**UNIVERSITE DE LIEGE**
**FACULTE DE MEDECINE VETERINAIRE**
**DEPARTEMENT OF ANIMAL RESOURCES**
**BIOSTATISTICS AND BIOINFORMATICS**

# Ressources génomiques pour la sélection du silure requin (*Pangasianodon hypophthalmus*) en eau saumâtre dans le Delta du Mékong au Vietnam

# Genomic resources for the selection of striped Catfish (*Pangasianodon hypophthalmus*) under saline conditions in the Mekong Delta, Viet Nam

**DAO Minh Hai**

**THESE PRESENTEE EN VUE DE L'OBTENTION DU GRADE DE**
**Docteur en Sciences Vétérinaires**

**ANNEE ACADEMIQUE 2021-2022**

**Supervisor**

Frédéric FARNIR (ULiège, Belgium)

**Co-supervisor**

Yen DUONG THUY (CTU, Vietnam),

**Thesis committee**

Johann DETILLEUX (ULiège, Belgium)

Patrick KESTEMONT (UNamur, Belgium)

**Chairman of the examination board**

Arnaud SARTELET (ULiège, Belgium)

**Examination board**

François ALLAL (IFREMER, France)

Filip VOLCKAERT (KU LEUVEN, Belgium)

Jean-Luc HORNICK (ULiège, Belgium)

Mutien-Marie GARIGLIANY (ULiège, Belgium)

Tom DRUET (ULiège, Belgium),

Nicolas ANTOINE-MOUSSIAUX (ULiège, Belgium)

## ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my principal supervisor, Prof. *Frédéric FARNIR* (Ulg, Belgium). If he had not accepted me as a PhD student in Liege University, I would not have had the chance for study in Liege University. His guidance, solicitude, and encouragement helped me immensely in my scientific research in a new environment, and his generosity and patience in teaching me, his erudition, and his amiable personality made my studies and life much easier. My understanding of science has become much more comprehensive as a result of his teaching. He trained me to become a geneticist. I would also like to thank Assoc. Prof. *DUONG Thuy Yen* (CTU, Vietnam), my co-promoter in Vietnam for her considerable support during the planning, sample analysis. My study was operated through the financial assistance of ARES-CCD project "PANGAGEN: Towards a sustainable pangasius breeding, a selection approach". I would especially like to thank Prof. *NGUYEN Thanh Phuong,* Vietnam-project leader for guiding, supporting, encouraging during my experiment in Viet Nam. I acknowledge the members of my committee, Prof. Johann DETILLEUX (ULg, Belgium) and Prof. *Patrick KESTEMONT* (UNamur, Belgium) for their kind guidance and suggestion.

I appreciate enormously Prof. *Mutien-Marie GARIGLIANY* (ULg, Belgium) and the staffs of his lab for their kind supports during my experiments in Liege University. Another special thank goes to our project members Prof. *Nicolas ANTOINE-MOUSSLAUX* (ULg, Belgium) and team in CTU (Vietnam) Assoc. Prof. *DO Thi Thanh Huong*, Dr. *BUI Thi Bich Hang*, Dr. TRAN Minh Phu, Dr. *NGUYEN Quoc Thinh* , Assoc. Prof. PHAM Thanh Liem, Assoc. Prof. BUI Minh Tam, Mr. NGUYEN Hong Quyet Thang, Mr NGUYEN Hoang Trong Phuc, Mr *NGUYEN Quang Hieu*, Ms. NGUYEN Thi Ngoc Tran, Ms HUYNH Thi Kim Duyen, Mr NGUYEN Van Loc and PhD student *NGUYEN Le Anh Dao* for their support, encouragement and advice during my PhD work. Special words of thank also go to my friends in our lab: Dr *Evelyne* MOYSE, Ms *Marie* LEONARD*,* Ms *Caroline* DE BACKER, Dr *FIKREMARIAM Geda* for their friendliness and kindness. All of them make me feel the warmth of friendship without borders.

Last but not least, I would exceptionally love to express my deepest gratitude and appreciation to my great family, my parents and parent in-laws, my wife, my son and my daughter, who always save unconditional affection and encouragement for myself.

# Abbreviations

| | |
|---|---|
| ASC | Aquaculture Stewardship Council |
| BAP | Best Aquaculture Practices |
| CNV | Copy number variants |
| DNA | Deoxyribonucleic acid |
| DWG | Daily weight gain |
| DPH | Days post-hatching |
| DWGS | Deep whole genome sequencing |
| FAO | Food and Agriculture Organization |
| FCR | Feed conversion ratio |
| GS | Genomic selection |
| HCG | Human chorionic gonadotrophin |
| LCA | Life cycle assessment |
| MAF | Minor Allelic Frequency |
| MAS | Markers-assisted selection |
| MRC | Mitochondria-rich cells |
| PE | Pair-end |
| PIT | Passive integrated transponders |
| PPT | Parts per thousand |
| QTL | Quantitative trait locus |
| RAS | Recirculating Aquaculture System |
| SGR | Specific growth rate |
| SLR | Sea level rise |
| SWGS | Shallow whole genome sequencing |
| TET | Ten Eleven translocation |
| USD | U.S Dollars |
| VietGAP | Vietnamese Good Agricultural Practices |
| GLOBALG.A.P. | Good Agricultural Practices |

# Table of Contents

**Summary……………………………………………………………………………………...…...1**

**Resumé ...…………………………………………………………...……………..………………5**

**Preamble………..………………………………………………...………………………………9**

**Chapter 1: Introduction………………...………………………………………………….…….11**

1. Overview of Mekong Delta, Viet Nam………….………………......……………………..……... 12

2. Impacts of saline water intrusion on Mekong Delta…………...…...………………………...…...15

3. Overview of the biology and aquaculture of striped catfish………………….……….……….....18

3.1 Biology and aquaculture of striped catfish….……………………………………...…..…18

3.2 Fillet and meat quality………..……………………..…………………………………......23

4. Impacts of saline water intrusion on striped catfish aquaculture in Mekong Delta………….....…24

4.1 Effects of salinity on striped catfish………....…………………………………………....24

4.2 Impacts of salinity intrusion on the striped catfish farming in coastal provinces of the Mekong Delta, Viet Nam……………………..………….………………………………………………..25

4.3 Adaptation to salinity intrusion of striped catfish farming……………………….…….27

5. Selection programs in aquaculture……..………………………...……………………………29

6. Selection program on striped catfish in Vietnam……………………......………………………...36

7. Parentage assignment in aquaculture…………….....…………………………………………37

8. Shallow whole genome sequencing for parentage assignment……………………….…………..40

9. The role of epigenetic phenomena in the improvement of aquatic species…..........………………42

10. The role of high-quality reference genomes in aquaculture……………....…………...…………45

**Chapter 2: Objectives…………………………………………………………………………....49**

1. General objectives……......………………...……………………………………………………50

2. Specific objectives…………………......…………………………………………………………50

**Chapter 3: Parentage assignment using shallow whole genome sequencing data……………………53**

**Chapter 4: Selective breeding of saline-tolerant striped catfish (*Pangasianodon hypophthalmus*) for sustainable catfish farming in climate vulnerable Mekong Delta, Vietnam …………………….………87**

**Chapter 5: A high-quality genome assembly of striped catfish (*Pangasianodon hypophthalmus*) based on highly accurate long-read HiFi sequencing data………………………………………………...…111**

**Chapter 6: DNA reprogramming and modification under a salinity selection programme during early development in striped catfish (*Pangasianodon hypophthalmus*)…..…………………………..129**

**Chapter 7: General discussion…………………………………………………………………....153**

1. Harnessing genomics in breeding program to support sustainable development of striped catfish aquaculture in Mekong Delta, Vietnam……………………………………………………………………..154
1.1 Improving the quality of current seed production of striped catfish………………..…..155
1.2 The potential for catfish farming in saline intrusion areas………………………………158
1.3 Reduction of environmental loads in striped catfish farming…………………………....160
2. Potential genetic adaptation of striped catfish to salinity………...……………………...………161
3. Epigenetic considerations for the improvement of salinity tolerance of striped catfish………...163
4. The future development of striped catfish aquaculture in Mekong Delta………..………………165
**Chapter 8: Conclusions and perspectives……………………………………………………......169**
1. Conclusions…………...…………………...……………………………………………….....170
2. Perspectives…………...…...……………………………………………………………….....171
**References……………………………………………………………………………...………………..175**
**Appendices…....…………………………………………………………………...………………..231**
1. Identification of parentage using shallow sequencing data (manual) of Chapter 3……………...232
2. Supplementary material for Chapter 4……………...………………………....…………………257
3. Supplementary materials of Chapter 5……………………………………...………..............266

# Summary

Striped catfish (*Pangasianodon hypophthalmus*) is a commercially important species cultured in the Mekong Delta region in Southern Vietnam. Farming of striped catfish is facing significant challenges, including salinity intrusion as a result of climatic changes. Given the increasing salinity in the Mekong Delta region, several options were suggested to minimize the impacts of salinity intrusion such as changing striped catfish farming practice, shifting to another species, in which selecting striped catfish strain tolerant to salinityis considered the most effective option to maintain an economically sustainable catfish production in this region. Therefore, the main purpose of my PhD was to implement a genetic selection program to develop a new strain of striped catfish better adapted to saline environment. As part of the overall aim, several relevant research projects have been carried out to support the selection program and supply basic knowledge for more understanding on mechanism of salinity tolerance as well as genomic resource for further studies on striped catfish. The research consisted of four main sections.

