high-qualitied genome assemblies of two marine arthropods as described in the second chapter of this dissertation. The *Nymphon striatum* and *Chionoecetes opilio* genomic assemblies showed greatly improved

genomic contiquity and completeness which were indicated with their scaffold N50 values and contents of complete BUSCO genes exceeding 100Kb and 90%, as described in Chapter 2-2. Their enhanced assembly qualities also enabled more informative downstream analyses of orthologous gene search and accurate phylogenetic tree reconstructions. In Chapter 2-3 of this dissertation, the discussions on the *Hox* genes characteristics and potential methodological improvements for future studies were developed. The annotated *Hox* genes of *C. opilio*, *N. striatum*, and *P. trituberculatus* suggested that *Hox* gene loss in arthropods is rare, and both of *C. opilio* and *N. striatum* *Hox* genes possibly underwent significant genomic rearrangement. Two improved DNA extraction protocols and Oxford MinION sequencing are discussed as the possible solution for the first bottleneck of obtaining sufficient amounts of genomic DNA with minimized degradation. To overcome the second bottleneck of effective scaffolding for highly heterozygous marine arthropod genome, Chapter 2-3 also discussed possible applications of superscaffolding by BioNano, 10X genomics, and Illumina Hi-C sequencing technologies. These discussions on optimizing *de novo* genome researches on marine arthropods were finally developed as the suggested workflow in Chapter 2-3.

The final chapter describes the preliminary chelicerate phylogenomics study which incorporated under-sampled pycnogonid and brachyuran *de novo* genome based data for the first time. Although the study of the final chapter requires further improvements in the number of sampled taxa and the substitution models used for analyses, methodologies and analyses were practiced with the best of efforts in currently available computering resoruces. The resulted consensus tree also strongly supported the recent hypothesis of Xiphosura nested with in the Arachnopulmonata instead of traditionally accepted sister group relationship between Xiphosura and Arachnida.

REFERENCES

Literatures cited

Abzhanov, A. and Kaufman, T. C. (2000). Embryonic expression patterns of the Hox genes of the crayfish *Procambarus clarkii* (Crustacea, Decapoda). *Evolution & development*, 2:5, 271-283.

Akam, M., Averof, M., Castelli-Gair, J., … and Ferrier, D. E. K. (1994). The evolving role of Hox genes in arthropods. *Development*, 1994(Supplement), 209-215.

Albrecht, G. T., Valentin, A. E., Hundertmark, K. J. and Hardy, S. M. (2014). PANMIXIA IN ALASKAN POPULATIONS OF THE SNOWCRAB *CHIONOECETES OPILIO* (MALACOSTRACA: DECAPODA) IN THE BERING, CHUKCHI, AND BEAUFORT SEAS. *Journal of Crustacean Biology*, 34:1. 31-39.

Albertin, C. B., Simakov, O., Mitros, T., … and Rokhsar, D. S. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, 524:7564, 220-224.

Alkan, C., Sajjadian, S. and Eichler,. E. E. (2011). Limitations of next-generation genome sequence assembly, *Nature Methods*, 8:1, 61-65.

Alvsvåg, J., Agnalt, A-L. and Jørstad, K. E. (2009). Evidence for a permanent establishment of the snow crab (*Chionoecetes opilio*) in the Barents Sea. *Biological invasions*, 11:3, 587-595.

Andrew, D.R. (2011). A new view of insect-crustacean relationships II. Inferences from expressed sequence tags and comparisons with neural cladistics. *Arthropod Structure and Development*, 40:3, 289-302.

Arseneau, J-R., Steeves, R. and Laflamme, M. (2016). Modified low-salt CTAB extraction of high-quality DNA from contaminant-rich tissues, *Molecular Ecology Resources*, 17, 686-693.

Azuma, N., Grant, W. S., Templin, W. D., … and Abe, S. (2011). Molecular Phylogeny of a Red-Snow-Crab Species Complex using Mitochondrial and Nuclear DNA Markers. *Zoological Science*, 28. 286-292.

Baldwin-Brown, J.G., Weeks, S.C., and Long, A. D. (2017). A new standard for crustacean genomes: the highly contiguous, annotated genome assembly of the clam shrimp *Eulimnadia texana* reveals *HOX* gene order and identifies the sex chromosome. *Genome biology and evolution*, 10:1, 143-156.

Ballesteros, J. A., López, C. E. S., Kováč, L., … and Sharma, P. P. (2019). Ordered phylogenomic subsampling enables diagnosis of systematic errors in the placement of the enigmatic arachnid order Palpigradi, *Proceedings of the Royal Society B*, 286:1917, 1-9.

Ballesteros, J. A., and Sharma, P. P. (2019). A critical appraisal of the placement of Xiphosura (Chelicerata) with account of known sources of phylogenetic error. *Systematic Biology*, 68:6, 896-917.

Barreto, F.S., Watson, E.T., Lima, T.G., … and Burton, R.S. (2018). Genomic signatures of mitonuclear coevolution across populations of *Tigriopus californicus*. *Nature ecology and evolution*, 2:8, 1250.

Basso, A., Babbucci, M., Pauletto, M., … and Negrisolo, E. (2017). The highly rearranged mitochondrial genomes of the crabs *Maja crispata* and *Maja squinado* (Majidae) and gene order evolution in Brachyura. *Scientific Reports*, 7:4096, 1-17.

Bernt, M., Donath, A., Jühling, F., … Stadler, P. F. (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, 69:2, 313-319.

Bitencourt, J.V.T., Roratto, P.A., Bartholomei-Santos, M.L., and Santos, S. (2007). Comparison of different methodologies for DNA extraction from *Aegla longirostri*. *Brazilian Archives of Biology and Technology*, 50:6, 989-994.

Boetzer, M., Henkel, C. V., Jansen, H. J., … and Pirovano, W. (2010). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 27, 578-579.

Boeuf, G. (2011). Marine biodiversity characteristics. *Comptes rendus biologies*, 334:5-6, 435-440.

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

Bradnam, K.R., Fass, J.N., Alexandrov, A., ... and Chitsaz, H. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2:1, 10.

Brown, S. J. and Coleman, M. (2018). Isolation of High Molecular Weight DNA from Insects, *In: Brown, S. J. (eds) Insect Genomics: Methods and Protocols*, Methods in Molecuclar Biology, vol. 1858, Humana Express, New York, NY, 27-32.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND, *Nature Methods*, 12(1), 59-63.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25:15, 1972-1973.

Chan, K. L., Rosli, R., Tatarinova, T. V., … and Low, E. T. L. (2017). Seqping: gene prediction pipeline for plant genomes using self-training gene models and transcriptomic data. *BMC bioinformatics*, 18: 29.

Chakraborty, S., Saha, A. and Ananthram, A. N. (2020). Comparison of DNA extraction methods for non-mairne molluscs: is modified CTAB DNA extraction method more efficient than DNA extraction kits?, *3 Biotech*, 10:69.1-6.

Chebby, M. A., Becking, T., Moumen, B., … and Cordaux, R. (2019). The Genome of Armadillidium vulgare (Crustacea, Isopoda) Provides Insights into Sex Chromosome Evolution in the Context of Cytoplasmic Sex Determination, *Molecular Biology and Evolution*, 36:4, 727-741.

Chen, D., Liu, Q., Zeng, X., and Su, Z. (1997). Catch composition and seasonal variation of setnet fisheries in the Yellow and Bohai Seas. *Fisheries Research*, 32:1, 61-68.

Chen, C-Y., Wu, K-M., Chang, Y-C., … and Tsai, S-F. (2003). Comparative Genome Analysis of *Vibrio vulnificus*, a Marine Pathogen, *Genome Research*, 13, 2577-2587.

Chernova, N. V., and Stein, D. L. (2004). A remarkable new species of *Psednos* (Teleostei: Liparidae) from the western North Atlantic Ocean. *Fishery Bulletin*, 102:2, 245-250.

Chernova, N. V. (2005). New species of Careproctus Liparidae from the Barents Sea and adjacent waters. *Journal of Ichthyology*, 45:9, 689-699.

Chin, C. S., Alexander, D. H., Marks, P., … and Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10:6, 563-569.

Chung, J. S., Ahn, I. S. and Kim, D. S., (2015). Crustacean hyperglycemic hormones of two cold water crab species, *Chionoecetes opilio* and *C. japonicus*: Isolation of cDNA sequences and localization of CHH neuropeptide in eyestalk ganglia. *General and Comparative Endocrinology*, 214:2015, 177-185.

Chyung, M. K. (1977). *The fishes of Korea*. Seoul: Ilji-sa, 1-727.

