WHOLE GENOME SEQUENCING FOR ESTABLISHMENT OF PHYLOGENETIC POSITION AND DISCOVERY OF MINIATURIZATION AND LIPID METABOLISM GENES IN *PAEDOCYPRIS*

SAM KA KEI

UNIVERSITI SAINS MALAYSIA

2022

WHOLE GENOME SEQUENCING FOR ESTABLISHMENT OF PHYLOGENETIC POSITION AND DISCOVERY OF MINIATURIZATION AND LIPID METABOLISM GENES IN *PAEDOCYPRIS*

by

SAM KA KEI

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

May 2022

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my supervisor, Prof. Alexander Chong Shu Chien, for his invaluable support, suggestions and guidance throughout the study. He has been giving me constructive advice and support with his expertise in the academic field. He is serious all the time but undeniable, he is a good philosopher, a good leader, and a great teacher to follow. “I cannot teach anybody anything, but I can only make them think” his strong mind, hard work, critical thinking and suggestion have given me effective training before stepping out into my career after graduation. I sincerely appreciate my best friend or senior in Chinese we called “shifu”, Dr. Lau Nyok Sean, for her constant support, guidance, patience, and encouragement throughout my PhD study. She is helping me all the time in terms of research and also giving me unlimited support.

I would like to thank the assistance offered by all members in the labs: Dr. Yasmin, Lishen, Melissa, MC Teoh, Wong She Cheng, Seng Yeat, Jana, Diyana, Wong, Yeap, Shima, Liana, Diana, Jess, Fiza, Fatanah, Aliza, Fadzli, Luqman, and all the staff from the Centre for Chemical Biology. Not forget to thanks to my best friends, Leong Kar Xin and Sumaliny. All of them are very supportive and helped me complete my study successfully. I also thank Universiti Sains Malaysia for providing financial assistance during my PhD study.

I am grateful for the unconditional love and support my parents gave, without which the completion of this study would be impossible. The encouragement and trust served as a source of motivation to overcome the challenges I have encountered during my PhD study.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS AND SYMBOLS

%	Percentage
^{o}C	Degree Celsius
Δ	Delta
α	Alpha
β	Beta
g	gram
mL	Millilitre
µg	Microgram
v/v	Volume per volume
C	Carbon
cDNA	Complementary deoxyribonucleic acid
DEPC	Diethylpyrocarbonate
dH_2O	Distilled water
DHA	Docosahexaenoic acid
gDNA	Genomic deoxyribonucleic acid
dNTP	Deoxyribonucleoside triphosphate
ALA	α – linolenic acid
ARA	Arachidonic acid
ATP	Adenosine triphosphate
BF_3	Boron trifluoride
EFA	Essential fatty acid
Elovl	Eicosapentaenoic acid
FA	Fatty acid
Fads	Fatty acid desaturase
FAME	Fatty acid methyl ester
GC-MS	Gas chromatography–mass spectrometry
LA	Linoleic acid
LB	Luria-bertani
LC-PUFA	Long-chain polyunsaturated fatty acid
LDL	Low density lipoprotein
mRNA	Messenger RNA
MUFA	Monounsaturated fatty acid
n-3	Omega-3
n-6	Omega-6
OD	Optical density
ORF	Open reading frame
PCR	Polymerase chain reactions
RNA	RNA
SCMM-U	*S. cerevisiae* minimal medium plates without uracil
VLC-PUFA	Very long-chain polyunsaturated fatty acid
VLC-SFA	Very long-chain saturated fatty acid
YPD	Yeast extract peptone dextrose

LIST OF APPENDICES

PENJUJUKAN KESELURUHAN GENOM UNTUK PENETAPAN KEDUDUKAN FILOGENETIK DAN PENEMUAN GEN PENGKERDILAN DAN METABOLISME LIPID DALAM *PAEDOCYPRIS*

ABSTRAK

Paedocypris adalah merupakan antara vertebrata terkecil di dunia dengan jumlah panjang ≤ 20 mm yang terhad kepada persekitaran hutan paya gambut yang berasid. Anatomi *Paedocypris* telah banyak diubahsuai dengan ciri perkembangan yang terpenggal. Fenotip kerdil yang unik menjadikan *Paedocypris* sebagai model yang sesuai untuk memahami perkembangan dan evolusi. Tiga *Paedocypris* spesies telah dihuraikan, termasuk *P. micromegethes* dari Sarawak dan dua spesies dari Indonesia, *P. progenetica* dan *P. carbunculus*. Setakat ini, *Paedocypris* tidak mempunyai genom skala kromosom, dan hanya draf genom dengan perancah N50 ~59-61 kb. Kajian ini telah menghasilkan genom bertahap kromosom berkualiti tinggi untuk *P. micromegethes* dengan perancah N50 saiz 29.8 Mb menggunakan gabungan teknologi penjujukan Illumina, PacBio dan Hi-C. Sementara itu, draf genom untuk *P. carbunculus* telah ditambah baik dengan panjang perancah N50 sebanyak 125 Kb. Kajian ini mendedahkan bahawa saiz genom yang kecil dan padat *Paedocypris* berciri kandungan ulangan yang rendah, intron pendek, dan kekurangan duplikasi genom keseluruhan. Sebagai tambahan, kajian ini juga menemui kehilangan gen daripada genom *Paedocypris*, seperti *fgf8b*, *mtmr14*, *scpp3a*, *scpp3b*, *scpp6*, *scpp7 scpp8, gsp37, spp2*, dan *tbx22*. Analisis filogenetik untuk menyelesaikan pertikaian kedudukan taksanomi ikan kerdil cyprinid dengan menggunakan 13 gen pengekodan protein (PCG) dan dua gen RNA ribosom (rRNA) daripada mitokondria menunjukkan nilai sokongan tinggi untuk hubungan kumpulan beradik bagi *Paedocypris* dan