Pedigree information is important in estimating genetic parameter and controlling inbreeding in selective breeding programs and hatchery management. Several previous studies on parentage assignment have used genetic data from SNP arrays. Obtaining genomic information from SNPs arrays is expensive, laborious and necessitates specific laboratory equipment. In addition, genomic information for developing SNPs arrays is not always available, as was the case for example for striped catfish. For this reason, in order to support the selection program on striped catfish, we have developed a new parentage assignment algorithm based on a likelihood approach that can use directly shallow whole genome sequencing (SWGS) data. The SWGS is similar to deep whole genome sequencing, but with much lower sequencing depth to reduce the cost. After testing our approach using simulation data, we have used the algorithm on empirical data from a striped catfish full factorial design involving 60 parents (30 males and 30 females) and 500 offspring. Around 21,000 high quality SNPs have been extracted from 5,900,000 genomic variants (obtained from mapping genomic data from the 60 parents on a draft genome) and used to infer the 500 offspring parents. As a result, 470 striped catfish were significantly allocated to one couple of parents. Our results show that using SWGS data can enable to generate highly accurate pedigree information using an appropriate algorithm. A standalone program called "Shallowped", has been written to implement this algorithm, and is now available for public use ( https://doi.org/10.5281/zenodo.6033536)

To implement the genetic selection program, we have first collected striped catfish broodstock from three different hatcheries located in three different provinces in Mekong Delta, Vietnam, with 10 unrelated males and 10 unrelated females from each hatchery (for a total of 60 breeders). Then, we have obtained the G1 generation from a full factorial cross of the broodstock, 30 females being crossed with 30 males to create

900 half and full-sib families. The larvae from these families have been reared in earthen ponds under freshwater conditions. After 47 days post-hatching (dph), the fry from all families were transferred from nursing ponds to a recirculating aquaculture system (RAS) to start the selection process under saline stress. In the RAS system, the selection program has been implemented through three sequential upward mass selections of approximately 50% of the fast-growing fish at three stages (148, 237, and 340 dph) under the targeted salinity of 10 ppt. In parallel, another set of fish has been reared under similar conditions (feed and biomass), and underwent a random selection process to serve as a control population for the selected group. After selection, all remaining fish have been identified using PIT tags and the corresponding DNA samples have been collected. This information has been used to identify the parentage relationships of all fish, using our newly developed program "Shallowped", to avoid inbreeding when producing the next generation (G2). To evaluate the effectiveness of selection after one generation in saline condition, several tests have been performed on the G2 generation. As a result, after one generation, average direct response to selection for growth was 18% and a moderate realized heritability (0.29) was estimated for body weight. These results indicate that our selection program succeeded in developing a productive strain of striped catfish with better tolerance to salinity.

With the increasing attention to epigenetic knowledge in the improvement of aquaculture production, an improved understanding of the epigenetic mechanisms in striped catfish should allow a better evaluation of the relationship between the phenotype and the external environment, as well as how a phenotype is generated and maintained. For this reason, an aim of this PhD thesis was to study DNA reprogramming in striped catfish during early development through the modifications in the DNA methylation pattern. Additionally, we have examined the effects of salinity and salinity tolerance-based selection program on the genes involved in DNA methylation. The results show that DNA reprogramming is dynamic during early striped catfish development and is likely regulated by changes in *dnmt*, *mecp2*, and *tet3* expression. Moreover, the significant increase demonstrated in *dnmt* genes expression of 6 dph larvae incubated at 2.5 ppt suggests an important role of the saline environment on the methylation process during early development of striped catfish. Furthermore, the findings from this study suggest that selective breeding program to salinity tolerance induces alterations in methylation levels, perhaps accounting for the regulation of growth genes and the growth performance in the offspring reared under favourable conditions.

The previous results suggest further research to increase our understanding of the physiology and genomic characteristics of tolerance salinity of striped catfish as well as new approaches for the improvement of tolerance salinity via breeding programs such as marker assisted selection (MAS) or genomic selection (GS) for the selection of the next generations. However, limited genetic and genomic resources for striped catfish at present time may restrict these studies. For example, a high-quality reference

genome could greatly facilitate fundamental research and application for genetic improvement programs in aquatic species in general, and in striped catfish in particular. For this reason, we have constructed a high-quality reference genome for striped catfish based on HiFi long reads. The resulting assembly is chromosome scale (30 chromosome groups) with a total length of 785.4 Mb across 381 scaffolds and a scaffold N50 length of 21.8 Mb. We expect that this chromosome level genome will promote fundamental studies not only on tolerance salinity but also for other important traits of striped catfish in the future.

**Resumé**

Le silure requin (*Pangasianodon hypophthalmus*) est une espèce commercialement importante cultivée dans la région du Delta du Mékong, dans le sud du Vietnam. L'élevage du pangasius est confronté à des défis importants, notamment en raison de l'intrusion de la salinité résultant des changements climatiques. Compte tenu de l'augmentation de la salinité dans la région du Delta du Mékong, la sélection de souches de pangasius tolérantes à la salinité est considérée comme une option pour maintenir une production de pangasius économiquement durable dans cette région. Par conséquent, l'objectif principal de mon doctorat était de mettre en œuvre un programme de sélection génétique pour développer une nouvelle souche de pangasius mieux adaptée à l'environnement salin. Dans le cadre de cet objectif global, plusieurs projets de recherche ont été menés pour soutenir le programme de sélection et fournir des connaissances de base pour mieux comprendre le mécanisme de la tolérance à la salinité ainsi que des ressources génomiques pour des études ultérieures sur le pangasius. La recherche a consisté en quatre sections principales.

Les informations généalogiques sont importantes pour estimer les paramètres génétiques et contrôler la consanguinité dans les programmes de sélection et la gestion des écloseries. Plusieurs études antérieures sur l'attribution de la filiation ont utilisé des données génétiques provenant de puces à SNP. L'obtention d'informations génomiques à partir de ces puces est coûteuse, laborieuse et nécessite un équipement de laboratoire spécifique. De plus, l'information génomique pour le développement de ces puces n'est pas toujours disponible, comme c'était le cas par exemple pour le pangasius. Pour cette raison, afin de soutenir le programme de sélection sur le pangasius, nous avons développé un nouvel algorithme d'attribution de parenté basé sur une approche de vraisemblance qui peut utiliser directement les données de séquençage du génome entier à faible profondeur (SWGS). Le SWGS est similaire au séquençage profond du génome entier, mais avec une profondeur de séquençage beaucoup plus faible pour en réduire le coût. Après avoir testé notre approche en utilisant des données de simulation, nous avons utilisé l'algorithme sur des données empiriques provenant d'un croisement factoriel complet impliquant 60 parents (30 mâles et 30 femelles) et 500 descendants. Environ 21 000 SNP de haute qualité ont été extraits de 5 900 000 variants génomiques (obtenus à partir de la cartographie des données génomiques des 60 parents sur un projet de génome) et utilisés pour inférer les 500 parents de la descendance. En conséquence, 470 poissons-chats rayés ont été attribués de manière significative à un couple de parents. Nos résultats montrent que l'utilisation des données SWGS peut permettre de générer des informations généalogiques très précises en utilisant un algorithme approprié. Un programme autonome appelé "Shallowped", a été écrit pour mettre en œuvre cet algorithme, et est maintenant disponible pour un usage public
( https://doi.org/10.5281/zenodo.6033536)

Pour mettre en œuvre le programme de sélection génétique, nous avons d'abord collecté des géniteurs de pangasius dans trois écloseries différentes situées dans trois provinces différentes du Delta du Mékong, au Vietnam, avec 10 mâles non apparentés et 10 femelles non apparentées dans chaque écloserie (pour un total de 60 reproducteurs). Ensuite, nous avons obtenu la génération G1 à partir d'un croisement factoriel complet du stock de géniteurs, 30 femelles étant croisées avec 30 mâles pour créer 900 familles de demi- et de plein-frères. Les larves de ces familles ont été élevées dans des étangs dans des conditions d'eau douce. Après 47 jours après l'éclosion (dph), les alevins de toutes les familles ont été transférés des étangs d'élevage vers un système d'aquaculture avec eau circulante (RAS) pour commencer le processus de sélection sous stress salin. Dans le système RAS, le programme de sélection a été mis en œuvre par trois sélections phénotypiques séquentielles d'environ 50 % des poissons à croissance rapide à trois stades (148, 237 et 340 dph) sous la salinité ciblée de 10 ppt. En parallèle, un autre ensemble de poissons a été élevé dans des conditions similaires (alimentation et biomasse), et a subi un processus de sélection aléatoire pour servir de population témoin pour le groupe sélectionné. Après la sélection, tous les poissons restants ont été identifiés à l'aide de transpondeurs (PIT tags) et les échantillons d'ADN correspondants ont été collectés. Ces informations ont été utilisées pour identifier les liens de parenté de tous les poissons, en utilisant notre nouveau programme "Shallowped", afin d'éviter la consanguinité lors de la production de la génération suivante (G2). Pour évaluer l'efficacité de la sélection après une génération en condition saline, plusieurs tests ont été effectués sur la génération G2. En conséquence, après une génération, la réponse directe moyenne à la sélection pour la croissance était de 18% et une héritabilité réalisée modérée (0,29) a été estimée pour le poids corporel. Ces résultats indiquent que notre programme de sélection a réussi à développer une souche productive de pangasius avec une meilleure tolérance à la salinité.