Cook, C. E., Akam, M., Smith, M. L., … and Bastianello, A. (2001). Hox genes and the phylogeny of the arthropods, *Current Biology*, 11:10, 759-763.

da Fonseca, R. R., Couto, A., Machado, A. M. … and Gilbert, M. T. P. (2020). A draft genome sequence of the elusive giant squid, *Architeuthis dux*. *GigaScience*, 9:1, giz152.

De Grave, S., Pentcheff, N. D., Ahyong, S. T., … and Wetzer, R. (2009). A CLASSFICATION OF LIVING AND FOSSIL GENERA OF DECAPOD CRUSTACEANS. *Raffles Bulletin of Zoology*, 1-109.

Demian, W. L. L., Jahouh, F., Stansbury, D., … and Banoub, J. H. (2014). Characterizing changes in snow crab (*Chionoecetes opilio*) cryptocyanin protein during molting using matrix-assisted laser desorption/ionization mass spectrometry and tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 28, 355-369.

Deutsch, J.S., and Mouchel-Vielh, E. (2003). *Hox* genes and the crustacean body plan. *BioEssays*, 25:9, 878-887.

Domingues, C.P., Creer, S., Taylor, M.I., ... and Carvalho, G.R. (2010). Genetic structure of *Carcinus maenas* within its native range: larval dispersal and oceanographic variability. *Marine Ecology Progress Series*, 410, 111-123.

Dunlop, J. A. and Selden, P. A. (2009). Calibrating the chelicerate clock: a paleontological reply to Jeyaprakash and Hoy. E*xperimental and Applied Acarology*, 48:3, 183.

English, A. C., Richards, S., Han, Y., … and Gibbs, R. A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE*, 7:e47768.

Earl, D., Bradnam, K., John, J. St., … and Haussler, D. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods, *Genome Research*, 21, 2224-2241.

Ellegren, H. (2013). Genome sequencing and population genomics in non-model organisms, *Trends in Ecology & Evolution*, 29:1, 51-63.

FAO, *The State of World Fisheries and Aquaculture 2016 (SOFIA): Contributing to food security and nutrition for all*, Food and Agriculture Organization of the United Nations (FAO), Rome, Italy. 2016. 1-190.

Fischer, S., Brunk, B. P., Chen, F., … and Stoeckert, C. J. (2011). Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups. *Current Protocols in Bioinformatics.* 35:6,12.1-6.12.19.

Garb, J. E., Sharma, P. P. and Ayoub, N. A. (2018). Recent progress and prospects for advancing arachnid genomics. *Current opinion in insect science*, 25, 51-57.

Giribet, G. (2003). Molecules, development and fossils in the study of metazoan evolution; *Articulata versus Ecdysozoa revisited. Zoology*, 106:4, 303-326.

Giribet, G. and Edgecombe, G. D. (2013). The Arthropoda: a phylogenetic framework. In: Minelli, A., Boxshall, G. and Fusco, G. (eds) *Arthropod biology and evolution*. Springer, Berlin, Heidelberg. 17-40.

Grabherr, M. G., Haas, B. J., Yassour, M., … and Regev, A., (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*. 29:7, 644-652.

Grimaldi, D.A. (2009). 400 million years on six legs: on the origin and early evolution of Hexapoda. *Arthropod structure and development*, 39:2-3, 191-203.

Gutekunst, J., Andriantsoa, R., Falckenhayn, C., ... and Lyko, F. (2018). Clonal genome evolution and rapid invasive spread of the marbled crayfish. *Nature Ecology and Evolution*, 2:3, 567.

Iliopoulos, I., Tsoka, S., Andrade, M. A., … and Ouzounis, C. (2003). Evaluation of annotation strategies using an entire genome sequence, *Bioinformatics*, 19:6, 717-726.

Hardy, S. M., Lindgren, M., Konakanchi, H. and Huettmann, F. (2011). Predicting the Distribution and Ecological Niche of Unexploited Snow crab (*Chionoecetes opilio*) Populations in Alaskan Waters: A First Open-Access Ensemble Model. *Integrative and Comparative Biology*, 51:4, 608-622.

Hare, E. E. and Johnston, J. S., (2014). Genome size determination using flow cytometry of propidium iodide-stained nuclei. *In: Orogogozo, V. and Rockman, M. (eds) Molecular Methods for Evolutionary Genetics*. Humana Press, 3-12.

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

Hughes, C. L., Kaufman, T. C., (2002). *Hox* genes and the evolution of the arthropod body plan. *Evolution & development*, 4:6, 459-499.

Jain, M., Koren, S., Miga, K. … and Malla, S. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, 36:4, 338.

Jin, X., Xu, B., and Tang, Q. (2003). Fish assemblage structure in the East China Sea and southern Yellow Sea during autumn and spring. *Journal of Fish Biology*, 62:5, 1194-1205.

Johnson, G. M. (2019). GENETIC DIVERSITY AND POPULATION GENETIC STRUCTURE OF TANNER CRAB *CHIONOECETES BAIRDI* IN ALASKAN WATERS (Master's thesis). Retrieved from http://dspace31b.library.uaf.edu:8080/bitstream/handle/11122/10506/Johnson_G_2019.pdf

Joshua, Q. and Loman, N. J. (2019). DNA Extraction Strategies for Nanopore Sequencing, *In: Branton, D. and Deamer, D. W. (eds) Nanopore Sequencing: An Introduction*, World Scientific, Singapore, Singapore, 91-100.

Jørgensen, T.S., Nielsen, B.L.H., Petersen, B., … and Hansen, L.H. (2019a). The whole genome sequence and mRNA transcriptome of the tropical cyclopoid copepod *Apocyclops royi*. *G3: Genes, Genomes, Genetics*, 9:5, 1295-1302.

Jørgensen, T.S., Petersen, B., Petersen, H.C.B., ... and Hansen, B.W. (2019b). The genome and mRNA transcriptome of the cosmopolitan calanoid copepod *Acartia tonsa* Dana improve the understanding of copepod genome size evolution. *Genome biology and evolution*, 11:5, 1440-1450.

Kajitani, R., Toshimoto, K., Noguchi, H., … and Itoh, T., 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*. 24:8, 1384-1395.

Kang, J-H., Park, J-Y. and Kim, E-M. (2013). Population genetic analysis and origin discrimination of snow crab (*Chionoecetes opilio*) using microsatellite markers. *Molecular biology reports*, 40, 5563-5571.

Kang, S., Ahn, D.H., Lee, J.H., … and Park, H. (2017). The genome of the Antarctic-endemic copepod, Tigriopus kingsejongensis. *GigaScience*, 6:1, giw010.

Kao, D., Lai, A.G., Stamataki, E., … and Bruce, H. (2016). The genome of the crustacean Parhyale hawaiensis, a model for animal development, regeneration, immunity and lignocellulose digestion. *Elife 5*, e20062.

Karagozlu, M. Z., Barbon, M. M., Dinh, T. D., … and Kim, C-B. (2018). Complete mitochondrial genome of *Atergatis integerrimus* (Decapoda, Xanthidae) from the Philippines. *MITOCHONDRIAL DNA PART B: RESOURCES*, 3:1, 205-206.

Katoh, K., Rozewicki, J., and Yamada, K. D. (2017). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics*. doi.org/10.1093/bib/bbx108

Keilwagen, J., Hartung, F., Paulini, M., … and Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals, and fungi. *BMC Bioinformatics*, 19:189, 1-12.

Kenny, N. J., Sin, Y. W., Shen, X., … and Hui, J. H. L. (2014). Genomic Sequence and Experimental Tractability of a New Decapod Shrimp Model, *Neocaridina denticulata*. *Marine drugs*, 12, 1419-1437.

Kenny, N.J., Chan, K.W., Nong, W., … and Hui, J.H. (2016). Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. *Heredity*, 116:2, 190.

Kim, H. S., Kim, K-Y., Lee, S-H., … and Yi, C-H. (2019). The complete mitochondrial genome of *Pseudohelice subquadrata* (Dana, 1851) (Crustacea: Decapoda: Varunidae). *MITOCHONDRIAL DNA PART B: RESOURCES*, 4:1, 103-104.

Kim, W-J., Jung, H. T., Chun, Y. Y., … and Cha, H. K. (2012). Genetic Evidence for Natural Hybridization Between Red Snow Crab (*Chionoecetes japonicus*) and Snow Crab (*Chionoecetes opilio*) in Korea. *Journal of Shellfish Research*, 31:1, 49-56.

Knudsen, S. W., Møller, P. R., and Gravlund, P. (2007). Phylogeny of the snailfishes (Teleostei: Liparidae) based on molecular and morphological data. *Molecular phylogenetics and evolution*, 44:2, 649-666.