Sundadanio kepada kelompok keluarga Danionidae. Perbezaan kedudukan dalam topologi pokok untuk ikan kerdil mencadangkan bahawa mereka tidak membentuk klad monofiletik. Kebolehan biosintesis asid lemak rantai panjang politaktepu dalam *P. micromegethes* juga telah dikaji. Walaupun saiz genom *Paedocypris* telah mengalami penyusutan, bilangan salinan gen desaturase dan elongases yang dijumpa dalam perhimpunan data genom dan transkrip adalah sama dengan *Danio rerio*. Tambahan pula, keputusan fungsi *in-vitro* menunjukkan bahawa *P. micromegethes* mempunyai keupayaan untuk biokonversi C18:2n-6 dan C18:3n-3 kepada asid arakidonik (ARA, 20:4n-6), asid eicosapentaenoic (EPA, 20:5n-3) dan asid docosahexaenoic (DHA, 22:6n-3), masing masing. Pengekalan laluan biosintesis LC-PUFA yang lengkap dalam *Paedocypris* adalah untuk memastikan penjajahan dan adaptasi dalam persekitaran akuatik yang berasid.

WHOLE GENOME SEQUENCING FOR ESTABLISHMENT OF PHYLOGENETIC POSITION AND DISCOVERY OF MINIATURIZATION AND LIPID METABOLISM GENES IN *PAEDOCYPRIS*

ABSTRACT

Paedocypris is the world's smallest vertebrate with a total length of ≤ 20 mm restricted to acidic freshwater peat swamp environments. The anatomy of *Paedocypris* is highly modified with developmental truncated characters. The unique miniature phenotypes make the *Paedocypris* a good model fish to study development and evolution. Three *Paedocypris* species have been described, including *P. micromegethes* from Sarawak and another two species, *P. progenetica* and *P. carbunculus* from Indonesia. Thus far, *Paedocypris* has no chromosome-scale genome, with only a draft genome with scaffold N50 ~59-61 kb. In this study, a high-quality chromosome level genome assembly for the *P. micromegethes* was generated with improved scaffold N50 of 29.8 Mb using a combination of Illumina, PacBio and Hi-C sequencing technologies. Meanwhile, the draft genome for the *P. carbunculus* was improved with N50 scaffold length of 125 Kb. This study revealed that the small and compact genome sizes of miniature *Paedocypris* are characterized by a low repeat content, short introns, and a lack of recent genome duplications. Additionally, this study discovered gene loss from the improved genome assemblies, such as *fgf8b, mtmr14, scpp3a, scpp3b, scpp6, scpp7 scpp8, gsp37, spp2,* and *tbx22*. Phylogenetic analysis to resolve the contentious taxonomy position of miniaturized cyprinids using mitochondrial 13 protein-coding genes (PCGs) and two ribosomal RNA genes (rRNAs) showed improved supportive values for the sister-group relationship for *Paedocypris* and *Sundadanio* to the Danionidae family. A different

position for the miniature fish within the tree topology suggested that they do not form monophyletic clade. The capacity for long-chain polyunsaturated fatty acid (LC-PUFAs) biosynthesis in *P. micromegethes* was also studied. Identical copy numbers of desaturase and elongases genes with *Danio rerio* were identified in *the Paedocypris* genome and transcriptome assemblies, despite its genome size reduction. *In-vitro* functional study results revealed that the *P. micromegethes* has the ability for bioconversion of C18:2n-6 and C18:3n-3 to arachidonic acid (ARA, 20:4n-6), eicosapentaenoic acid (EPA, 20:5n-3), and docosahexaenoic acid (DHA, 22:6n-3), respectively. The retaining of a complete LC-PUFA biosynthesis pathway in this species most likely ensures successful colonization and adaptation in the acidic aquatic environment.

CHAPTER 1

INTRODUCTION

1.1 Overview

The evolution of miniaturization is observed in all vertebrate lineages. However, among the miniature organisms, *Paedocypris* comprised the world's smallest vertebrate. Therefore, to understand the evolution of miniaturization, the two best candidates, *P. micromegethes* and *P. carbunculus*, were selected for this study. *Paedocypris* species are restricted to the highly acidic (pH 3~5) dark-coloured freshwater peat swamp forest distributed in Southeast Asia. The acidic freshwater ecosystem in the peat swamp is depleted in nutrients but high in tannins (Yule and Gomez, 2009;Page and Rieley, 2016). Interestingly, adaptive evolution has emerged specialized life forms to adapt to this harsh aquatic environment, and most of them are restricted to this habitat (Leete, 2006;Posa et al., 2011). Acidic peat swamp forest consists of many miniature creatures, including fish (Kottelat et al., 2006). *Paedocypris* represents a phenotypic extreme in adult body size reduction among vertebrates and is classified as the smallest vertebrate known in the world. They have a sexually matured body size for an adult fish that can reach only 7.9-11.5 mm standard length (Kottelat et al., 2006;Britz and Kottelat, 2008). Despite body size reduction, *Paedocypris* possesses developmentally truncated characters, resulting in massive bones loss and reduction (Britz and Conway, 2009). Owing to its simplified morphology and remarkable biological characteristics, including its well-adaptation in

such a hostile environment, *Paedocypris* serve as an excellent model for the studies of evolution, development, and acidic freshwater adaptation.

Sequencing the genome is an important task that can help provide a clue to scientists on the gene functions and how genes work together to direct an organism's growth, development and maintenance through analyzing the sequenced genome. The advent of next-generation sequencing technology has enabled scientists to discover the organism's genetic information by sequencing its whole genome and now can even sequence up to a chromosome level. Nevertheless, the paucity of genome information of this tiny teleost fish has hindered understanding of its ecology and evolution. Furthermore, over-exploitation of peat swamp forests has threatened the survival of the fish population, urging the need for a genomic study to conserve the genomic information before the fish is going to be extinct. To date, a draft whole-genome sequence of *Paedocypris* (*P. micromegethes* and *P. carbunculus*) was released. Nevertheless, the employment of Illumina short-read sequencing technology has led to the creation of highly fragmented genome assembly with scaffold N50 of ~59-61 Kb (Malmstrøm et al., 2018). Without a high-quality reference genome and the lack of completeness and continuity of the current *Paedocypris's* draft genome may become one of the most critical factors limiting its application for in-depth downstream analysis. Therefore, it is essential to improve the genome assembly and create a model reference genome for *Paedocypris* species to get more detailed information on their genome architectures and gene contents to understand the genotype-phenotype relationships towards the evolution of miniaturization.