Avec l'attention croissante portée aux connaissances épigénétiques dans l'amélioration de la production aquacole, une meilleure compréhension des mécanismes épigénétiques chez le pangasius devrait permettre une meilleure évaluation de la relation entre le phénotype et l'environnement externe, ainsi que la façon dont un phénotype est généré et maintenu. Pour cette raison, un des objectifs de cette thèse était d'étudier la reprogrammation de l'ADN chez le pangasius au cours du développement précoce par le biais des modifications du modèle de méthylation de l'ADN. De plus, nous avons examiné les effets de la salinité et du programme de sélection basé sur la tolérance à la salinité sur les gènes impliqués dans la méthylation de l'ADN. Les résultats montrent que la reprogrammation de l'ADN est dynamique pendant le développement précoce du pangasius et qu'elle est probablement régulée par des changements dans l'expression de dnmt, mecp2 et tet3. De plus, l'augmentation significative démontrée dans l'expression des gènes dnmt des larves de 6 dph incubées à 2,5 ppt suggère un rôle important de l'environnement salin sur le processus de méthylation pendant le développement précoce du pangasius. En outre, les résultats de cette

étude suggèrent que le programme d'élevage sélectif pour la tolérance à la salinité induit des altérations dans les niveaux de méthylation, expliquant peut-être la régulation des gènes de croissance et la performance de croissance dans la progéniture élevée dans des conditions favorables.

Les résultats précédents suggèrent des recherches supplémentaires afin d'accroître notre compréhension de la physiologie et des caractéristiques génomiques de la tolérance à la salinité du pangasius, ainsi que de nouvelles approches pour l'amélioration de la tolérance à la salinité via des programmes d'élevage tels que la sélection assistée par marqueurs (MAS) ou la sélection génomique (GS) pour la sélection des prochaines générations. Cependant, les ressources génétiques et génomiques limitées pour le pangasius à l'heure actuelle peuvent restreindre ces études. Par exemple, un génome de référence de haute qualité pourrait grandement faciliter la recherche fondamentale et l'application des programmes d'amélioration génétique des espèces aquatiques en général, et du pangasius en particulier. Pour cette raison, nous avons construit un génome de référence de haute qualité pour le pangasius basé sur des lectures longues HiFi. L'assemblage résultant est à l'échelle des chromosomes (30 groupes de chromosomes) avec une longueur totale de 785,4 Mb à travers 381 échafaudages et une longueur N50 des échafaudages de 21,8 Mb. Nous espérons que ce génome à l'échelle chromosomique favorisera les études fondamentales non seulement sur la tolérance à la salinité mais aussi pour d'autres traits importants du pangasius à l'avenir.

# Preamble

Striped catfish (*Pangasianodon hypophthalmus*) is currently the most economically important freshwater indigenous species in aquaculture sector in the Mekong Delta, Vietnam and in other Asian countries. Benefits related to optimal climatic conditions, a complex network of rivers and channels across the area and the development of an efficient technology for artificial seed production, have allowed striped catfish culture to develop rapidly in the Mekong Delta.

However, striped catfish aquaculture in the Mekong Delta is now facing significant challenges. These include changes in the flood regime due to the construction of upstream dams, the eutrophication of the environment due to non-sustainable aquaculture practices, theenvironmental pollution due to human activities (including aquaculture and agriculture), the appearance of new markets... Another major concern is the salinity intrusion resulting from climate changes. In Vietnam, the Mekong Delta is considered one of the most vulnerable places in the worlds. Forecasts show an increase of an expanded salinity intrusion in the dry season both at spatial and temporal scales in the Mekong Delta in the future due to subsidence and rising sea levels. Striped catfish culture is largely dependent on the availability of abundant water supply from the Mekong river. Therefore, the increasing salinization of freshwater areas will negatively impact striped catfish culture, which is a freshwater species. Given the situation of increasing salinity in the region, coupled with the requirement of new expansions to the traditional catfish farming areas to meet the growing demand, selecting a striped catfish strain tolerant to salinity is considered a promising approach to maintain sustainable catfish aquaculture in the region in the current changing environmental context.

In selective breeding, pedigree information plays an important role. However, maintaining pedigree information is one of the challenges in aquaculture. Genetic information generated from a SNP array for parentage assignment has been implemented on several species. However, this approach requires considerable labor, time and cost for developing the SNP array, and dedicated equipment to exploit the chips. Hence such approach is a big challenge for the striped catfish industry because a SNP array is not currently available. Moreover, although there were several studies investigating the effect of salinity on growth performances of striped catfish, the molecular mechanisms underpinning the salinity tolerance of striped catfish is not well understood. In addition, the lack of a high-quality reference genome may hinder research concerning the genome and impact significantly breeding progress.

For these reasons, the main goal in the present study was to carry out a genetic selection program targeting fish showing improved adaptation to salinity. Fish better adapted to salinity will be selected

through a sequential upward genetic selection program. We have developed a new parentage assignment algorithm based on a likelihood approach allowing to use directly low depth sequencing data to support this selection program. We have tested this algorithm on simulation data first, and then used it on real data. Besides, to better understand the mechanism underlying the variable salinity tolerance of striped catfish, we have designed some experiments to evaluate the role the epigenetic factors in the salinity tolerance of striped catfish. In parallel, we have assembled a high-quality genome of striped catfish using HiFi sequencing data. The novel knowledge and genomic (sequencing) information can be used to identify molecular DNA markers associated with economically important traits, including salinity tolerance. This should facilitate genetic improvement of striped catfish through marker assisted selection and genomic selection in the future. Moreover, this information will contribute to international databases to increase the scientific knowledge of striped catfish.

# Chapter 1

# Introduction

# Chapter 1: Introduction

## 1. Overview of Mekong Delta, Viet Nam

The Mekong River is recognized as one of the few great, largely unregulated rivers of the world; it is the 12$^{th}$ biggest river worldwide. Its delta is both agriculturally and aquatically highly productive and a major contributor to the region's food production and export earnings (Mathieu et al., 2021; White, 2002). The Mekong Delta region in southwestern Vietnam (Figure 1) is home to over 20 million people, and is also a unique natural habitat, teeming with thousands of unique plant and animal species. It is recognized as one of the largest populous deltas in the world and is considered the "rice bowl" of Southeast Asia, supplying 50% of the Vietnamese food (MRC, 2010). A majority of people in the Mekong Delta rely on the riverine aquatic resources and rice production for their subsistence. For many, 40 to 60% of their protein intake is from fish from the Mekong, and the prodigious fish resources rely on the annual flooding of the Mekong (Sebesvari et al., 2011; White, 2002). Although, economically, the delta is very important for the country as a whole, the region remains one of the poorest when compared to other regions in Vietnam. In spite of the rapid economic growth of Vietnam in recent years and important improvement in agricultural systems in the region, many farmers in the delta have to deal with low profitability and high economic and environmental risks forcing them into insecure livelihoods (Renaud et al., 2012). Moreover, the Mekong Delta is facing several emerging issues including dams construction, climate change, poor water quality, delta subsidence, and loss of biodiversity (Smajgl et al., 2015 ; Mackay and Russell, 2011; Mathieu et al., 2021).

The rapid economic development of Southeast Asian countries has increased the use of natural resources and is now accompanied by a high demand in electricity. Consequently, a large number of dams were built by the governments bordering of the Mekong Basin (Kuenzer et al., 2013). In total, 176 dams are now being built or planned through the Mekong River Basin (Nhan and Cao, 2019). These dams have considerable impacts on the ecosystem and on human life. One of the first impacts is the changes of the sediment load as a dam traps up to 95% of sediments and consequently, reduces the amount of sediment downstream. Reports from MONRE, (2016) indicated that the future of the Mekong Delta seems to face a dramatic loss of sediment load into the delta of about 16–40 Mt./yr, that is around 10–25% of the pre-dam sediment load of 160 Mt./yr. The changing hydrology of the Mekong Delta river is another consequence of the dams. Hydro-power operations have already altered the natural hydrological cycle with lower flood peaks and higher dry season flows (Rasanen et al., 2012; Hoanh et al., 2010).

The Mekong estuary grew rapidly due to high sediment supply transforming it into a delta from 5,300 to 3,500 years ago. However, the past few years (from 2005), the area is experiencing considerable coastal erosion According to Anthony et al. (2015), between 2003 and 2012, the erosion affected more than 50% of the 600 km long delta shoreline. The sediment trapping by dams and sand mining for construction is clearly identified as the main risk enhance coastal erosion; other reasons may be largely attributed to processes such as natural shore sediment redistribution by the ocean forces (waves, currents and tides), mangrove destruction, sea level rise and/or subsidence (Mathieu et al., 2021). It has been reported that subsidence exacerbates the delta sink and associated risks of coastal erosion, extended duration of flooding and salinization (Minderhoud et al., 2019; Erban et al., 2014; Erkens et al., 2015).