Kono, N. and Arakawa, K. (2019). Nanopore sequencing: Review of potential applications in functional genomics, *Development, Growth & Differentiation*, 61, 316-326.

Kosugi, S., Hirakawa, H., and Tabata, S. (2015). GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. *Bioinformatics*, 31. 3733-3741.

Kwak, W. (2017). *Establishing strategies of genome assembly for unprecedented species* (Doctoral dissertation). Seoul National University, Seoul, Republic of Korea. Retrieved from http://s-space.snu.ac.kr/handle/10371/125387

Lécher, P., Defaye, D. and Noel, P. (1995). Chromosomes and nuclear DNA of Crustacea. *Invertebrate Reproduction & Development*, 27:2. 85-114.

Lee, B.Y., Choi, B.S., Kim, M.S., ... and Lee, J.S. (2019). The genome of the freshwater water flea *Daphnia magna*: A potential use for freshwater molecular ecotoxicology. *Aquatic Toxicology*, 210, 69-84.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13, 2178-2189.

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12:323.

Li, W-N., Shao, R., Zhang, Q., Deng, W. and Xue, X-F. (2019). Mitochondrial genome reorganization characterizes various lineages of mesostigmatid mites (Acari: Parasitiformes), *Zoologica Scriptra*, 48, 679-689.

Lienhard, A. and Schäffer, S. (2019). Extracting the invisible: obtaining high quality DNA is a challenging task in small arthropods, *PeerJ*, http://dx.doi.org/10.7717/peerj.1147

Liu, S., Sun, J. and Hurtado, L.A. (2013). Genetic differentiation of *Portunus trituberculatus*, the world's largest crab fishery, among its three main fishing areas. *Fisheries research*, 148, 38-46.

Liu, L., Cui, Z., Song, C., … and Wang, C. (2016). Flow cytometric analysis of DNA content for four commercially important crabs in China. *Acta Oceanologica Sinica*, 35, 7-11.

Lozano-Fernandez, J., Carton, R., Tanner, A. R., … and Pisani, D. (2016). A molecular palaeobiological exploration of arthropod terrestrialization, *Philosphical Transactions of Royal Society B*, 371, 1-12.

Lozano-Fernandez, J., Tanner, A. R., Giacomelli, M., … and Pisani, D. (2019). Increasing species sampling in chelicerate genomic-scale datasets provides support for monophyly of Acari and Arachnida. *Nature communications*, 10:2295.

Lu, H., Giordano, F. and Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly, *Genomics Proteomics Bioinformatics*, 14, 265-279.

Luo, R., Liu, B., Xie, Y., … and Wang, J., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1:18, 1-6.

Lv, J., Gao, B., Liu, P., ... and Meng, X. (2017). Linkage mapping aided by *de novo* genome and transcriptome assembly in *Portunus trituberculatus*: applications in growth-related QTL and gene identification. *Scientific Reports*, 7874, 1-13.

Machner, J. and Scholtz, G. (2010). A scanning electron microscopy study of the embryonic development of *Pycnogonum litorale*, (Arthropoda, Pycnogonida). *Journal of morphology*, 271:11, 1306-1318.

Madoui, M.A., Poulain, J., Sugier, K., … and Jamet, J.L. (2017). New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod *Oithona*. *Molecular ecology*, 26:17, 4467-4482.

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics*, 20:16, 2878-2879.

Manuel, M., Jager, M., Murienne, J., … and Le Guyader, H. (2006). Hox genes in sea spiders (Pycnogonida) and the homology of arthropod head segments. *Development Genes and Evolution*, 216, 481-491.

Mapleson, D., Accinelli, G. G., Kettleborough, G., … and Clavijo, B. J. (2017). KAT: a K−mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, 33:4, 574-576.

Mardis, E. R. (2008). Next-Generation DNA Sequencing Methods, *Annual Review of Genomics and Human Genetics*, 9, 387-402.

Marshall, C. R. (2016). Explaning the Cambrian “Explosion” of Animals, *The Annual Review of Earth and Planetary Science*, 34, 355-384.

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27:6, 764-770.

Márquez, E. J., Hurtado-Alarcón, J. C., Isaza, … and Campos, N. H. (2016). Mitochondrial genome of the Caribbean king crab Damithrax spinosissimus (Lamarck, 1818) (Decapoda: Majidae). *MITOCHONDRIAL DNA PART A*, 27:3, 1724-1725.

Meader, S., Hillier, L D. W., Locke, D. … and Lunter, G. (2010). Genome assembly quality: Assessment and improvement using the nerutral indel model, *Genome Research*, 20, 675-684.

Meng, G., Li, Y., Yang, C. and Liu, S. (2019). MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acid Research*, 47:11, e63.

Meusemann, K., von Reumont, B. M., Simon, S., … and Misof, B. (2010). A phylogenomic approach to resolve the arthropod tree of life. *Molecular biology and Evolution*, 27:11, 2451-2464.

Mikheyev, A.S., and Tin, M.M. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular ecology resources*, 14:6, 1097-1102.

Miller, J. R., Koren, S. and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data, *Genomics*, 95, 315-327.

Minelli, A., Boxshall, G. and Fusco G. (2013). An Introduction to the Biology and Evolution of Arthropods, *In: Minelli, A., Boxshall, G. and Fusco G. (eds) Arthropod Biology and Evolution*, Molecules, Development, Morphology, Springer-Verlag, Heidelberg, Berlin, 1-16.

Mouchel-Vielh, E., Rigolot, C., Gibert, J. M. and Deutsch, J. S. (1998). Molecules and the body plan: the *Hox* genes of Cirripedes (Crustacea). *Molecular phylogenetics and evolution*, 9:3, 382-389.

Mora, C., Tittensor, D. P., Adl, S. … and Worm, B. (2011). How Many Species Are There on Earth and in the Ocean?, *PLoS Biology*, 9:8, e1001127.

Mullowney, D. R., Dawe, E. G., Morado, J. F. and Cawthorn, R. J. (2011). Sources of variability in prevalence and distribution of bitter crab disease in snow crab (*Chionoecetes opilio*) along the northeast coast of Newfoundland. *ICES Journal of Marine Science*, 68:3, 463-471.

Nachtweide, S. and Stanke, M. (2019). Multi-Genome Annotation with AUGUSTUS. *In: Kollmar, M. (eds) Gene Prediction. Methods in Molecular Biology, vol 1962*. Humana, New York, NY

Narzisi, G. and Mishra, B. (2011). Comparing De Novo Genome Assembly: The Long and Short of It, *PLos ONE*, 6:4, e19175.

Nossa, C.W., Havlak, P., Yue, J.X., … and Putnam, N.H. (2014). Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication. GigaScience, 3:1.

Ng, P. K. L., Guinot, D. and Davie, P. J. F. (2009). SYSTEMA BRACHYRORUM: PART I. AN ANNOTATED CHECKLIST OF EXTANT BRACHYURAN CRABS OF THE WORLD. *Raffles Bulletin of Zoology*, 1-286.

Nguyen, T. V., Jung, H., Rotllant, G., … and Ventura, T. (2018). Guidelines for RNA-seq projects: applications and opportunities in non-model decapod crustacean species. *Hydrobiologia*, 825, 5-27.

Niiyama, H. (1966). THE CHROMOSOMES OF TWO SPECIES OF EDIBLE CRABS (Brachyura, Decapoda, Crustacea) With Two Textfigures. *Bulletin of Fisheries Sciences*, Hokkaido University, 16:4, 201-205.

Oakley, T.H., Wolfe, J.M., Lindgren, A.R., and Zaharoff, A.K. (2012). Phylotranscriptomics to bring the understudied into the fold: monophyletic ostracoda, fossil placement, and pancrustacean phylogeny. *Molecular biology and evolution*, 30:1, 215-233.

Pace, R. M., Grbić, M. and Nagy, L. M. (2016). Composition and genomic organization of arthropod Hox clusters, *EvoDevo*, 7:11, 1-11.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes, *Bioinformatics*, 23:9, 1061-1067.

Paszkiewicz, K., and Studholme, D. J. (2010). *De novo* assembly of short sequence reads, *Briefings in Bioinformatics*, 2:5, 457-472.

Pepato, A. R. and Klimov, P. B. (2015). Origin and higher-level diversification of acariform mites – evidence from nuclear ribosomal genes, extensive taxon sampling, and secondary structure alignment, *BMC Evolutionary Biology*, 15:178, 1-20.

Phillippy, A. M., Schatz, M. C. and Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly, *Genome Biology*, 9, R55.

Phillippy, A. M. (2017). New advances in sequence assembly, *Genome Research*, 27, xi-xii.