Mitochondria DNA is the small circular chromosome found in all eukaryotes and serves as an essential component responsible for supplying energy for all eukaryotes (Saraste, 1999). This membrane-bound organelle is an effective genetic

marker in phylogeny, taxonomy and biological evolution analyses, owing to its maternal inheritance, simple genetic structure, easy detection, low recombination frequencies and rapid evolutionary rate (Miya et al., 2003;Gissi et al., 2008). Mitochondria genome (mitogenome) is a popular DNA molecule that has been successfully employed in studying the phylogenetic taxonomy among fishes (Miya and Nishida, 2015). However, the mitogenome of the miniature fish is still less well-understood, with only a few *Paedocypris* mitogenomes have been deposited in the GenBank database. The inconsistencies of the *Paedocypris* phylogenetic position arise from contradicting morphological identification and molecular analyses. No study has been done using a complete mitochondria gene set to examine the phylogenetic position of this miniaturized fish species among cyprinids. However, this method has been successfully applied to improve the controversial phylogenetic position of fish (Sun et al., 2021). Therefore, it is essential to assemble and study the mitogenome of *P. micromegethes* and *P. carbunculus* to provide a good resolution into its features, evolution and phylogenetic relationships among miniature fish and other cyprinids.

Long-chain polyunsaturated fatty acids (LC-PUFAs; ≥ C20), consisting of eicosapentaenoic acid (EPA; 20:5n-3), docosahexaenoic acid (DHA; 22:6n-3), and arachidonic acid (ARA; 20:4n-6), are important compounds for normal cell functions in vertebrates, including humans. These essential LC-PUFAs play vital roles in manipulating the cell's membrane phospholipid composition, modulating gene expression, regulating eicosanoid synthesis, and as cell-signaling molecules (Jump, 2002;Schmitz and Ecker, 2008). Due to the absence of D12 and D15 desaturases, fish, like all vertebrates, are incapable of *de novo* synthesizing linolenic acid (LA) and α-linolenic acid (ALA) and are solely dependent on dietary intake. Limited or inability of marine fish species in LC-PUFA biosynthesis has been hypothesized to adapt to the

DHA-rich marine environment. In contrast, enzymatic functions required for C18 PUFAs to LC-PUFAs conversion are conserved in most fish species living in DHA-deficient freshwater environments (Tocher, 2010). Studies on fish suggested that evolutionary history, environmental factors (habitats), trophic levels, and ecological backgrounds can influence the fatty acid biosynthesis capability (Kabeya et al., 2017;Xie et al., 2020). The miniature *Paedocypris* species mainly depend on planktonic rotifers and cladocerans for food (Kottelat et al., 2006). *Paedocypris* species are adapted well to this acidic freshwater environment. They might possess a functional LC-PUFA biosynthesis capacity to cope with the stress environment with limited food resources for LC-PUFA production. Thus, it is interesting to investigate the *Paedocypris* LC-PUFA biosynthesis pathways from the genome, and transcriptomic assemblies and then functional characterize the isolated fatty acid genes to provide an insight into how they colonize and adapt this acidic aquatic environment.

1.2 Problem statement

Understanding the genome features and genotype-phenotype relationship is vital for answering the molecular basis of evolutionary history and diversification of life; however, the current *Paedocypris* draft genomes are fragmented, and part of the genome sequences are missing or consisting of gaps. Thus, improving the contiguity of the genome assembly is essential to capture complete genetic information for deeper insights into the characteristics of the genome, the evolution of miniaturization and acidic adaptation, which can tell us much about the organisms in question. In this study, Malaysia's species *P. micromegethes* was selected to sequence up to a chromosomal-scale assembly using Illumina, PacBio and Hi-C sequencing techniques. Meanwhile, the contiguity of the *P. carbunculus* draft genome was improved to examine the genotype difference between the two miniature *Paedocypris* species.

Besides that, the mitogenome of the miniature fish is still less well-understood. Limited mitogenome resources are reported in the GenBank database, encouraging the assembly of the complete mitogenome from different *Paedocypris* species. *Paedocypris* has reduced genome size; thus, examining the miniaturization effects toward the mitogenome features is interesting. Mitogenome is a good marker for studying the phylogenetic relationships among organisms. Therefore, a complete set of mitochondrial genes was applied in this study to examine the phylogeny topology of miniature *Paedocypris* to explore the evolutionary relationships between different fish species.

Furthermore, the capability of LC-PUFA biosynthesis in *Paedocypris* is still unknown. This is due to the non-discovery of *fads* and *elovl* genes in this miniature species. Therefore, it is important to investigate the capacity of LC-PUFA in *Paedocypris* species based on the assembled genome and transcriptome to understand its LC-PUFA biosynthesis capability. Acquiring this knowledge could provide fundamental insight into how fatty acid biosynthesis capacity contributes to acidic freshwater adaptation in organisms.

1.3 Research objectives

The objectives of this study are stated below:

- To improve genome assemblies for *P. micromegethes* and *P. carbunculus*.
- To identify potential factors contributing to the evolution of miniaturization.
- To generate *Paedocypris* complete mitogenomes and determine their phylogenetic position.
- To functionally characterize putative *fads2* and *elovls* genes of *Paedocypris* in adaptation to the acidic peat swamp.