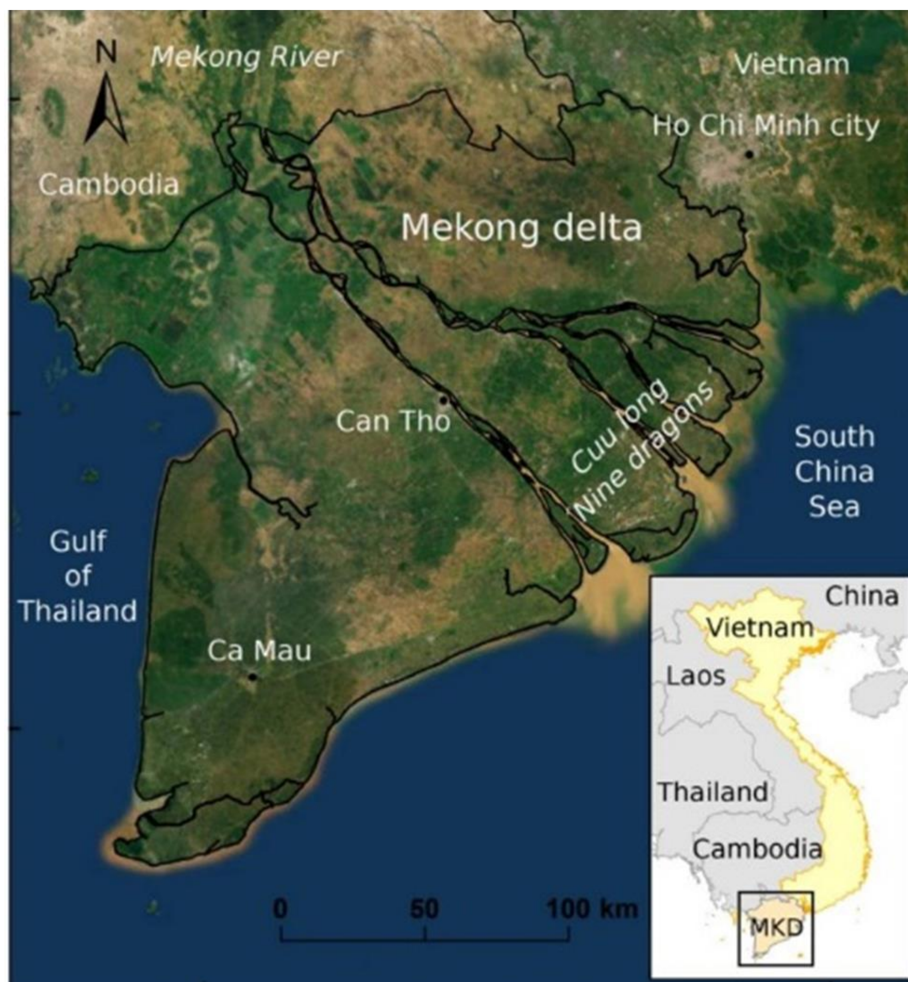


Figure 1: Satellite image of the Mekong Delta in Vietnam (https://www.preventionweb.net/new)

Water quality degradation is also another concern for the Mekong Delta. Dams are considered one of the reasons that lead to eutrophication and pollution of the water resources locally and downstream, as a result of the regulated water flow and water stagnation (Li et al., 2013). The other sources of water pollution come from human activities such as the increasing amount of industrial, agricultural and urban wastewater (Snidvongs and Teng, 2006; Swain., et al. 2007). The main agricultural activities linked to water pollution in the Mekong Delta are land preparation (e.g. puddling), fertilizer and pesticide use, waste water release from animal production (mainly pig and poultry) and aquaculture (mainly fish and shrimp) (Nhan et al., 2007; Renaud et al., 2012)

The Mekong River is recognized as the second most biodiverse river in the world (after the Amazon River). In the Mekong Delta Basin, 877 fish species have been identified (not including estuary marine species) (Ziv et al., 2012). Unfortunately, the delta is changing from a sparsely populated free flowing riffle pool river to a densely populated basin with slow flowing river, with devastating impact on the biodiversity. Threats to the biodiversity of the delta include growing human populations and intensification of agriculture, with increasing use of fertilizers and pesticides and altering flood water levels during the high flow season (Campbell, 2012). Moreover, the Mekong River system is severely threatened by hydroelectric dams too fulfill the growing needs for energy in the region. These artefacts could cause ecosystem to collapse and biodiversity loss, devastating freshwater species and the communities that rely on the rivers to live. The dams also hinder movements of migratory fish up and down the river, reduce the level of sediments and nutrients that are  the natural resource for biodiversity of this region and the building blocks of the delta’s productivity (Renaud et al., 2012; Mathieu et al., 2021; Ziv et al., 2012).

The Mekong Delta is one of the most vulnerable places on earth to the impacts of climate change, causing existing and projected threats (Hai et al., 2020; Mackay and Russell, 2011; IPCC, 2007). Global warming is the main indicator of climate change due to greenhouse gas emissions from human activities. Climate change results in a strong fluctuation in rainfall, an increase in weather and climate extremes such as floods and droughts, and rising sea levels that will also directly affect coastal areas, potentially inundating land or increasing salinity (Mackay and Russell, 2011). Climate change therefore affects the region’s people, the biodiversity and the natural resources. For instance, water scarcity leads to reduced agricultural productivity, leading to food scarcity, unemployment and poverty (Renaud et al., 2012). Across the region, temperatures have risen and continue to rise. Sea level rise is threatening the region’s coastal communities. Degradation of ecosystems from climate change will decrease their productivity and capacity to provide livelihoods for people of the Mekong Delta (Campbell, 2012; Tuan and Chinvanno, 2011; Wassmann et al., 2004). The Mekong Delta is flat and low-lying with a maximum elevation of less than 4.0 m above

mean sea level, making it one of the 3 most vulnerable deltas in the world to sea level rise (Mackay and Russell, 2011; Sebesvari et al., 2011).

**2. Impacts of saline water intrusion on Mekong Delta**

The Mekong Delta is projected to be heavily affected by salinity intrusion as a result of climate change (IPCC, 2007). Sea level rise over the next decades will increase salinity intrusion further upstream of the Mekong Delta and consequently affect freshwater culture, including the striped catfish culture industry. The Mekong Delta is one of the three large low topographical regions in Vietnam, with an area of 4.06 million ha. It comprises a complex network of rivers and channels (Sebesvari et al., 2011) and has a total freshwater area of 641,350 ha, which comprises 67.2% of the total water surface (Nguyen and Dang, 2010). The Mekong Delta has been predicted to be among the top three most seriously affected regions around the world, and seawater level was also predicted to raise of 1 m in this century (IPCC, 2007; Thi et al., 2015). If the predictions of future climate change are realized, a 1 m sea level rise will result in approximately 1,000 $km^2$ of cultivated land and farming area in Vietnam becoming salt marshland. In addition, 15,000 to 20,000 $km^2$ of the Mekong Delta is likely to be inundated by seawater, with a loss of 76% of arable land (De Silva and Soto, 2009) (Figure 2).

As mentioned earlier, the numerous dams on the main stream and its tributaries of the Mekong River, constructed to respond to the high demand for hydropower of upstream countries, also may play an important role in the salinity increase of the Mekong Delta (Thi et al., 2015). These structures are believed to reduce the water inflow in the delta that may lead to increase in salinity intrusion in the Mekong Delta (Smajgl et al., 2015). Until now, few published studies document the environmental impact of upstream dams on the salinity of the Mekong Delta. However, according to Lu and Siew (2006), the water discharge regime of the Mekong River has been influenced by the construction of the Manwan dam in the upper stream (China), although the extent of influence remains small and the seasonal discharge regime has largely remained within the historical range. Nevertheless, four mega dams are planned within China and larger dams are under construction in Laos and Cambodia. Therefore, further studies are needed to examine the potential impact of these dams on freshwater inflow as well as on salinity in the Mekong Delta (Thi et al., 2015).