Pisani, D., Carton, R., Campbell, L. I., … and Rota-Stabelli, O. (2013). An Overview of Arthropod Genomics, Mitogenomics, and the Evolutionary Origins of the Arthropod Proteome. *In: Minelli, A., Boxshall, G. and Fusco, G. (eds) Arthropod Biology and Evolution*. Springer, Berlin, Heidelberg. 41-61.

Poelchau, M., Childers, C., Moore, G., … and Hackett, K. (2014). The i5k Workspace@ NAL-enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic acids research*, 43(D1), D714-D719.

Pop, M., Kosack, D. S. and Salzberg, S. L. (2004). Hierachical Scaffolding With Bambus, *Genome Research*, 14, 149-159.

Poynton, H.C., Hasenbein, S., Benoit, J.B., ... and Werren, J.H. (2018). The Toxicogenome of *Hyalella azteca*: a model for sediment ecotoxicology and evolutionary toxicology. Environmental science and technology 52(10), 6009-6022.

Rahman, A. M. A., Kamath, S. D., Lopata, A. L., … and Helleur, R. J. (2011). Biomolecular characterization of allergenic proteins in snow crab (*Chionoecetes opilio*) and de novo sequencing of the second allergen arginine kinase using tandem mass spectrometry. *Journal of Proteomics*, 74, 231-241.

Reese, M. G., Hartzell, G., Harris, N. L., … and Lewis, S. E. (2003). Genome Annotation Assessment in *Drosophila melanogaster*, *Genome Research*, 10, 483-501.

Regier, J.C., Shultz, J.W., … and Cunningham, C.W. (2010). Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature*, 463:7284, 1079.

Rehm, P., Borner, J., Meusemann, K., … and Burmester, T. (2011). Dating the arthropod tree based on large-scale transcriptome data. *Molecular Phylogenetics and Evolution*, 61:3, 880-887.

Reumont, B.M. von, Jenner, R.A., Wills, M.A., … and Niehuis, O. (2011). Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of Hexapoda. *Molecular biology and evolution*, 29:3, 1031-1045.

Rhoads, A. and Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13:5, 278-289.

Rhodes, K. L. (1998). Seasonal trends in epibenthic fish assemblages in the near-shore waters of the western yellow sea, Qingdao, People's Republic of China. Estuarine, *Coastal and Shelf Science*, 46:5, 629-643.

Roach, M. J., Schmidt, S. A., and Borneman, A. R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, 19:460. doi.org/10.1186/s12859-018-2485-7

Ronquist, F., Teslenko, M., van der Mark, P., … and Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology*, 61:3, 539-542.

Rota-Stabelli, O. and Telford, M. J. (2008). A multi criterion approach for the selection of optimal outgroups in phylogeny: Recovering some support for Mandibulata over Myriochelata using mitogenomics, *Molecular Phylogenetics and Evolution*, 48, 103-111.

Rota-Stabelli, O., Daley, A. C. and Pisani, D. (2013). Molecular timetrees reveal a Cambrian colonization of land and a new scenario for ecdysozoan evolution. *Current Biology*, 23:5, 392-398.

Rotllant, G., Palero, F., Mather, P. B., Bracken-Grissom, H. D. and Santos, M. B. (2018). Preface: Recent advances in Crustacean Genomics. *Hydrobiologia*, 825, 1-4.

Ruan, J. and Li, H. (2019). Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, 17, 155-158.

Ryazanova, T. V., Fedotov, P. A. and Kharlamenko, V. I. (2016). The Bitter Crab Syndrome in Commercial Crabs in the Western Part of the Bering and Chukchi Seas. *Russian Journal of Marine Biology*, 42:5, 409-413.

Sabroux, R., Audo, D., Charbonnier, S., Corbari, L. and Hassanin, A. (2019). 150-million-year-old sea spiders (Pycnogonida: Pantopoda) of Solnhofen. *Journal of Systematic Palaeontology*, 1-12.

Sasaki, M., Akiyama-Oda, Y., and Oda, H. (2017). Evolutionary origin of type IV classical cadherins in arthropods. *BMC evolutionary biology*, 17:1, 142.

Savojardo, C., Luchetti, A., Martelli, P.L., ... and Mantovani, B. (2019). Draft genomes and genomic divergence of two *Lepidurus* tadpole shrimp species (Crustacea, Branchiopoda, Notostraca). *Molecular ecology resources*, 19:1, 235-244.

Sharma, P. P., Kaluziak, S. T., Pérez-Porro, A. R., … and Giribet, G. (2014). Phylogenomic interrogation of Arachnida reveals systemic conflicts in phylogenetic signal. *Molecular Biology and Evolution*, 31:11, 2963-2984.

Shin, S. C., Ahn, D. H., Kim, S. J. … and Park, H. (2013). Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes, *PLoS ONE*, 8:7, 1-9.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., … and E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31:19, 3210-3212.

Song, I., Lee, S. and Son, J. (2000). Molecular Cloning of Novel Genes Specifically Expressed in Snailfish, *Liparis tanakae*, *Development & Reproduction*, 4:1. 67-77.

Song, I., Lee, S. and Son, J. (2002). Molecular Cloning and Identification of Novel Genes, Gomsin, Characteristically Expressed in Snailfish, *Liparis tanakae*. *Development & Reproduction*, 6:1, 7-16.

Song, L., Bian, C., Luo, Y., … and Wang, Z. (2016). Draft genome of the Chinese mitten crab, *Eriocheir sinensis*. *GigaScience*, 5:1.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30:9, 1312-1313.

Stanke, M., Keller, O., Gunduz, I., … and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research*, 34:1, W435-W439.

Stein, L. (2001). GENOME ANNOTATION: FROM SEQUENCE TO BIOLOGY, *Nature Genetics*, 2, 493-592.

Stein, D. L. (2006). New and rare species of snailfishes (Scorpaeniformes: Liparidae) collected during the ICEFISH cruise of 2004. *Polar Biology*, 29:8, 705-712.

Stork, N.E. (2018). How many species of insects and other terrestrial arthropods are there on Earth?. *Annual review of entomology*, 63, 31-45.

Studholme, D. J. (2015). Genome Update, Let the consumer beware: *Streptomyces* genome sequence quality, *Micribial biotechnology*, 9:1, 3-7.

Tang, B., Wang, Z., Liu, Q., … and Li, Y. (2020a). High-Quality Genome Assembly of *Erichier japonica sinensis* Reveals Its Unique Cenome Evolution, *Frontiers in Genetics*, 17.

Tang, B., Zhang, D., Li, H., … and Ren, Y. (2020b). Chromosome-level genome assembly reveals the unique genome evolution of the swimming crab (*Portunus trituberculatus*), *GigaScience*, 9:1, giz161.

Tarailo-Graovac, M. and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols of Bioinformatics*, 25, 4-10.

Tomiyama, T., Ebe K., Kawata G. and Fujii T. (2009) Post-release predation on hatchery-reared Japanese flounder *Paralichthys olivaceus* in the coast of Fukushima, Japan. *Journal of Fish Biology*, 75, 2629-2641.

Tomiyama, T., Uehara, S., and Kurita, Y. (2013). Feeding relationships among fishes in shallow sandy areas in relation to stocking of Japanese flounder. *Marine Ecology Progress Series*, 479, 163-175.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25:9, 1105-1111.

Ustadi, U., Kim, K. Y., and Kim, S. M. (2005). Purification and identification of a protease inhibitor from glassfish (Liparis tanakai) eggs. *Journal of agricultural and food chemistry*, 53:20, 7667-7672.

Van Dam, M. H., Trautwein, M., Spicer, G. S. and Esposito, L. (2018). Advancing mite phylogenomics: Designing ultraconserved elements for Acari phylogeny, *Molecular Ecology Resources*, 19, 465-475.

Visel, A., Bristow, J., and Pennacchio, L.A. (2007). Enhancer identification through comparative genomics. In Seminars in cell and developmental biology. *Seminars in Cell & Developmental Biology*, 18:1, 140-152.

Wallberg, A., Bunikis, I, Pettersson, O. V. … and Webster, M. T. (2019). A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds, *BMC Genomics*, 20:275, 1-19.

Woolfe, A., and Elgar, G. (2008). Organization of conserved elements near key developmental regulators in vertebrate genomes. *Advances in genetics*, 61, 307-338.

Yuan, J., Gao, Y., Zhang, X., … and Xiang, J. (2017). Genome Sequences of Marine Shrimp *Exopalaemon carinicauda* Holthuis Provide Insights into Genome Size Evolution of Caridea, *Marine Drugs*, 15, 1-18.

Yuan, J., Zhang, X., Liu, C., … and Xiang, J. (2018). Genomic resources and comparative analyses of two economical penaeid shrimp species, *Marsupenaeus japonicus* and *Penaeus monodon*. *Marine Genomics*, 39, 22-25.