CHAPTER 2

LITERATURE REVIEW

2.1 Miniaturization

Miniaturization, the evolutionary reduction of adult body size, occurs among vertebrate lineages such as fish, amphibians, reptiles, mammals (Hanken and Wake, 1993) and within invertebrates, including insects (Polilov, 2015). Miniaturization is commonly associated with novel ecology, physiology and morphology (Hanken and Wake, 1993). The miniaturized phenotypes are unique because of the complex derived traits with reduction and structural simplification not developed in their larger relatives or ancestors. The consequences of miniaturization for the organism in physical (obvious) features are reduction and structural simplification, morphological novelty and increased morphological variability, whereas biologically features (nonobvious) are genome and cell size. Miniaturization has emerged as an adaptive evolution for living organisms to survive under critical environments. The evolution of miniaturization has influenced organismal physiology, ecology, life history, and behaviour (Hanken and Wake, 1993). Dwarfism can help organisms avoid predators, obtain new food sources, reduce intra-and interspecific competitions, invade new ecological niches or habitats and promote sex maturity at an early age (Hanken and Wake, 1993;Blanckenhorn, 2000). Extreme body size reduction is reported in teleosts, where species maturing at sizes less than 20 mm are defined as miniature fishes (Weitzman and Vari, 1988). Miniaturization has independently evolved several times within the bony fish from family Cyprinidae, where abundant miniature cyprinids are under subfamily Danioninae, the zebrafish relatives, such

as *Boraras*, *Danionella*, *Horadandia*, *Microboraras*, *Sundadanio*, and *Paedocypris* (Rüber et al., 2007). In addition, there are some miniature cyprinids from the subfamily Cyprininae, including African genera "*Barbus*" and *Barboides* and South Asian Sawbwa (Conway et al., 2017). Cyprinidae (Teleostei: Ostariophysi: Cypriniformes) is the most diverse vertebrate family consisting mainly of freshwater fishes with more than 3,000 species distributed in Africa, Eurasia, and North America (Nelson et al., 2016;Fricke et al., 2018). Two main phenotypic forms of dwarfism were observed and distinguished in danionine cyprinids: (1) proportionate dwarfism, tiny at the adult stage but almost identical copies to their larger relatives; and (2) developmental truncation, the acceleration of maturation, leading to dwarf sexually matured adults (progenetic miniatures), resembling the juvenile/larvae stage (an early developmental stage) of their larger relatives. These two miniature groups are similar in size range, but the proportioned dwarfs exhibit few reductions compared to their large ancestors, whereas developmentally truncated miniatures possess numerous losses (Rüber et al., 2007;Britz and Conway, 2009;Britz et al., 2014). Miniaturized cyprinids with remarkably developmental truncation features are *Barboides*, *Danionella*, *Sundadanio*, and *Paedocypris* (Rüber et al., 2007;Conway et al., 2017). Generally, reduction and structural simplifications are common among miniature cyprinids. However, morphological novelties are only exhibited in miniature species with a high degree of developmental truncation, and the majority are sexually dimorphic (Rüber et al., 2007;Britz, 2009;Britz and Conway, 2009). Thus, there is a possible relationship between developmental truncation and the novelty in morphological evolution (Britz, 2009;Britz and Conway, 2009;Britz et al., 2009).

2.2 Peat swamp forest

Peat swamp forest (PSF) is formed by accumulating partially decomposed dead leaves and wood over time under waterlogged conditions. This forest is one of the unique and harsh ecosystems in the tropical rainforest biome distributed extensively in Southeast Asia. About 60% of the world's peat swamp forests are distributed in Southeast Asia, and Malaysia is the second-largest country where the peat swamp forests are localized (Posa et al., 2011). PSF is also the most endangered ecosystem globally, under legal and illegal logging, fire and land conversion. Nevertheless, the biodiversity in this forest is still poorly understood. Peat swamp, also called "black water", is highly acidic (pH 3 to 5). The water is almost dark in colour due to the high concentration of humic acids and other phenolic compounds like tannins in this waterlogged environment. Peat swamp is a poor nutrient environment since leaf litters are the main build-up for the peat, and the water was attributed to the rainfall with no inflow of nutrients from either stream or river (Yule and Gomez, 2009). Highly acidic, low nutrients and poor light conditions (primarily due to the dark brown colour of water and shared with the forest canopy) have restricted algae or plankton growth in this harsh environment (Wehr and Sheath, 2003;Yule, 2008;Yule and Gomez, 2009). Surprisingly, this unique acidic ecosystem supports many flora and fauna despite unfavourable conditions like low pH, anaerobic, and poor nutrients (Yule, 2008). PSF endemism is likely due to the inhabitant of organisms in rigorous microhabitats provided by the peat. Adaptive evolution enables them to uniquely adapted to this stressed environment (Posa et al., 2011). Most of the fish species found in PSF are mainly under the Cyprinidae family, followed by Osphronemidae, Bagridae and Siluridae (Sule et al., 2016). Meanwhile, a high number of miniature fish species found in the aquatic ecosystem of PSF are under the Cyprinidae family (Kottelat et al., 2006),

such as *Boraras* (Conway and Kottelat, 2011), *Sundadanio* (Conway et al., 2011), *Fangfangia* (Britz et al., 2011), and the smallest recorded fish belong to *Paedocypris* species (Kottelat et al., 2006). In PSF, small size is advantageous for fish, as they can survive droughts in shallow pools with low water levels, animal burrows, or soil (Kottelat et al., 2006).