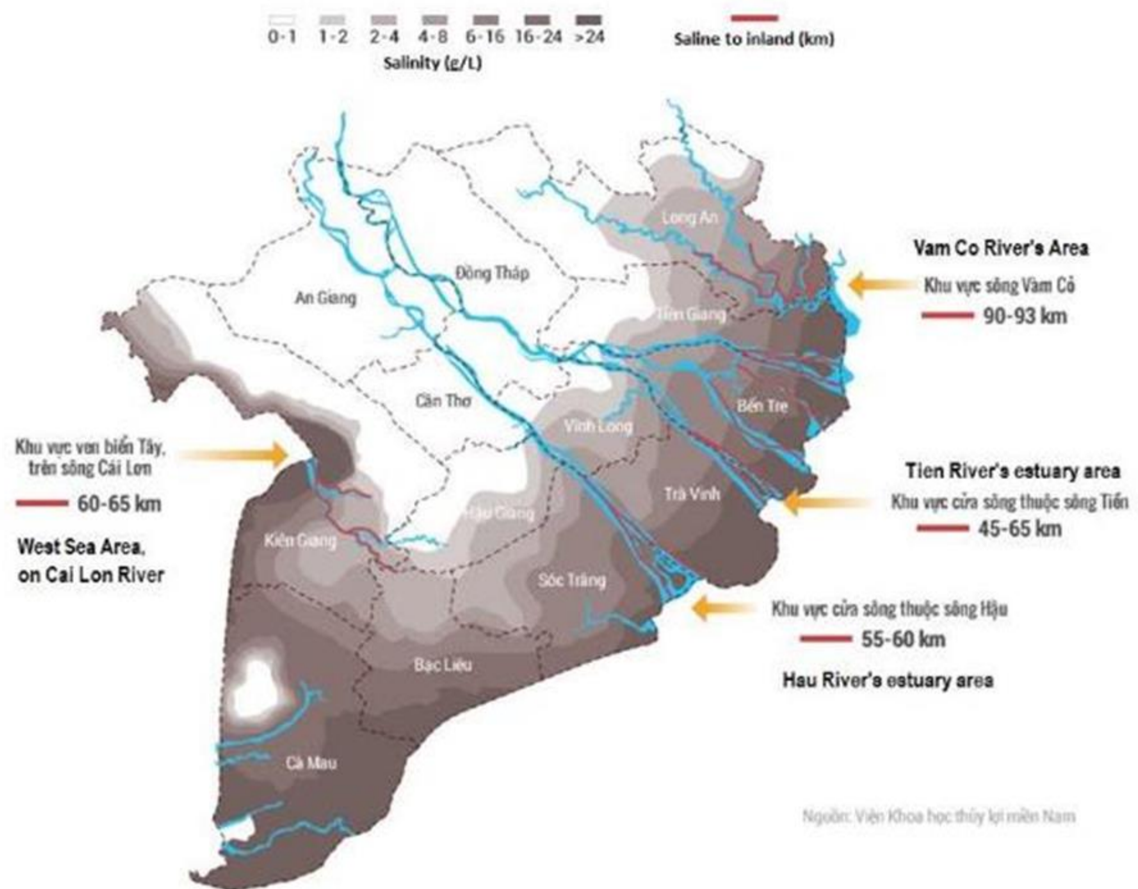


Figure 2 Saline water intrusion in the Mekong Delta in 2016
(http://www.tongcucthuyloi.gov.vn/, accessed on March 24, 2016)

The sea level around the Mekong River Delta has also risen by 20 cm since 1901, with an average sea level rise of 3 mm per year over the last 30 years (CGIAR, 2016). Saline intrusion is reaching further land and affecting wider areas of the region (Figure 2). The increasing salinity levels in the Mekong Delta have substantially reduced agricultural productivity and caused declining rice production, which is contributing to Vietnamese national food security (Smajgl et al., 2015), in particular for crops and varieties with a low tolerance to salt (Le Dang et al., 2014; Wassmann et al., 2004) (Figure 3). However, the salinity intrusions is positive for shrimp aquaculture, because it could provide additional areas for shrimp farming (Hai et al., 2020). Shrimp aquaculture provides key livelihoods widely adopted by coastal communities in response to increasing salinity levels. The results from responses of 1,265 randomly sampled households indicated approximately 50% increase of annual household income using the shrimp–rice rotation system (USD 1,735) as compared to a two-rice-crop system (USD 1,019) (Smajgl et al., 2015). Although, shrimp farming is a much more highly valued commodity than many agriculture products and has greater market

potential, it is associated with high production risks due to diseases outbreaks, especially in intensive shrimp culture with high level of salinity (Aranguren Caro et al., 2021; Ha et al., 2013; Son et al., 2011).

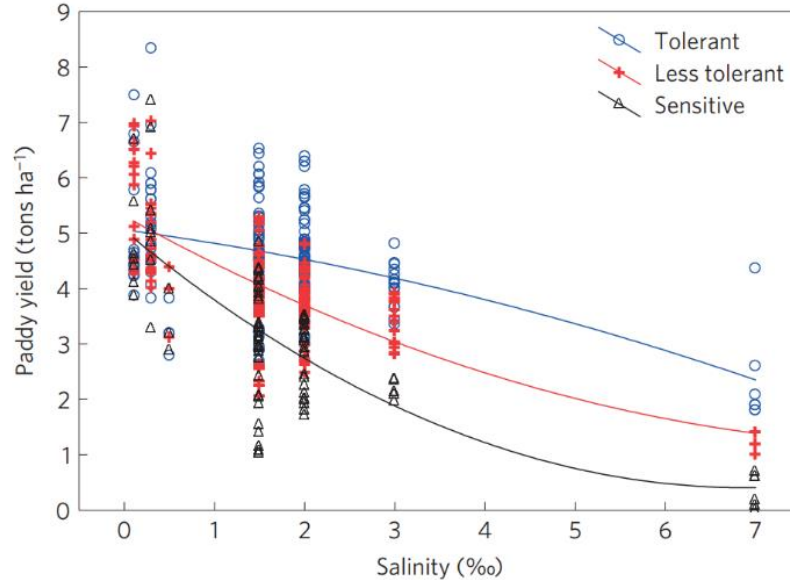


Figure 3: Relationship between soil salinity and yields of rice varieties of variable salinity tolerances based on multiple field trials in 2005–2010 (Smajgl et al., 2015).

The greatest long-term threat for the Mekong Delta is undoubtedly climate change. The increase in sea level into the Delta has very serious consequences not only for the biodiversity of the Delta, but also for the economy of Viet Nam (Campbell, 2012). In 2016, the Mekong Delta faced serious problems of drought and saline water intrusion. In that year, the water discharge from upstream river declined by 900 $m^3/s$. Meanwhile, the water level in the middle and downstream Mekong River increased by 0.1 - 1.5 m due to tidal rise. Therefore, salinity intrusion led to an increased salinity of water in rivers and irrigation canals. Salinity (4 ppt) expanded deeply through Tien and Hau rivers by up to 45-65 km and 55-60 km, respectively (Figure 2). The issue strongly affected 11 of the 13 provinces in the Mekong Delta, causing considerable damage to aquaculture (194.163 ha), rice (405.000 ha) and fruit tree production (28.457 ha), with economic loss of approximately 0.34 billion USD (VAWR, 2017).

## 3. Overview of the biology and aquaculture of striped catfish

### 3.1 Biology and aquaculture of striped catfish

Striped catfish (*Pangasianodon hypophthalmus* (Sauvage, 1878); Pangasiidae; Siluriformes; Vietnamese: cá tra; Khmer: trey pra; French: silure requin) is a freshwater siluriform (Figure 4). It is commonly cultured in the Mekong Delta region, Southern Vietnam, because of the favourable environment, simple grow-out techniques, the availability of seeds and feeds, reliable logistics, a well developed whole sale network, fish packing plants, access to export markets as well as governmental support (De Silva and Phuong, 2011; MARD, 2013). It has developed from a species farmed in small ponds using wild-caught seed to large-scale commercial culture followed in cages, pens and ponds commencing with the development of artificial mass seed production (Nguyen et al., 2013). Critical in the expansion has been the closure of the reproductive cycle (Legendre et al. 2000). Striped catfish in the Mekong Delta, Vietnam, is one of the fastest growing species used in aquaculture. In recent decades, striped catfish has been translocated for aquaculture to Bangladesh, India, China, Indonesia, Myanmar, Nepal, Pakistan and many more countries (FAO, 2019).

Figure 4: Striped catfish (*Pangasianodon hypophthalmus)* (Nguyen and Dang, 2010)

Striped catfish migrates upstream the Mekong River at the end of flooding season (from October to February), and their offspring returns to the mainstream at the beginning of the rainy season (from June to August) (Nguyen and Dang, 2010). The striped catfish life-cycle is intimately tied to the annual monsoon flood cycle, with spawning in May-June at the start of the monsoon season. The spawning ground of striped

catfish is upstream of the Mekong River Delta, more specifically, below the Khone Falls on the Laos-Cambodia border (Sopha et al. 2002). Striped catfish spawn at the beginning of the rainy season and the adhesive eggs are deposited on roots of aquatic macrophytes and other substrates. The newly hatched larvae drift downstream and are swept into the floodplain in southern Cambodia, Tonle Sap and the Mekong Delta (Nguyen et al., 2013). Striped catfish is natively distributed in the Mekong and Chao Phraya rivers in the Lower Mekong basin (Rainboth, 1996; Roberts and Vidthayanon, 1991), involving four countries: Thailand, Laos, Cambodia and Vietnam. Domestication of this species started in 1967 in Thailand (Na-Nakorn and Moeikum, 2009) and in 1979 in Vietnam (Trong et al., 2002). Commercial production, however, has taken off recently, especially in Vietnam since 1999 (Trong et al., 2002). According to So et al. (2006), genetically differentiated populations of striped catfish coexisted in the Mekong River, with notably different spawning times. In Thailand, Na-Nakorn and Moeikum (2009) reported insignificant genetic differentiation of broodstock between hatcheries and wild striped catfish. Likewise, Ha and colleagues (2009) found no difference between wild versus hatcheries stocks in Vietnam. Recent study showed however two distinct populations in the lower Mekong river: (i) the wild stock from the upper part of lower Mekong river in Thailand and (ii) the other populations from Cambodia and Vietnam in the Mekong river (Vu et al., 2020). In addition, both farmed and wild stocks from these countries showed reasonable level of genetic diversity. The close genetic relationship between wild and farmed stocks in Cambodia and Vietnam (maybe due to the recent domestication of striped catfish (Nguyen and Dang, 2010 ; Vu et al., 2019)) indicated that the small changes observed in farmed stocks in these countries were unlikely to be associated to a deterioration of the genetic potential of the farmed stocks with respect to the wild animals (Vu et al., 2020). These results pointed out that the captive stock in Vietnam can be used as an alternative to the wild to produce seeds for commercial production.