Zhang, C., Lim, J. H., Kwon, Y., … and Seo, Y. I., 2014. The current status of west sea fisheries resources and utilization in the context of fishery management of Korea. *Ocean & Coastal Management*. 102, 493-505.

Zhang, X., Yuan, J., Sun, Y., … and Ma, K.Y. (2019). Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. *Nature communications*, 10:1, 356.

Zhang, Y-X., Chen, X., Wang, J-P., … and Liu, M. (2019). Genomic insights into mite phylogeny, fitness, development, and reproduction. *BMC genomics*, 20:954, 1-22.

Zhang, Z-Q. (2013). Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. *Zootaxa*, 3703, 1-82.

Zhu, D. F., Wang, C. L. and Li, Z. Q. (2005). Karyotype analysis on *Portunus trituberculatus*. *Journal of Fisheries of China*, 29, 649-653.

Online-columns and webpages cited

FAO Fisheries and Aquaculture Department, (2019). Species Fact Sheets - Chionoecetes opilio (O. Frabricius, 1788). Retrieved 2020.04.21. from http://www.fao.org/fishery/species/2644/en

NCBI. (2020.05.03.). NCBI Genome List. Retrieved 2020.05.03. from https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/

NCBI. (2020.06.03.). NCBI ftp /genomes/. Retrieved from ftp://ftp.ncbi.nlm.nih.gov/genomes/

National Human Genome Research Institute (NIH). (2019.10.30.). The Cost of Sequencing a Human Genome. Retrieved 2020.06.03. from https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

PACBIO BLOG. (2020.05.05.). Beyond Contiguity - Assessing the Quality of Genome Assemblies with the 3 C's. Retrieved 2020.06.03. from https://www.pacb.com/blog/beyond-contiguity/.

Wikipedia. (2020.06.03.). List of sequenced animal genomes. Retrieved 2020.06.03. from https://en.wikipedia.org/wiki/List_of_sequenced_animal_genomes

Bioinformatics tools without publication cited with URLs

Babraham Bioinformatics. (2019.08.01.). FastQC, Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Hass, B. J. (2018.10.22.). TransDecoder, Retrieved from https://github.com/TransDecoder/TransDecoder/

Li, B., Dewey, C. and Liu, P. (2020.02.14.). RSEM (RNA-Seq by Excpectation-Maximization). Retrieved from http://deweylab.github.io/RSEM/

Zhu, Q. (2014.09.30.). BeforePhylo, Retrieved from https://github.com/qiyunzhu/BeforePhylo

Appendix

Appendix 1.

Appendix 1. Detailed list of sequenced animal genomes with their Scientific names visible, modified from Wikipedia article, List of Sequenced animal genomes (Latest update at 2020.05.24. Retrieved at 2020.06.03.)

Clade (deep branched)		Phylum / Subphylum	Class	Order	Scientific name	Habitat type
Non-Bilaterila	Non-Eumetazoa	Porifera	Demospongiae	Haplosclerida	*Amphimedon queenslandica*	Marine
					Xestospongia testudinaria	Marine
				Scopalinida	*Stylissa carteri*	Marine
	Basal Eumetazoa	Ctenophora	Tentaculata	Cydippida	*Pleurobrachia bachei*	Marine
				Lobata	*Mnemiopsis leidyi*	Marine
	ParaHoxozoa	Placozoa	N/A	N/A	*Trichoplax adhaerens*	Marine
					Hoilungia hongkongensis	Marine
		Cnidaria	Anthozoa	Actiniaria	*Aiptasia pallida*	Marine
					Nematostella vectensis	Marine
				Alcyonacea	*Dendronephthya gigantea*	Marine
				Pennatulacea	*Renilla muelleri*	Marine
				Scleractinia	*Acropora digitifera*	Marine
					Pocillopora damicornis	Marine
					Stylophora pistillata	Marine
			Cubozoa	Carybdeida	*Alatina alata*	Marine
			Hydrozoa	Leptothecata	*Clytia hemisphaerica*	Marine
				Anthoathecata	*Hydra vulgaris*	Marine
			Scyphozoa	Rhizostomeae	*Nemopilema nomurai*	Marine

Appendix 1. continued

					Rhopilema esculentum	Marine
					Cassiopea xamachana	Marine
				Semaeostomeae	*Aurelia aurita*	Marine
			Staurozoa	Stauromedusae	*Calvadosia cruxmelitensis*	Marine
		Chordata / Hemichordata	Enteropneusta	Enteropneusta	*Ptychodera flava*	Marine
					Saccoglossus kowalevskii	Marine
		Chordata / Echinodermata	Asteroidea	Valvatida	*Acanthaster planci*	Marine
			Echinoidea	Echinoida	*Strongylocentrotus purpuratus*	Marine
			Holothuroidea	Synallactida	*Apostichopus japonicus*	Marine
		Chordata / Urochordata	Ascidiacea	Enterogona	*Ciona intestinalis*	Marine
				Enterogona	*Ciona savignyi*	Marine
			Appendicularia	Copelata	*Oikopleura dioica*	Marine
		Chordata / Cephalochordata	Leptocardii	Amphioxiformes	*Branchiostoma floridae*	Marine
Bilateria	Deuterostomia	Chordata / Vertebrata	Hyperoartia	Petromyzontiformes	*Petromyzon marinus*	Marine
			Chondrichthyes	Chimaeriformes	*Callorhinchus milii*	Marine
				Orectolobiformes	*Rhincodon typus*	Marine
				Lamniformes	*Carcharodon carcharias*	Marine
				Orectolobiformes	*Chiloscyllium punctatum*	Marine
				Carcharhiniformes	*Scyliorhinus torazame*	Marine
			Actinopterygii	Anabantiformes	*Betta splendens*	Terrestrial(freshwater)
				Beloniformes	*Oryzias latipes*	Marine
				Centrarchiformes	*Oplegnathus fasciatus*	Marine
				Characiformes	*Astyanax mexicanus*	Terrestrial(freshwater)
				Cichliformes	*Oreochromis niloticus*	Terrestrial(freshwater)
					Metriaclima zeb	Terrestrial(freshwater)

Appendix 1. continued

	Clupea harengus	Marine
Clupeiformes	*Coilia nasus*	Marine
	Sardina pilchardus	Marine
	Anabarilius grahami	Terrestrial(freshwater)
	Danio rerio	Terrestrial(freshwater)
Cypriniformes	*Heterandria formosa*	Terrestrial(freshwater)
	Oxygymnocypris stewartii	Terrestrial(freshwater)
	Megalobrama amblycephala	Terrestrial(freshwater)
	Heterandria formosa	Terrestrial(freshwater)
Cyprinodontiformes	*Nothobranchius furzeri*	Terrestrial(freshwater)
	Xiphophorus maculatus	Terrestrial(freshwater)
Esociformes	*Esox lucius*	Marine
Gadiformes	*Gadus morhua*	Marine
Gasterosteiformes	*Gasterosteus aculeatus*	Marine
Gymnotiformes	*Electrophorus electricus*	Terrestrial(freshwater)
Lepisosteiformes	*Lepisosteus oculatus*	Terrestrial(freshwater)
Osmeriformes	*Protosalanx hyalocranius*	Marine
	Channa argus	Terrestrial(freshwater)
	Dissostichus mawsoni	Marine
	Eleginops maclovinus	Marine
Perciformes	*Larimichthys crocea*	Marine
	Lutjanus campechanus	Marine
	Parachaenichthys charcoti	Marine
	Seriola dumerili	Marine
	Sillago sinica	Marine

Appendix 1. continued

			Sparus aurata	Marine
			Oncorhynchus mykiss	Terrestrial(freshwater)
		Salmoniformes	*Oncorhynchus tshawytscha*	Marine
			Salmo salar	Marine
		Scorpaeniformes	*Sebastes schlegelii*	Marine
		Siluriformes	*Ictalurus punctatus*	Marine
			mola mola	Marine
		Tetraodontiformes	*Takifugu rubripes*	Marine
			Tetraodon nigroviridis	Marine
	Sarcopterygii	Coelacanthiformes	*Latimeria chalumnae*	Marine
			Leptobrachium leishanense	Terrestrial(freshwater)
			Nanorana parkeri	Terrestrial(freshwater)
			Oophaga pumilio	Terrestrial(freshwater)
	Amphibia	Anura	*Rana (Lithobates) catesbeiana*	Terrestrial(freshwater)
			Rhinella marina	Terrestrial(freshwater)
			Vibrissaphora ailaonica	Terrestrial(freshwater)
			Xenopus tropicalis	Terrestrial(freshwater)
		Urodela	*Ambystoma mexicanum*	Terrestrial(freshwater)
			Alligator mississippiensis	Terrestrial(freshwater)
		Crocodylia	*Alligator sinensis*	Terrestrial(freshwater)
			Crocodylus porosus	Marine
	Reptilia		*Gavialis gangeticus*	Terrestrial(freshwater)
			Anolis carolinensis	Terrestrial
		Squamata	*Dopasia gracilis*	Terrestrial
			Emydocephalus ijimae	Terrestrial