2.3 *Paedocypris* species

Paedocypris (Teleostei; Order: Cypriniformes) is the genus of the family Cyprinidae (Rüber et al., 2007;Britz and Conway, 2009) and *P. progenetica*, *P. micromegethes* and *P. carbunculus* are the three described species to date (Kottelat et al., 2006;Rüber et al., 2007). *Paedocypris* is derived from the Greek words "*Paideios*" (children) and "*Cypris*" (Venus), a common suffix for cyprinid genera. *Paedocypris* is the world's smallest vertebrate and was first discovered in Southeast Asia's peat swamp forests (Kottelat et al., 2006). These miniature fish species are restricted and stereotopic to acidic peat swamp forests, for example, *P. progenetica* at Sumatra and Bintan island, Indonesia, *P. micromegethes* at Matang, Sarawak, and *P. carbunculus* at Kalimantan, Indonesia (Kottelat et al., 2006;Rüber et al., 2007). Besides that, there are some undescribed *Paedocypris* species, such as *Paedocypris* sp "North Selangor", "Pahang", and "Pondok Tanjung" were distributed in the peat swamp forests of Peninsular Malaysia (Sule et al., 2016;Ng et al., 2019), and also from Indonesia, namely *Paedocypris* sp "Pulau Singkep" and "Banka" (Rüber et al., 2007).

Paedocypris is the smallest fish and vertebrate known in the world, displaying a remarkably larval-like appearance with a sexually matured body size for adult fish that can reach only 7.9-11.5 mm standard length. (Kottelat et al., 2006;Britz and

Kottelat, 2008). The anatomy of *Paedocypris* is simplified, combined with developmentally truncated features, including loss of scales, reduction in the number of fin rays, absence of some bones, poorly ossified skeleton, presence of post-anal-fin fold along the ventral edge of caudal peduncle that resembles the fish larvae, and a marked reduction in size (Kottelat et al., 2006;Britz and Kottelat, 2008;Britz and Conway, 2009). In addition, *Paedocypris* is sexually dimorphic. This unique morphological novelty was observed in the modified pelvic fins in males associated with the greatly enlarged and supported keratinized pads of skin, termed "flange and hook" (Kottelat et al., 2006).

2.3.1 *Paedocypris* serve as an important model fish to study

Model organisms play a vital role to provide insight into the general principles in biological research related to genetics, development and evolution. They usually possess unique or interesting characteristics helped to decipher specific processes. In the past decades, model organisms often possess favourable traits such as short generation times, easy handling and culturing, and facilitating experimental laboratory research. However, due to the advancement of sequencing technologies, many whole-genome assemblies, "genomic" models (from organisms with unique/favourable traits) are available, and most of the research studies are based on the analyzed data obtained from the assembled genome (Hedges, 2002). *Paedocypris* species is miniature fish, truncated in development with unusual reproductive mode and is adapted well in an acidic environment, making this fish a potential model organism for genetic study of vertebrate development, the evolution of miniaturization, acidic adaptation, and reproductive behaviour (Liu et al., 2012).

2.3.2 Morphological classification and phylogenetic position of *Paedocypris*

There was considerable interest to resolve the phylogenetic position of *Paedocypris* using morphological characteristics (Britz and Conway, 2009; Mayden and Chen, 2010; Britz et al., 2014), a combination of nuclear and/- or mitochondrial markers (Fang et al., 2009;Mayden and Chen, 2010;Tang et al., 2010;Tang et al., 2013; Stout et al., 2016;Hirt et al., 2017) or single marker gene alone (Rüber et al., 2007). As a result, it was suggested that *Paedocypris* was a member of the Danioninae subfamily (Rüber et al., 2007;Fang et al., 2009;Tang et al., 2010;Tang et al., 2013), within Cyprinidae family but not within Danioninae subfamily (Yang et al., 2015), or as a lineage sister to all Cypriniformes (Mayden and Chen, 2010;Stout et al., 2016;Hirt et al., 2017). The differences in *Paedocypris* phylogenetic position observed from several molecular studies could be due to the high evolutionary rate and base compositional heterogeneity in miniature fishes. It is known that base compositional heterogeneity and long branches (associated with high molecular evolution rates) can influence phylogenetic reconstruction (Hirt et al., 2017).

2.4 Genome assembly

Each species adapts and lives in a specific environmental condition, an ecological niche, that shapes its unique genome sequence and expression (Stange et al., 2021). Thus, it is crucial to sequence and assembles the organisms' genome to provide insights into their genome diversity. Genome assembly constructs a representative genome sequence of that particular organism through a computational process of taking a large number of short or long DNA fragments and aligning them together to form the original genome sequence (Baker, 2012). Genome assembly is a hierarchical process with continuous improvement of sequence length; for example, the shortest assembly at the beginning are contigs derived from the assembly of sequence reads. Hereafter, the contigs are assembled into longer scaffolds, and then the scaffold lengths are further improved to chromosome-scale (Choudhuri, 2014). There are *de novo* and reference-based of genome assemblies, *de novo* refers to sequencing a novel genome without using a reference sequence, whereas reference-based assembly uses a reference-guided assembly strategy (with the availability of a related genome) during mapping/alignment (Lischer and Shimizu, 2017).

2.4.1 Fish genome study

Many fish genomes were sequenced and assembled to investigate or identify genetic factors underpinning evolution that contributed to the diverse traits among fish species. Tine et al., 2014 generated a high-quality chromosome-scale genome of European sea bass *Dicentrarchus labrax,* a temperate-zone euryhaline teleost. Gene families related to ion and water regulation were identified expanded in the genome, providing insights into euryhaline adaptation. Mudskippers are amphibious teleost fishes that are uniquely evolved to live on mudflats. You et al., 2014 sequenced and assembled the genome of four mudskippers (*Boleophthalmus pectinirostris*, *Scartelaos histophorus*, *Periophthalmodon schlosseri* and *Periophthalmus magnuspinnatuto*) to examine genetic changes underlying adaptation towards terrestrial life. Comparative genomic analyses, including gene family and positive selection, were carried out. Expansion of genes involved in the innate immune system was identified from gene family analysis, suggesting their roles in defence against terrestrial pathogens. For positive selection, genes involved in the ammonia excretion pathway were positively selected that could help mudskippers tolerate environmental ammonia. Genomic analysis also revealed genetic changes in mudskipper genomes associated with aerial vision, such as the absence or mutation of certain vision-related genes. The genome from another economically important marine fish species, the large yellow croaker *Larimichthys crocea* was sequenced and assembled (Ao et al., 2015). This fish species displays peculiar behavioral and physiological features, particularly sensitive to environmental stresses, a good model to study. Analysis of the genome revealed expansions of several gene families, specifically associated with vision-related crystallins, olfactory receptors, and the auditory sense related genes, providing insights into the genetic mechanisms of environmental stress adaptation. Whole-