In the wild, striped catfish can reach a maximum length of about 90 cm and achieve a maximum weight of approximately 44 kg (Nguyen, 2015). Under pond conditions with commercial feed, striped catfish were harvested at the size of approximately 1 kg after a growth period of 6-7 months (Phan et al., 2009). Striped catfish is omnivorous; its feeding habit is based on on algae, higher plants, zooplankton, and insects, while larger specimens also take fruit, crustaceans and small fish (Rainboth, 1996). Mature fish can reach an age of three years in the wild or more in captivity, with a minimum weight and length of 4 kg and 54 cm at 1st maturity, respectively. The fish is highly fecund, a female of 10 kg could produce 1 kg of eggs (approx. 1.2 million eggs) (Poulsen and Valbo-Jørgensen, 2000). Striped catfish is a very tolerant species that adapts to a wide range of environments and this characteristic allows it becoming one of the most important species in global aquaculture. First of all, striped catfish possess an air-breathing capacity that

allows fish to withstand low levels of dissolved oxygen in water at certain times of their life (Lefevre et al., 2011). In addition, according to Phuc et al. (2017), striped catfish can grow well in a variety of temperatures from 25̊C to 35̊C, with even better growth at 35̊C. Moreover, the species can highly tolerate environmental nitrite concentration with a 50% lethal dose reaching 1.95 mM after 96h (Graham and Wegner, 2010). Finally, the striped catfish can adapt to different food sources and types, such as homemade food formulated from agricultural and fishery by-products, pellets, fish waste, and animal and human wastes due to its omnivorism (Graham and Wegner 2010). The chromosome number of striped catfish is n=30 (Sreeputhorn et al., 2017) and a result on study of Wen et al. (2022) suggested that striped catfish has a XY sex determination system.

Figure 5: Checking quality of eggs of striped catfish before breeding

Adequate quantities of high quality seed stocks play an important role for the development of striped catfish aquaculture in Vietnam. Initially, the seed stock for striped catfish farming was wild caught and of limited quantity and seasonal availability (Nguyen, 2009). In late 1978, researchers were initially successful in the artificial propagation of striped catfish, but the results were not sufficiently reliable for mass seed production and the program was discontinued. Research was restarted in 1995 with the

collaboration of Vietnamese with French and Belgian researchers under the EU funded project *Catfish Asia* to develop artificial propagation techniques for striped catfish (Legendre et al. 2000; Tuan et al., 2003). In 1996, induced spawning of striped catfish was established, and transferred to hatchery operators from 2000 onwards (Caco, 1998; Cacot et al., 2002). Since then, striped catfish production has expanded (Figure 5) and followed the industry demand. The success of the artificial seed production is one of the key reasons for the explosive growth of striped catfish industry in Vietnam (Nguyen et al., 2013).

Figure 6: Striped catfish pond in the Mekong Delta

The production chain of striped catfish industry in Vietnam can be subdivided into four main components including seed production or hatchery, nursery, grow-out, and processing (Nguyen and Dang, 2010). The hatchery sector produces large numbers of larvae, which are mostly sold to nursery farms, while the nurseries grow fingerlings for sale to grow-out farms. As of December 2019, there were 230 hatcheries and more than 4,000 nursery farms (covering 3,000 hectares). These farms were producing 21 billion larvae, and more than 2.1 billion fingerlings, to meet the regional seed demand, with a total production area of 6,600 ha (DOF, 2019). In the hatcheries, the mature broodstock (Figure 5) are induced to spawn using HCG

(Human Chorionic Gonadotropin) or a combination of HCG and pituitary gland extract. The injection can be two or three times in peak spawning season and three or four times in the off season. Total doses of HCG vary from 5500 to 6500 IU/kg of females. The males receive only one injection of 1000 IU/kg, coinciding with the time of the third injection of the females. Broodstock are spawned in single pairs or batches, and are usually dry stripped. The eggs are incubated in conical shaped jars made either of stainless steel or glass, with up-welling water flow to keep the eggs in suspension. Depending on water temperature, the eggs hatch usually within 22-24 h. Yolk sac absorption takes a further 24 h. The larvae are transferred to nursing ponds just prior to full yolk sac absorption (Bui et al., 2010). The nursing pond can be a part of the hatchery but, in most cases, larvae are nursed to fry and fingerling stages by a nursery sector separately from the hatcheries activities. The nursing stage lasts from 3 to 4 months and then fingerlings (with average size 25 g per individual) are transferred to growth-out ponds (Bui et al., 2010; Phan et al., 2009).

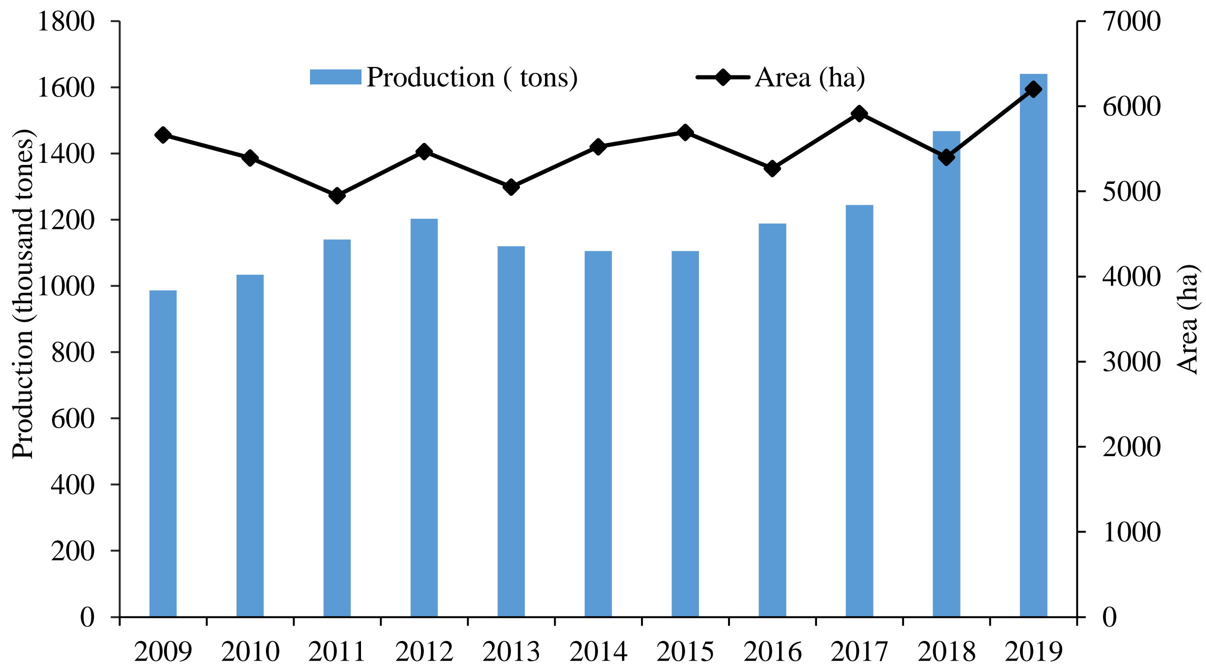


Figure 7: Production and culture area of striped catfish in Mekong Delta from 2009 to 2019

In the grow-out stage, the species is mostly farmed in earthen ponds (Figure 6), with pond water depths of 3- 4.5 m, in nine provinces, which are located adjacent to or near Hau River and Tien River, two branches of the Mekong River (Phan et al., 2009). Stocking density ranges from 18-125 fish $m^{-2}$ (average:

48 fish $m^{-2}$), depending on several factors such as the size and availability of fingerlings and the financial capacity of farmers. Farm size and yields ranged from 0.2 to 30 ha (average: 4.09) and 70 to 850 ton $ha^{-1}$ $crop^{-1}$ (average: 406), respectively (De Silva and Nguyen, 2011; Phan et al., 2009). Catfish culture is operated by individual farmers, horizontally and/or vertically integrated large companies, entrepreneurs or cooperatives. They are mostly certified for good aquaculture practices such as by VietGAP (Vietnamese Good Agricultural Practices) and GlobalGAP (Good Agricultural Practices), Global Aquaculture Alliance's Best Aquaculture Practices (BAP), and the Aquaculture Sustainability Council (ASC) (Hai et al., 2020). By the year 2018, the catfish production reached nearly 1.42 million tons (for a total production area of approximately 5,400 ha) and an export value of 2.26 billion USD (contributing to about 1% GDP of Vietnam). Striped catfish is exported to more than 140 countries and territories, with the main export markets including the EU, Russia, China and the USA (VASEP, 2019) (Figure 7). The sector provides direct employment to more than 180,000 people in Vietnam, with a vast majority of women, primarily in the processing sector (De Silva and Nguyen, 2011). Striped catfish is considered as one of the strategic products of Vietnam. In the conference on the establishment of national striped catfish brand for export, Vietnamese Ministry of fisheries emphasized that the genetic improvement of this species for some specific quality traits called for immediate actions in order to create a national brand (Sang et al., 2010).