Appendix 1. continued

			Eublepharis macularius	Terrestrial
			Hydrophis melanocephalus	Terrestrial
			Laticauda colubrina	Marine
			Laticauda laticaudata	Marine
			Ophiophagus hannah	Terrestrial
			Pantherophis guttatus	Terrestrial
			Pogona vitticeps	Terrestrial
			Python molurus bivittatus	Terrestrial
			Salvator Merinae	Terrestrial
			Shinisaurus crocodilurus	Terrestrial
		Testudines	*Chelonia mydas*	Marine
			Chrysemys picta bellii	Terrestrial(freshwater)
			Pelodiscus sinensis	Terrestrial(freshwater)
			Platysternon megacephalum	Terrestrial(freshwater)
	Aves	Accipitriformes	*Aegypius monachus*	Terrestrial
			Aquila chrysaetos	Terrestrial
			Haliaeetus albicilla	Terrestrial
			Haliaeetus leucocephalus	Terrestrial
		Anseriformes	*Anas platyrhynchos*	Terrestrial
		Apodiformes	*Chaetura pelagica*	Terrestrial
		Bucerotiformes	*Buceros rhinoceros silvestris*	Terrestrial
		Caprimulgiformes	*Antrostomus carolinensis*	Terrestrial
		Cariamiformes	*Cariama cristata*	Terrestrial
		Cathartiformes	*Cathartes aura*	Terrestrial
		Charadriiformes	*Charadrius vociferus*	Terrestrial

Appendix 1. continued

	Himantopus novaezelandiae	Terrestrial
	Himantopus himantopus	Terrestrial
	Recurvirostra avosetta	Terrestrial
Ciconiiformes	*Nipponia nippon*	Terrestrial
Coliiformes	*Colius striatus*	Terrestrial
Columbiformes	*Columba livia*	Terrestrial
Coraciiformes	*Merops nubicus*	Terrestrial
Cuculiformes	*Cuculus canorus*	Terrestrial
	Tauraco erythrolophus	Terrestrial
Eurypygiformes	*Eurypyga helias*	Terrestrial
Falconiformes	*Falco cherrug*	Terrestrial
	Falco peregrinus	Terrestrial
Galliformes	*Gallus gallus*	Terrestrial
	Meleagris gallopavo	Terrestrial
	Pavo cristatus	Terrestrial
	Syrmaticus mikado	Terrestrial
	Tetrao tetrix	Terrestrial
Gaviiformes	*Gavia stellata*	Terrestrial
Gruiformes	*Balearica regulorum*	Terrestrial
	Chlamydotis macqueenii	Terrestrial
Leptosomiformes	*Leptosomus discolor*	Terrestrial
Mesitornithiformes	*Mesitornis unicolor*	Terrestrial
Opisthocomiformes	*Opisthocomus hoazin*	Terrestrial
Passeriformes	*Acanthisitta chloris*	Terrestrial
	Corvus brachyrhynchos	Terrestrial

Appendix 1. continued

	Corvus hawaiiensis	Terrestrial
	Eopsaltria australis	Terrestrial
	Ficedula albicollis	Terrestrial
	Ficedula hypoleuca	Terrestrial
	Geospiza fortis	Terrestrial
	Hirundo rustica	Terrestrial
	Lonchura striata domestica	Terrestrial
	Manacus vitellinus	Terrestrial
	Lycocorax pyrrhopterus	Terrestrial
	Manacus vitellinus	Terrestrial
	Paradisaea rubra	Terrestrial
	Pteridophora alberti	Terrestrial
	Ptiloris paradiseus	Terrestrial
	Taeniopygia guttata	Terrestrial
Pelecaniformes	*Egretta garzetta*	Terrestrial
	Pelecanus crispus	Terrestrial
Phaethontiformes	*Phaethon lepturus*	Terrestrial
Phoenicopteriformes	*Phoenicopterus ruber ruber*	Terrestrial
Piciformes	*Picoides pubescens*	Terrestrial
Podicipediformes	*Podiceps cristatus*	Terrestrial
Procellariiformes	*Fulmarus glacialis*	Terrestrial
Pterocliformes	*Pterocles gutturalis*	Terrestrial
Psittaciformes	*Amazona leucocephala*	Terrestrial
	Amazona ventralis	Terrestrial
	Amazona vittata	Terrestrial

Appendix 1. continued

	Ara macao	Terrestrial
	Melopsittacus undulatus	Terrestrial
	Nestor notabilis	Terrestrial
Struthioniformes	*Struthio camelus australis*	Terrestrial
Sphenisciformes	*Aptenodytes forsteri*	Marine
	Aptenodytes patagonicus	Marine
	Eudyptes chrysocome	Marine
	Eudyptes chrysolophus chrysolophus	Marine
	Eudyptes chrysolophus schlegeli	Marine
	Eudyptes filholi	Marine
	Eudyptes moseleyi	Marine
	Eudyptes pachyrhynchus	Marine
	Eudyptes robustus	Marine
	Eudyptes sclateri	Marine
	Eudyptula minor albosignata	Marine
	Eudyptula minor	Marine
	Eudyptula novaehollandiae	Marine
	Megadyptes antipodes antipodes	Marine
	Pygoscelis adeliae	Marine
	Pygoscelis antarctica	Marine
	Pygoscelis papua	Marine
	Spheniscus demersus	Marine
	Spheniscus humboldti	Marine
	Spheniscus magellanicus	Marine
Strigiformes	*Spheniscus mendiculus*	Terrestrial

Appendix 1. continued

			Strix occidentalis caurina	Terrestrial
			Strix varia	Terrestrial
			Phalacrocorax auritus	Marine
			Phalacrocorax brasilianus	Marine
		Suliformes	*Phalacrocorax carbo*	Marine
			Phalacrocorax harrisi	Marine
			Phalacrocorax pelagicus	Marine
		Tinamiformes	*Tinamus guttatus*	Terrestrial
		Trochiliformes	*Calypte anna*	Terrestrial
		Trogoniformes	*Apaloderma vittatum*	Terrestrial
		Monotremata	*Ornithorhynchus anatinus*	Terrestrial(freshwater)
		Didelphimorphia	*Monodelphis domestica*	Terrestrial
			Thylacinus cynocephalus	Terrestrial
		Dasyuromorphia	*Macropus eugenii*	Terrestrial
			Phascolarctos cinereus	Terrestrial
			Sarcophilus harrisii	Terrestrial
		Erinaceomorpha	*Erinaceus europaeus*	Terrestrial
	Mammalia	Eulipotyphla	*Solenodon Parodoxus*	Terrestrial
			Megaderma lyra	Terrestrial
			Eidolon helvum	Terrestrial
		Chiroptera	*Myotis lucifugus*	Terrestrial
			Pteronotus parnellii	Terrestrial
			Pteropus vampyrus	Terrestrial
			Rhinolophus ferrumequinum	Terrestrial
		Primates (Callitrichidae)	*Callithrix jacchus*	Terrestrial

Appendix 1. continued

Primates (Cercopithecidae)	*Macaca mulatta*	Terrestrial
	Macaca fascicularis	Terrestrial
	Rhinopithecus roxellana	Terrestrial
Primates (Galagidae)	*Otolemur garnettii*	Terrestrial
Primates (Hominidae)	*Pongo pygmaeus*	Terrestrial
	Pongo abelii	Terrestrial
	Gorilla gorilla	Terrestrial
Primates (Hominidae, Homininae)	*Homo sapiens*	Terrestrial
	Homo neanderthalensis	Terrestrial
	Pan troglodytes	Terrestrial
	Pan paniscus	Terrestrial
Carnivora (Felidae)	*Acinonyx jubatus*	Terrestrial
	Felis silvestris catus	Terrestrial
	Panthera leo	Terrestrial
	Panthera pardus	Terrestrial
	Panthera tigris altaica	Terrestrial
	Panthera tigris tigris	Terrestrial
	Panthera uncia	Terrestrial
	Prionailurus bengalensis	Terrestrial
Carnivora (Canidae)	*Canis lupus familiaris*	Terrestrial
	Canis lupus lupus	Terrestrial
	Lycaon pictus	Terrestrial
Carnivora (Ursidae)	*Ailuropoda melanoleuca*	Terrestrial
	Ursus arctos ssp. Horribilis	Terrestrial
	Ursus americanus	Terrestrial