genome sequencing of three endemic cavefish *Sinocyclocheilus* species: the surface-dwelling *S. grahami*; the semi-cave-dwelling *S. rhinocerous*; and the cave-restricted *S. anshuiensis* was performed by Yang et al., 2016 to investigate and understand how genomic change in adaptation to the isolated cave life. Comparative analyses of the *Sinocyclocheilus* genomes showed many genetic changes, including loss of gene, pseudogenes, mutations, and down-regulation that contributed to the regressive characters, such as degeneration of eye, albinism, rudimentary scales, and low fecundity. They also found the expansion of genes associated with the sense of taste in the cave-restricted species, providing a better understanding of cavefish biology. Besides that, the genome of the largest bony fish, ocean sunfish *Mola mola* was sequenced and assembled by Pan et al., 2016. This world's largest fish can grow up to 2.7 m and weigh 2.3 tons. It possessed notable characteristics, such as fast growth rate and largely cartilaginous skeleton. Genomic and comparative analyses showed that growth hormone and insulin-like growth factor 1 (GH/IGF1) were positively selected, pointing to its fast growth rate and large body size. Meanwhile, several genes associated with the extracellular matrix involved in the regulation of bone and cartilage development were under positive selection, which could be the factors contributing to the cartilaginous bone. In Braasch et al., 2016, spotted gar *Lepisosteus oculatus*, whose lineage diverged from teleost before teleost genome duplication, was chosen to be sequenced to facilitate in the connection between human and teleost biomedical models. Analyses of the slowing evolving gar genome showed that many entire gar chromosomes conserved with some tetrapods, including human and some undetectable human-teleost conserved noncoding elements become apparent when using the gar genome. Thus, the gar genome could assist in identifying potential regulatory gene elements shared by teleost and humans. The genomes of the world's smallest

vertebrate, *Paedocypris* micromegethes and *P. carbunculus* were sequenced and assembled by Malmstrøm et al., 2018. The unique features of this miniature species are developmental truncation and remaining larval phenotype with an adult body size of < 8mm. Analyses of the two miniature *Paedocypris* genomes showed evolutionary simplification through extensive loss of *Hox* genes and genome reduction associated with their developmental truncation features. In addition, loss of genes related to the development of muscle, nervous system and skeletal were identified from the genomes, contributing to their progenetic phenotype. Genome size reduction of the *Paedocypris* was associated with short intron and low repeat content. Amazonian freshwater fish *Arapaima gigas* is the largest freshwater fish in the world. Du et al., 2019 have sequenced and assembled the *Arapaima* genome to understand this giant fish's biology. Gene family and positive selection analyses identified genes related to its unique characters, especially its large size and fast growth. The application of genome sequence from both sexes (male and female) and RAD-tag analyses enable the identification of male- (Y-) specific scaffolds, which is useful in developing the XY sex-determination system in *Arapaima*. The Antarctic environment is the coldest place on Earth where blackfin icefish *Chaenocephalus aceratus* live. Icefishes are "white blooded" because they are the only vertebrate without functional haemoglobin genes and red blood cells. Whole-genome sequencing on blackfin icefish was performed by Kim et al., 2019. Gene family analysis on the icefish genome identified genes involved in protection from ice damage and control cellular redox state are highly expanded, as an evolutionary adaptation to this cold and high oxygen concentration Antarctic environment, respectively. Loss of genes involved in regulating circadian homeostasis was detected in the blackfin icefish genome. The availability of the blackfin icefish genome enables the understanding of how they adapt

to an extreme environment. Anglerfish *Lophius litulon* is a widely consumed fish with unique features, including a scaleless body, diverse feeding habits and large liver. A high-quality genome assembly of anglerfish *L. litulon* was generated (Lv et al. 2020a). Comparative genomic analyses, including gene family and positive selection, were performed. Gene family analysis revealed expanded and contracted gene families are correlated with adaptation to the benthic environment and its scaleless phenotype; for positive selection, genes involved in metabolic processes are positively selected, which could help diversify the prey for anglerfish. Hadal zone is an extreme environment with high hydrostatic pressure, low temperatures, lack of food supply, and limitation of light. Snailfish is the fish inhabiting the hadal environment with a depth below 6,000 m, and its genome has been sequenced and assembled by Mu et al., 2021. Comparative genomic analyses showed that genes involved with DNA repair were positively selected and expanded in the snailfish genome, showing the importance of maintaining DNA integrity under high hydrostatic pressure. Gene families associated with taste receptors, olfactory receptors, and vision-related genes were significantly changed in the snailfish genome, suggesting these changes might help snailfish adapt to nutrient-poor and dark hadal environments. All findings from fish genome studies mentioned above were summarized in Table 2.1.

Table 2.1 Summary of the fish genome study.