### 3.2 Fillet and meat quality

The fillet of striped catfish contains a relatively high concentration of eicosapentaenoic acid (20:5n-3, EPA) and docosahexaenoic acid (22:6n-3, DHA) (Domiszewski et al., 2011; Men et al., 2005), fatty acids that have been identified as enhancing brain development and learning in children, while protecting vision and promoting eye health (Kwak et al., 2012). Striped catfish grown in Vietnam has rapidly become an affordable "white fish" substitute in the Western world, and, conceivably, its acceptability and popularity have been growing (Phan et al. (2009). According to Tartila et al. (2021), a preferable striped catfish fillet is known to have properties, such as a white or light color with a solid and hard texture and a total lipid content of 1– 3%. Vietnamese suchi catfish is mostly marketed in fresh seafood markets and supermarkets as frozen or thawed fillets (120–260 g). Frozen fillets (generally individually quick-frozen) with glazing are sold individually or packed according to size into 1 kg packages (Orban et al., 2008). Striped catfish fillets were characterised by high moisture levels (80–85%) and low protein (12.6–15.6%) and lipid (1.1–3.0%) contents. The total lipids of fillets contained low cholesterol levels (21–39 mg/100 g), high percentages of saturated fatty acids (41.1–47.8% of total fatty acid) and low percentages of polyunsaturated fatty acids (12.5–18.8% of total fatty acids). The latter were mainly represented by linoleic acid (44–59% of total polyunsaturated fatty acids) (Guimarães et al., 2016; Orban et al., 2008). A study of Bland et al.

(2018) indicated that texture and color play a massive role in the fish fillet quality following consumer preference. Therefore, to increase the marketability of striped catfish fillets, improvement of fillet properties (namely, texture, color, and lipid content) should be considered when developing the improved culture technique for striped catfish (Tartila et al., 2021).

## 4. Impacts of saline water intrusion on striped catfish aquaculture in Mekong Delta

### 4.1 Effects of salinity on striped catfish

Salinity is one of the critical factors affecting metabolism and distribution of fish species, as all fish need to osmoregulate and maintain osmotic difference between their extracellular fluid and either freshwater or seawater (Taugbøl et al., 2014). Fish are unique among vertebrates in that their body fluids directly contact the environment water across the thin respiratory epithelia. An ability to regulate body fluids independently of the external environment, both in freshwater and seawater is essential for fish survival (Takei and Balment, 2009). In most organisms, stress is a common physiological response to unfavorable condition or noxious stimuli. When they perceive unfavorable changes to their natural habitats, they activate a suite of defence mechanisms and often divert energy in their body away from growth to counteract the new environmental conditions in an attempt to maintain their homoeostatic equilibrium (Fuzzen et al., 2011). To a large extent, fish are protected from changes to their external environment by a thin skin with mucus secreted outside the skin layer. Moreover, individuals maintain homoeostatic equilibrium via a self-regulating control system and a suite of homoeostatic mechanisms including the respiratory, digestive and urinary systems (Takei and Balment, 2009), but efficient ionic regulation has an associated large energy cost. Between 20 and 50% of total energy budget of teleost fish is dedicated to a specific common response under stressful conditions whereby they increase plasma cortisol levels to regulate their osmotic balance (Gilles and Patrick, 2001). In most freshwater fish, the blood osmolality is regulated around 280-360 mOsm/kg to maintain homoeostasis and the iso-osmotic salinity in fish ranges from 10 to 12 g/L (Varsamos et al., 2005). Striped catfish, like most teleosts, can limit ion and water exchange using barriers at the skin surface via scales and a mucus layer when they live in their natural environment (Evans, 2011).

Several studies investigated the effects of salinity on striped catfish. Striped catfish do not appear to be efficient osmoregulators and the survival in saline condition is limited due to the absence of efficient electrolyte excretion. In juvenile stage, progressive transfer of striped catfish from freshwater to brackish water up to 20 ppt (parts per thousand) induced an elevation in plasma osmolality to a level higher than in

fish exposed to freshwater environment while gill NaK-ATPase only slightly increased (Schmitz et al., 2016). Moreover, chronic exposure of this fish up to 20 ppt was responsible for up regulation of mitochondrial proteins and ion transporters typical from chloride cells in kidney. These results suggest that, in response to an increased internal osmotic pressure, striped catfish may increase the size or the number of mitochondria-rich cells (MRCs) in kidney tubules, in order to increase the excretory capacity of the kidney while minimizing water loss. Embryos of striped catfish develop and hatch in water salinity up to 11 ppt and the isomotic point of fish larvae is 9 ppt (283 ± 34.6 mOsm/kg) (Huong and Quyen, 2012). In addition, the survival, growth, and feed conversion ratio of striped catfish fingerlings are the best at 6 ppt (Huong et al., 2009). Moreover, Lam et al. (2011) found that striped catfish fingerlings of 23.5 g grew fastest at the salinity of 9 ppt in indoor condition. Besides, Nguyen et al. (2014) also reported that the salinity conditions at 10 ppt did not affect growth performance, while salinities larger than 14 ppt gave poor survival rates at fingerling stage. The osmolality and concentration of $Na^+$ and $Cl^-$ was increased remarkably at high salinities (Nguyen et al., 2021). Moreover, recent studies of Hieu et al. (2022) showed that the intestinal microbiota of striped catfish was significantly disrupted in salinities higher than 10 ppt and these effects were proportional to the exposure time. In addition, the modifications of intestinal gene expression related to ion exchange and stressful responses may help the fish to adapt to hyperosmotic environment.

### 4.2 Impacts of salinity intrusion on the striped catfish farming in coastal provinces of the Mekong Delta, Viet Nam

Striped catfish culture is largely dependent on the availability of a plentiful water supply from the Mekong river (Nguyen and Dang, 2010). Therefore, the increasing salinization of freshwater areas will negatively impact the striped catfish culture model, as a result of changes to both soil and local freshwater resources (Sebesvari et al., 2011). These major changes have a significant direct impact on striped catfish such as exposing the fish to environmental stress and affecting the growth rate of catfish, which is a freshwater species (De Silva and Soto, 2009). The striped catfish needs freshwater to survive. An increase of 0.73 m of the sea level, combined with a decrease of 29% of the Mekong River flow, a scenario expected in Vietnam by 2100, will impact the area dedicated to the striped catfish farming by 11% (Thi et al., 2015).

Although temperature increase may also affect the culture of striped catfish, it is not much of a concern because striped catfish has a large temperature comfort zone (Department of Aquaculture, 2008). Consequently, only salinity intrusion induced by sea level rise are critical, while temperature increase due to global climate change will not have strong impact on striped catfish farming (Nguyen and Vu, 2020). Numerous studies have been conducted to evaluate the predicted effects of salinity intrusion on aquaculture

in the Mekong Delta. Nguyen et al. (2014) built a model that examined the effects of sea level rise (SLR) in three scenarios (increasing +30, +50 and +75 cm) on the striped catfish culture industry in the Mekong Delta. They concluded that rising salinity level with +75 cm would reduce the window appropriate areas for culture striped catfish in Soc Trang and Ben Tre province, and in Tien Giang's coastal districts (Figure 8). According to the study of Nguyen and Vu (2020), increased seawater intrusion in coastal regions, increased flooding in upstream regions caused by sea level rise and exacerbated by reduced river flow in the dry season or increased water discharge during the rainy season would altogether increase salinity and water level significantly. This would affect catfish farming due to salinity stress. Additionnally, new risks result from flooding when water levels are higher than the pond dyke, or pond dyke could be destroyed.