Appendix 1. continued

	Ursus maritimus	Terrestrial
Carnivora (Odobenidae)	*Odobenus rosmarus*	Terrestrial
Carnivora (Mustelidae)	*Enhydra lutris kenyoni*	Marine
	Mustela erminea	Terrestrial
	Mustela putorius furo	Terrestrial
	Pteronura brasiliensis	Terrestrial
Cetacea (Delphinidae)	*Tursiops truncatus*	Marine
	Balaenoptera acutorostrata	Marine
	Balaenoptera physalus	Marine
	Neophocaena phocaenoides	Marine
	Orcinus orca	Marine
	Sousa Chinensis	Marine
Cetacea (Monodontidae)	*Delphinapterus leucas*	Marine
Cetacea (Physeteridae)	*Physeter macrocephalus*	Marine
Proboscidea	*Elephas maximus*	Terrestrial
	Loxodonta africana	Terrestrial
Sirenia	*Trichechus manatus*	Marine
Perissodactyla	*Equus ferus caballus*	Terrestrial
Artiodactyla (Antilocapridae)	*Antilocapra americana*	Terrestrial
Artiodactyla (Suidae)	*Sus scrofa*	Terrestrial
Artiodactyla	*Ammotragus lervia*	Terrestrial

Appendix 1. continued

(Bovidae)	*Antidorcas marsupialis*	Terrestrial
	Bison bonasus	Terrestrial
	Bos grunniens	Terrestrial
	Bos primigenius indicus	Terrestrial
	Bos primigenius taurus	Terrestrial
	Bubalus bubalis	Terrestrial
	Capra ibex	Terrestrial
	Cephalophus harveyi	Terrestrial
	Connochaetes taurinus	Terrestrial
	Damaliscus lunatus	Terrestrial
	Gazella thomsoni	Terrestrial
	Hippotragus niger	Terrestrial
	Kobus ellipsiprymnus	Terrestrial
	Litocranius walleri	Terrestrial
	Oreotragus oreotragus	Terrestrial
	Oryx gazella	Terrestrial
	Ourebia ourebi	Terrestrial
	Ovis ammon	Terrestrial
	Ovis ammon polii	Terrestrial
	Nanger granti	Terrestrial
	Neotragus moschatus	Terrestrial
	Neotragus pygmaeus	Terrestrial
	Philantomba maxwellii	Terrestrial
	Procapra przewalskii	Terrestrial
	Pseudois nayaur	Terrestrial

Appendix 1. continued

		Raphicerus campestris	Terrestrial
		Redunca redunca	Terrestrial
		Syncerus caffer	Terrestrial
		Sylvicapra grimmia	Terrestrial
		Tragelaphus buxtoni	Terrestrial
		Tragelaphus strepsiceros	Terrestrial
		Tragelaphus imberbis	Terrestrial
		Tragelaphus spekii	Terrestrial
		Tragelaphus scriptus	Terrestrial
		Taurotragus oryx	Terrestrial
		Cervus albirostris	Terrestrial
		Elaphurus davidianus	Terrestrial
	Artiodactyla (Cervidae)	*Muntiacus crinifrons*	Terrestrial
		Muntiacus muntjak	Terrestrial
		Muntiacus reevesi	Terrestrial
		Rangifer Tarandus	Terrestrial
		Giraffa camelopardalis	Terrestrial
	Artiodactyla (Giraffidae)	*Giraffa camelopardalis tippelskirchi*	Terrestrial
		Okapia johnstoni	Terrestrial
	Artiodactyla (Moschidae)	*Moschus berezovskii*	Terrestrial
		Moschus chrysogaster	Terrestrial
	Artiodactyla (Tragulidae)	*Tragulus javanicus*	Terrestrial
	Rodentia	*Hydrochoerus hydrochaeris*	Terrestrial
		Mus musculus	Terrestrial

Appendix 1. continued

				Rattus norvegicus	Terrestrial
				Peromyscus leucopus	Terrestrial
			Lagomorpha	*Oryctolagus cuniculus*	Terrestrial
				Blattella germanica	Terrestrial
				Cryptotermes secundus	Terrestrial
			Blattodea	*Macrotermes natalensis*	Terrestrial
				Periplaneta americana	Terrestrial
				Zootermopsis nevadensis	Terrestrial
				Aquatica lateralis	Terrestrial
				Dendroctonus ponderosae	Terrestrial
			Coleoptera	*Photinus pyralis*	Terrestrial
				Protaetia brevitarsis	Terrestrial
				Tribolium castaneum	Terrestrial
Protostomia	Arthropoda / Hexapoda	Insecta	Diptera (Calliphoridae)	*Aldrichina grahami*	Terrestrial
				Dasypogon diadema	Terrestrial
			Diptera (Chironomidae)	*Parochlus steinend*	Terrestrial
				Proctacanthus coquilletti	Terrestrial
				Aedes aegypti	Terrestrial
				Aedes albopictus	Terrestrial
				Anopheles darlingi	Terrestrial
				Anopheles gambiae Strain: PEST	Terrestrial
			Diptera (Culicidae)	*Anopheles gambiae Strain: M*	Terrestrial
				Anopheles gambiae Strain: S	Terrestrial
				Anopheles sinensis	Terrestrial
				Anopheles stephensii	Terrestrial

Appendix 1. continued

	Anopheles arabiensis	Terrestrial
	Anopheles quadriannulatus	Terrestrial
	Anopheles merus	Terrestrial
	Anopheles melas	Terrestrial
	Anopheles christyi	Terrestrial
	Anopheles epiroticus	Terrestrial
	Anopheles maculatus	Terrestrial
	Anopheles culicifacies	Terrestrial
	Anopheles minimus	Terrestrial
	Anopheles funestus	Terrestrial
	Anopheles dirus	Terrestrial
	Anopheles farauti	Terrestrial
	Anopheles atroparvus	Terrestrial
	Anopheles sinensis	Terrestrial
	Anopheles albimanus	Terrestrial
	Culex quinquefasciatus	Terrestrial
	Drosophila albomicans	Terrestrial
	Drosophila ananassae	Terrestrial
	Drosophila biarmipes	Terrestrial
	Drosophila bipectinata	Terrestrial
Diptera (Drosophilidae)	*Drosophila erecta*	Terrestrial
	Drosophila elegans	Terrestrial
	Drosophila eugracilis	Terrestrial
	Drosophila ficusphila	Terrestrial
	Drosophila grimshawi	Terrestrial

Appendix 1. continued

	Drosophila kikkawai	Terrestrial
	Drosophila melanogaster	Terrestrial
	Drosophila mojavensis	Terrestrial
	Drosophila neotestacea	Terrestrial
	Drosophila persimilis	Terrestrial
	Drosophila pseudoobscura	Terrestrial
	Drosophila rhopaloa	Terrestrial
	Drosophila santomea	Terrestrial
	Drosophila sechellia	Terrestrial
	Drosophila simulans	Terrestrial
	Drosophila takahashi	Terrestrial
	Drosophila virilis	Terrestrial
	Drosophila willistoni	Terrestrial
	Drosophila yakuba	Terrestrial
Diptera (Phoridae)	*Megaselia abdita*	Terrestrial
Diptera Psychodidae)	*Clogmia albipunctata*	Terrestrial
Diptera (Sarcophagidae)	*Sarcophaga Bullata*	Terrestrial
Diptera (Syrphidae)	*Episyrphus balteatus*	Terrestrial
Hemiptera	*Acyrthosiphon pisum*	Terrestrial
	Ericerus pela	Terrestrial
	Laodelphax striatellus	Terrestrial
	Lycorma delicatula	Terrestrial
	Rhodnius prolixus	Terrestrial
	Rhopalosiphum maidis	Terrestrial

Appendix 1. continued

	Sitobion miscanthi	Terrestrial
	Triatoma rubrofasciata	Terrestrial
	Acromyrmex echinatior	Terrestrial
	Apis mellifera	Terrestrial
	Atta cephalotes	Terrestrial
	Camponotus floridanus	Terrestrial
	Cerapachys biroi	Terrestrial
	Harpegnathos saltator	Terrestrial
	Lasius niger	Terrestrial
Hymenoptera	*Linepithema humile*	Terrestrial
	Nasonia girault	Terrestrial
	Nasonia longicornis	Terrestrial
	Nasonia vitripennis	Terrestrial
	Nomia Melanderi	Terrestrial
	Pogonomyrmex barbatus	Terrestrial
	Solenopsis invicta	Terrestrial
	Antharaea yamamai	Terrestrial
	Bicyclus anynana	Terrestrial
	Bombyx mori	Terrestrial
	Cydia pomonella	Terrestrial
Lepidoptera	*Danaus plexippus*	Terrestrial
	Eudocima phalonia	Terrestrial
	Heliconius melpomene	Terrestrial
	Melitaea cinxia	Terrestrial
	Megathymus ursus violae	Terrestrial