Fish species	Findings from genome
European sea bass (*Dicentrarchus labrax*) (Tine et al., 2014)	Expansion of gene families contribute to euryhaline adaptation
Mudskipper *Boleophthalmus pectinirostris, Scartelaos histophorus, Periophthalmodon schlosseri* and *Periophthalmus magnuspinnatuto* (You et al., 2014)	Genome analyses revealed expansion of genes involved in the innate immune system, positive selection on genes involved in the ammonia excretion pathway, and absence or mutation of certain vision-related genes associated with terrestrial life adaptation in mudskipper.
Large yellow croaker *Larimichthys crocea* (Ao et al., 2015)	Expansion of gene families associated to the adaptation and response to environmental stress
Sinocyclocheilus cavefish (Yang et al., 2016)	Genetic changes, including loss of gene, pseudogenes, mutations, down-regulation, and gene expansion is associated with the evolved cavefish features for adaptation
Ocean sunfish *Mola mola* (Pan et al., 2016)	Identify positively selected genes corresponding to the fast growth rate and largely cartilaginous skeleton
Spotted gar *Lepisosteus oculatus* (Braasch et al., 2016)	Many entire gar chromosomes are conserved from bony vertebrate ancestors, facilitating the comparison of human and teleost medical models.
Miniature fish *Paedocypris* (Malmstrøm et al., 2018)	Developmental truncation of *Paedocypris* or evolutionary simplification through the loss of *Hox* genes or genes related to muscle, nervous system, and skeleton development, whereas short intron and low repeat content contribute to their small genome size. Findings from the genomic analyses have provided insights into miniaturization.
Amazonian freshwater fish *Arapaima gigas* (Du et al., 2019)	Gene family and positive selection analyses identified genes related to its unique characters, especially its large size (gigantism) and fast growth. Isolation of the male- (Y-) specific scaffolds and supports the XY sex determination system.
Antarctic blackfin icefish *Chaenocephalus aceratus* (Kim et al., 2019)	Discovered the expanded genes involved in protection from ice damage and control cellular redox state and absence of genes involved in regulating circadian homeostasis in the blackfin icefish genome, providing insights on Antarctic adaptation.
Anglerfishes *Lophius litulon* (Lv et al. 2020a)	Comparative genomic analyses, including gene family and positive selection, discovered the genes that contribute to the scaleless phenotype, adaptation to the benthic environment and diverse feeding habits.
Yap hadal snailfish (YHS) (Mu et al., 2021)	Comparative genomic analyses showed that genes involved with DNA repair were positively selected and expanded in the snailfish genome; the gene families associated with taste receptors, olfactory receptors, and vision-related genes were significantly changed in the snailfish genome providing insight into adaptation in hadal environments.

2.5 DNA sequencing

DNA sequencing is the process to determine nucleotide content (Adenosine, Guanine, Cytosine, and Thymine) in a DNA fragment. The earliest sequencing (first generation) strategy, known as Sanger sequencing, was developed in 1977 by F. Sanger and colleagues based on a chain-termination approach (Sanger et al., 1977). The Sanger sequencing was commonly applied in single sequencing reactions using a specific DNA primer, especially in verifying PCR products or plasmid constructs. Since then, the sequencing technologies have been constantly improved, and "next-generation" sequencing has begun to emerge.

2.5.1 Next generation sequencing

Next-generation sequencing methods enable the examination of billions of DNA and RNA templates via a high-throughput sequencing technique. Short-read-based second-generation sequencing provides low cost and higher accurate data (Goodwin et al., 2016;Levy and Myers, 2016;Slatko et al., 2018). Short-reads paired-end is being used for genome survey, genomic base correction and assembly in fish, such as *Oxygymnocypris stewartii* (Liu et al., 2019), golden pompano *Trachinotus ovatus* (Zhang et al., 2019) and *Labeo rohita* (Das et al., 2020). Meanwhile, this low-cost and high-throughput Illumina sequencing approach has been successfully used to sequence and assemble the small organelle, the chloroplast and mitochondria (Nunez and Oleksiak, 2016; Cheng et al., 2020). The sequencing approach has been applied to sequence and assemble many fish mitochondria DNA (mtDNA), and the studies are provided in Table 2.2.

Table 2.2 Sequencing approach utilized in the fish mitochondrial genome assembly.

Fish species	Sequencing approach	Genome assembly	Application
Tonguefish, *Cynoglossus trigrammus* (Mu et al., 2015)	Illumina 2 ×500 bp paired-end reads	Complete	Unusual mitochondrial genome structure and phylogenetic analysis
Endangered species Japanese Swellshark *Cephalloscyllium umbratile* (Zhu et al., 2017)	Illumina 2 ×101 bp paired-end reads	Complete	mtDNA features and phylogenetic analysis
Banded cusk-eel *Raneya brasiliensis* (Fromm et al., 2019)	Illumina 2 ×150 bp paired-end reads	Complete	Gene rearrangement and phylogenetic reconstructions
Medicinal fish *Cyprinion semiplotum* (Sharma et al., 2020)	Illumina 2 ×150 bp paired-end reads	Complete	mtDNA features and phylogenetic implications

Besides that, in short-insert paired-end sequencing, Illumina also provides long-insert mate-pair sequencing, which can sequence more than 1kb DNA fragments. A combination of short-insert paired-end and long-insert mate-pair sequences can improve genome assembly quality. Long insert reads (mate pairs) facilitate mapping across a greater distance that can better cover highly repetitive regions. In contrast, short insert reads (paired-end) can fill in gaps missed by the long inserts (van Heesch et al., 2013). This combination approach (Figure 2.1) is widely adopted in many whole-genome sequencing projects, including vertebrate and invertebrate organisms. For vertebrates, especially in fish, both Illumina paired-end and mate-pair were applied in the generation of the draft genome in whale shark (Weber et al., 2020), forming a hybrid assembly approach with Nanopore in *Danionella translucida* to achieve high assembly contiguity (Kadobianskyi et al., 2019), and to perform initial genome assembly prior to chromosome construction in Anglerfish genome (Lv et al., 2020a). In invertebrates, both Illumina paired-end and mate-pair were applied to generate draft genome with PacBio in jellyfish (Gold et al., 2019) and cricket (Ylla et al., 2021) for

evolutionary study. Thousands of eukaryotic genomes were assembled based on Illumina reads; however, the mate-pair libraries are still unable to span all the repetitive regions, resulting in highly fragmented assemblies. Therefore, long-read sequencing approaches such as Pacific Bioscience (PacBio) and Oxford Nanopore Technology (ONT) and chromosome conformation capture methods, including Hi-C and Dovetail Genomics Chicago libraries were, emerged to resolve and span the repetitive regions in order to improve the continuity of genome assemblies (Elbers et al., 2019).