Appendix 1. continued

			Papilio bianor	Terrestrial
			Pieris rapae	Terrestrial
			Plutella xylostella	Terrestrial
			Spodoptera frugiperda	Terrestrial
		Orthoptera	*Locusta migratoria*	Terrestrial
		Phthiraptera	*Pediculus humanus*	Terrestrial
		Trichoptera	*Stenopsyche tienmushanensis*	Terrestrial
Arthropoda / Crustacea	Hexanauplia	Calanoida	*Acartia tonsa dana*	Marine
		Harpacticoida	*Tigriopus kingsejongensis*	Marine
	Branchiopoda	Anomopoda	*Daphnia pulex*	Terrestrial(freshwater)
		Spinicaudata	*Eulimnadia Texana*	Terrestrial(freshwater)
	Malacostraca	Amphipoda	*Parhyale hawaiensis*	Marine
		Decapoda	*Neocaridina denticulata*	Marine
		Decapoda	*Procambarus virginalis*	Terrestrial(freshwater)
		Decapoda	*Portunus trituberculatus*	Marine
Arthropoda / Chelicerata	Merostomata	Xiphosura	*Carcinoscorpius rotundicauda*	Marine
			Limulus polyphemus	Marine
	Arachnida	Araneae	*Acanthoscurria geniculata*	Terrestrial
			Dysdera silvatica	Terrestrial
			Nephila clavipes	Terrestrial
			Parasteatoda tepidariorum	Terrestrial
			Stegodyphus mimosarum	Terrestrial
		Ixodida	*Ixodes scapularis*	Terrestrial
		Mesostigmata	*Tropilaelaps mercedesae*	Terrestrial
		Scorpiones	*Mesobuthus martensii*	Terrestrial

Appendix 1. continued

		Trombidiformes	*Tetranychus urticae*	Terrestrial
Myriapoda	Chilopoda	Geophilomorpha	*Strigamia maritima*	Terrestrial
Tardigrada	Eutardigrada	Parachaela	*Hypsibius dujardini*	Terrestrial(freshwater)
Mollusca	Bivalvia	Arcida	*Scapharca broughtonii*	Marine
		Mytilida	*Bathymodiolus platifrons*	Marine
			Limnoperna fortunei	Marine
			Modiolus philippinarum	Marine
			Mytilus galloprovincialis	Marine
		Ostreida	*Crassostrea gigas*	Marine
			Saccostrea glomerata	Marine
		Pectinida	*Argopecten purpuratus*	Marine
			Chlamys farreri	Marine
			Patinopecten yessoensis	Marine
			Pecten maximus	Marine
		Pteriida	*Pinctada fucata*	Marine
		Unionida	*Venustaconcha ellipsiformis*	Terrestrial(freshwater)
		Venerida	*Ruditapes philippinarum*	Marine
	Cephalopoda	Octopoda	*Octopus bimaculoides*	Marine
			Octopus minor	Marine
			Octopus vulgaris	Marine
		Oegopsida	*Architeuthis dux*	Marine
		Sepiida	*Euprymna scolopes*	Marine
	Gastropoda	Caenogastropoda	*Pomacea canaliculata*	Terrestrial(freshwater)
		Heterobranchia	*Biomphalaria glabrata*	Terrestrial(freshwater)
			Elysia chlorotica	Marine

Appendix 1. continued

			Patellogastropoda	*Lottia gigantea*	Marine
			Sigmurethra	*Achatina fulica*	Terrestrial(freshwater)
			Vetigastropoda	*Haliotis discus hannai*	Marine
				Echinococcus granulosus	Terrestrial(freshwater)
				Echinococcus multilocularis	Terrestrial(freshwater)
				Hymenolepis microstoma	Terrestrial(freshwater)
		Cestoda	Cyclophyllidea	*Schistosoma haematobium*	Terrestrial(freshwater)
	Platyhelminthes			*Schistosoma japonicum*	Terrestrial(freshwater)
				Schistosoma mansoni	Terrestrial(freshwater)
				Taenia solium	Terrestrial(freshwater)
			Tricladida	*Schmidtea mediterranea*	Terrestrial(freshwater)
		Rhabditophora	Plagiorchiida	*Clonorchis sinensis*	Terrestrial(freshwater)
		Trematoda	Diplostomida	*Schistosoma haematobium*	Terrestrial
			Ascaridida	*Ascaris suum*	Terrestrial
				Ancylostoma ceylanicum	Terrestrial
				Bursaphelenchus xylophilus	Terrestrial
				Caenorhabditis angaria	Terrestrial
				Caenorhabditis brenneri	Terrestrial
				Caenorhabditis briggsae	Terrestrial
	Nematoda	Chromadorea	Rhabditida	*Caenorhabditis elegans*	Terrestrial
				Caenorhabditis remanei	Terrestrial
				Dirofilaria immitis	Terrestrial
				Haemonchus contortus	Terrestrial
				Heterorhabditis bacteriophora	Terrestrial
				Pristionchus pacificus	Terrestrial

Appendix 1. continued

				Brugia malayi	Terrestrial
			Spirurida	*Loa loa*	Terrestrial
				Onchocerca volvulus	Terrestrial
				Wuchereria bancrofti	Terrestrial
			Strongylida	*Necator americanus*	Terrestrial
				Globodera pallida	Terrestrial
			Tylenchida	*Heterodera glycines*	Terrestrial
				Meloidogyne hapla	Terrestrial
				Meloidogyne incognita	Terrestrial
		Enoplea	Mermithida	*Romanomermis culicivorax*	Terrestrial
				Trichuris suis	Terrestrial
			Trichocephalida	*Trichuris muris*	Terrestrial
				Trichuris trichiura	Terrestrial
	Annelida	Polychaeta	Capitellidae	*Capitella teleta*	Marine
		Clitellata	Rhynchobdellida	*Helobdella robusta*	Terrestrial(freshwater)
			Haplotaxida	*Eisenia fetida*	Terrestrial
	Brachiopoda	Lingulata	Lingulada	*Lingula anatina*	Marine
	Rotifera	Eurotatoria	Bdelloidea	*Adineta vaga*	Terrestrial(freshwater)

ABSTRACT (In Korean)

포스트게놈 시대의 도래에 따라 드노보 유전체 조립은 비모델 생명체의 생명현상을 연구하는데 필수적인 과정이 되었다. 비모델 절지동물의 드노보 조립된 유전체의 사례는 근래에 들어 급격하게 증가했다. 그러나, 해양 절지동물은 놀라울 정도로 다양한 분류군과 형태를 가짐에도 불구하고, 가장 드노보 유전체 조립 연구가 미흡한 분류군 중 하나이다. 현재까지 보고된 해양 절지동물의 드노보 유전체 조립 연구는 대부분이 그 양과 질 모두가 제한적이다. 그러므로, 본 연구는 국내에서 최초로 선행 연구가 미흡한 해양 절지동물 분류군인 바다거미 강과 단미 하목에 초점을 맞춰 드노보 유전체 조립 및 분석을 실시하였다. 본 연구의 결과로, 1건의 미토콘드리아 유전체와 4건의 전장유전체가 드노보 조립되었으며, 조립된 유전체의 특징이 기술되었다. 단서열 염기서열결정법으로 조립된 두 건의 유전체의 품질은 비교적 낮았으나, 장서열 염기서열결정법을 주로하여 조립된 *Nymphon striatum*과 *Chionoecetes opilio* 유전체가 매우 풍부한 고품질 유전체 정보를 제공한다는 것이 밝혀졌다. 본 연구에서 수행된 기초적인 계통유전체학 연구는 바다거미 강과 십각 목을 각각 대표하는 드노보 조립된 유전체를 최초로 포함했으며, 이를 통해 최근 논란의 대상인 거미강에 속하는 투구게류 가설을 지지하는 결과를 나타내는 것으로 밝혀졌다. 더 나아가, 비생물정보학 연구실 환경에서 이루어지는 드노보 유전체 연구에서 발생하는 제한요인들을

분석함으로써 비모델 해양 절지동물의 드노보 유전체 연구에 최적화된 안정적인 연구방법론을 제시하였다.

주요어 : 계통유전체학, 드노보 유전체 조립, 미토콘드리아 유전체, 비교유전체학, 전장유전체, 해양 절지동물