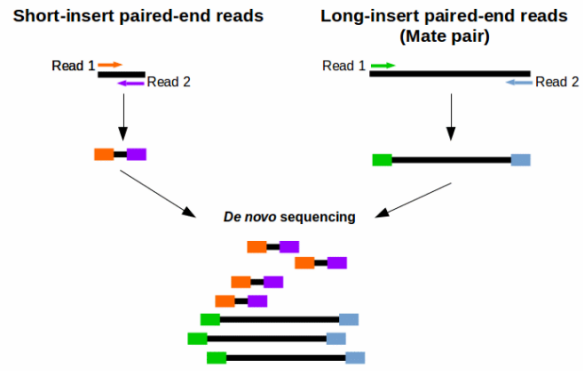

Figure 2.1 Combination of long inserts from mate pair sequencing with that from short-insert paired-end sequencing for *de novo* assembly. This figure was adopted from website: https://www.ecseq.com/support/ngs/what-is-mate-pair-sequencing-useful-for

With the rapid evolution of sequencing technologies, third-generation sequencing (also called long-read sequencing) technology focuses on generating long reads DNA (or can be RNA) longer than second-generation sequencing was developed. The technology adopted is called Single-Molecule Sequencing in Real-Time (SMRT), which enables the sequencing of very long fragments up to 30 to 50 kb or longer (average 10-25 kb). The long-continuous reads can help produce large scaffolds to overcome limitations encountered when using second-generation sequencing, specifically to assemble the highly repetitive genomic regions that mainly caused numerous short breaks in short-reads assemblies. Long-read PacBio or ONT can generate long sequence reads for prokaryotic genome assemblies. However, for eukaryotic genomes, this technique can improve their genome assemblies to scaffold-level, but insufficient decipher to the chromosome-scale level (Belser et al., 2018).

2.5.2 Chromosome conformation capture techniques

New technologies utilize long-range linking information, including linked reads (10x Genomics) (Zheng et al., 2016;Yeo et al., 2017), optical mapping (BioNano) (Schwartz et al., 1993;Shelton et al., 2015;Vij et al., 2016) and chromosome conformation capture such as Hi-C and Dovetail Genomics Chicago libraries (Dovetail) (Simonis et al., 2006;Burton et al., 2013;Elbers et al., 2019) were being adopted to validate orientation and further improve draft genome assemblies by spanning the previously unsaffolded contigs. Hi-C data can provide long-range linkage information across a variety of length scales, spanning tens of megabases. Meanwhile, Hi-C data can be mapped to draft genome assemblies to scaffold the contigs, generating super/ultra-long-range-scaffolds in the size range of chromosomes (Elbers et al., 2019;Ghurye et al., 2019). Therefore, Hi-C data is being adopted in many

genome projects to improve the draft genome assemblies up to a chromosome scale to obtain a high-quality reference genome (Burton et al., 2013;Marie-Nelly et al., 2014;Dudchenko et al., 2017;Zhang et al., 2019). Illumina, PacBio, and Hi-C high-throughput sequencing technologies have been used in *de novo* fish genome studies, and chromosome-level genome assemblies were reported, as summarized in Table 2.3.

Table 2.3 Sequencing approach applied for fish chromosome-level genome assembly.

Fish species	**Sequencing approach**	**Genome assembly**
Takifugu favidus (Zhou et al., 2019)	Illumina short reads*, PacBio Sequel**, Hi-C***	Chromosome-level
Ancherythroculter nigrocauda (Sun et al., 2020b)	Illumina short reads*, PacBio Sequel**, Hi-C***	Chromosome-level
Argyrosomus japonicus (Zhao et al., 2021)	Illumina short reads*, PacBio Sequel**, Hi-C***	Chromosome-level

*Genome survey and genomic base correction; **Genome assembly; ***Chromosome construction

2.6 Fatty acids

Fatty acids are building blocks of lipids, made up of an aliphatic chain with a carboxyl group (-COOH) at one end and methylene group at the other end that can be classified into four categories: saturated, monounsaturated, polyunsaturated, and *trans* fats (Rustan and Drevon, 2005;White, 2009). Saturated fatty acids (SFAs) have no double bonds of linear hydrocarbon chains (CH2), with one carboxyl group attached at the terminal end. A common source of SFAs can be derived from animals and plants (i.e. palm oil and coconut oil). High consumption of SFAs is not suitable for human health that can cause health problems such as raised cholesterol levels, cardiovascular disease, obesity, inflammation, and increased diabetes risk (Iggman and Risérus, 2011). However, a study using a tropical fish, barramundi, which is living in a habitat with high temperature, showed no effect on the growth of fish with a dietary supplement of SFAs, suggesting the SFAs might be able to utilize and digest for energy in some fish species (Salini et al., 2017).

Unsaturated fatty acids have one or more double bonds (C=C) in their carbon chains. Monounsaturated fatty acids (MUFAs) consist of only one double bond, existing in either cis (Z-configuration) or trans (E-configuration) configuration. In cis-configuration, the two adjacent hydrogens lie on the same side of the double-bonded carbons, whereas in trans configuration, the two adjacent hydrogens lie on opposite sides double-bonded carbons (Rustan and Drevon, 2005;Christie and Han, 2012). In nature, *cis*-MUFAs are the predominant form compared with *trans* and oleic acid (C18:1n-9), palmitoleic acid (C16:1n-7), and vaccenic acid (C18:1n-7) are common *cis*-MUFAs found in the dietary supplements. For daily nutrition, more MUFAs can obtain from vegetable oils, including olive oil, canola oil, hazelnut oil, mustard oil, almond oil and avocado oil (Schwingshackl and Hoffmann, 2014). Dietary