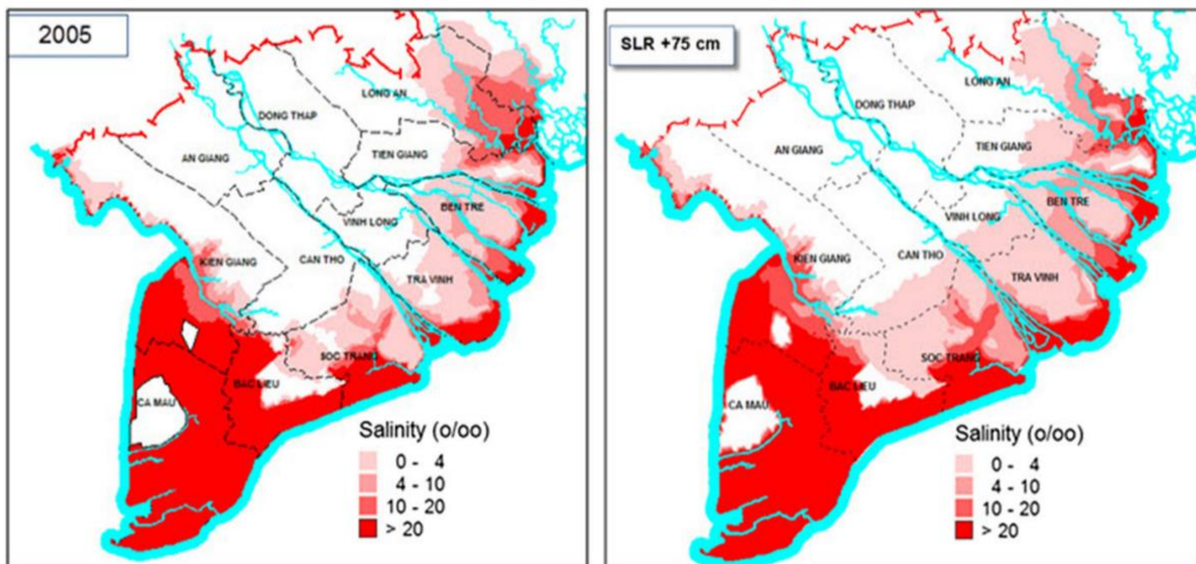


Figure 8: The areas affected by salinity intrusion according the SLR +75 cm scenarios (Nguyen et al., 2014)

Results from De Silva and Soto (2009) study predicted that aquaculture activities occurring in deltaic areas of major rivers in Asia, South America, and the Caribbean will encounter saline water intrusion caused by climate change. In the Mekong Delta, the risk from saltwater intrusion reduces the suitable area for catfish farming in the coastal provinces (Tien Giang, Ben Tre, Tra Vinh, and Soc Trang), from January to April. These provinces cover approximate 952 ha with a production of 183,666 tonnes, which accounts for 16.7 % of the total catfish farming area in the delta and contributed 11.8% to the striped catfish production in 2020 (DOF, 2020). The farmers in these provinces experience currently minor salinity variations depending on tidal amplitude. Most farmers do not stock fingerlings in months with high salinity

(January to April), and thus their window for farming is reduced. In the study of Phan et al. (2009), the annual yield from catfish farms located in the lower reaches of the Mekong River branches was significantly lower than that of upstream regions. These results suggested that the lowered yields could be a result of the diurnal fluctuations in the salinity. In 2011, the increased saline water intrusion in Ben Tre province, resulting in salinity up to 15 ppt, has resulted in increased mortality and in reduced growth rate compared to the normal years (Nguyen et al., 2014).

### 4.3 Adaptation to salinity intrusion of striped catfish farming

Aquaculture and fisheries are a top priority for the development in the Mekong Delta. In the last decade, in order to adapt with climate change and saltwater intrusion in this region, many national strategies, plans, actions, decisions, resolutions were issued and carried out. The Vietnam Government Resolution No. 120 NQ-CP (dated 17 Jan 2017) put aquaculture and fisheries as the top priorities for agri-aquaculture production structure for sustainable development of the Mekong Delta under climate change. On August 12, 2021, the Ministry of Agriculture and Rural Development approved the “The Sustainable Aquaculture Development Plan in the Mekong Delta to 2030” through Decision No. 3550/QD-BNN-TCTS. The plan was developed in line with the “Fisheries Development Strategy in Vietnam for the period of 2021 - 2030, vision to 2045” and Resolution 120/NQ-CP on “Sustainable development of the Mekong Delta adapting to climate change”, promoting proactive adaptation to climate change, fostering advanced technology application, and targeting economic growth in the Mekong Delta.

In the case of striped catfish farming, several options are possible for the sustainable development in the Mekong Delta region, and the adaptation to minimize the impacts of salinity intrusion.

**Changing striped catfish farming practice**: In the coastal areas, the striped catfish farmers have been encouraged to extend the nursing period of fish fingerlings, thus reducing the grow-out period in ponds during the months of high salinity intrusion (Nguyen et al., 2015). However, this change will result in slightly higher cost for transport, as juveniles will be heavier and perhaps at an increased risk in transportation of juveniles (De Silva and Phuong, 2011). In the Mekong Delta, some researches have started to experiment on the use of recirculating aquaculture systems (RAS) for the nursery and grow-out phases, which could also be regarded as an autonomous adaptation (Liem et al., 2017). In the RAS, water intake is very restricted except for the grow-out period with water exchange rate in first months being only 3% of the total volume/day, and the final months about 5-10%/day which was much lower than farming in soil pond (10-20% for first months and 30-40% for the final months) (Liem et al., 2017). Thus, RAS

simultaneously reduces water pollution and contributes an added benefit for the mitigation of environmental impacts of the striped catfish farming (Bosma et al., 2011; De Silva et al., 2010), and ultimately contribute to the mitigation of the impacts of climate change too. Although the results seem promising, the full costs and benefits are needed to be considered, especially with cost for building and energy to operating RAS system (Nhut et al., 2013; Liem et al., 2017).

**Shifting to another species**: Another option in dealing with saltwater intrusion for striped catfish farmers in the coastal provinces is to culture other aquatic species (Nguyen et al., 2016). However, farming other species would be expensive and would have a serious socio-economic impact on the involved communities. Such change would necessitate capacity building among rural farming communities, and even changes in infrastructure, in particular for ponds. For example, a water depth of 4.0 – 5.0 m is preferred in striped catfish farming ponds (Phan et al., 2009), but such deep ponds are unsuitable for other commonly farmed species in brackish environment such as Asian sea bass or even shrimp. Besides, pond restructuring would be necessary especially for those located directly next to rivers or main canals, and is likely to be very costly as lowering the water level will increase, for example, the pressure on the dykes by river water at high tides. The cost for dyke maintenance of striped catfish farms accounted for 0.3% of the total variable cost (per ha and per crop) (Nguyen et al., 2016). While investigating autonomous adaptation measures of striped catfish farmers to climate change in the Mekong Delta, Kam et al. (2012) estimated the costs for individual dyke upgrading in period of 2010-2020 to be about 14.6 million USD in the up and mid-stream regions and 3.0 million USD in the downstream region. Nonetheless, both economic and technical aspects of the feasibility of shifting to other species need further and thorough studies (Nguyen et al., 2016).

**Developing salinity tolerant strains of striped catfish**: According to Nguyen et al. (2015), most striped catfish farmers, who have been confronted with the risk from salinity intrusion prefer to continue producing striped catfish rather than shifting to another species. They believe that only by farming catfish, they can maintain their revenues at a sufficiently high level, enabling them to recover their investments. Along with this line, many researchers suggest that breeding a salinity-tolerant strain of striped catfish would be a good option as well (Nguyen et al., 2016). De Silva and Soto (2009) suggested that farming a tolerant strain will require minimal changes in the farming techniques and in the related infrastructures, and would not require the development of new market chains (De Silva and Phuong, 2011). The estimation from Nguyen and Vu (2020) showed that, the generic cost of a salinity-tolerant striped catfish breeding program, starting with 150 individuals of wild broodstock of various origins in the Mekong River, is about USD 120,000 (Table 1). Due to the generation interval, the actual cost in the long run, could be fourfold (USD 480,000). This program will serve only the farmers in the coastal provinces who produce about 10%

of the 1.2 million tonnes produced annually by Vietnam. The additional cost per kg of pangasius produced in the coastal provinces from this program would be USD 0.004 $kg^{-1}$. This is slightly less than 0.4 % of the present production cost (USD 1.1 $kg^{-1}$) and this appears to be a feasible investment (Nguyen et al., 2016). This selective breeding approach is widely applied in livestock and crop selection. In aquatic species, several selection programs have been implemented to improve (for example) the growth rate of tilapia (O. *niloticus*) (Ninh et al., 2014; Hamzal et al., 2014) or of Atlantic salmon (Flynn et al., 1999; Thodesen et al., 1999), or the stress tolerance of rainbow trout (Pottinger and Carrick, 1999). Recently, the use of modern molecular genetic techniques in selective breeding such as genomic selection has been used to reduce the time period required to develop a desired trait (Hollenbeck and Johnston, 2018; Zenger et al., 2019). Using these approaches, for traits such as salinity tolerance could significantly shorten the breeding time needed to achieve significant improvements with respect to the traditional selective breeding programs used in the past (Nguyen et al., 2016).

Table 1: Generic cost (USD) of a salinity-tolerant catfish breeding program in Viet Nam (Nguyen et al., 2016) (respecting the principles of an effective population size as established by Ponzoni et al. (2011))

| Time | Year 1 | Year 2 | Year 3 | Year 4 |
|---|---|---|---|---|
| Fixed cost of infrastructure (rental of a 1.5 ha farm with 1.1 ha of ponds)* | 7,200 | 7,200 | 7,200 | 7,200 |
| Salary | 4,700 | 6,400 | 6,000 | 4,250 |
| Materials: broodstock and feeds | 4,000 | 6,800 | 16,100 | 16,000 |
| Accessories, disposable tools | 480 | 2,850 | 950 | 950 |
| Electricity, gasoline, diesel | 320 | 1,700 | 320 | 320 |
| Equipment | 1,400 | 250 | 12,700 | 0 |
| Maintenance cost | 700 | 2,100 | 1,600 | 1,400 |
| Total | 18,800 | 27,300 | 44,870 | 20,120 |

**Cost of land: about USD 30,000 and infrastructure: USD 23,0000; interest rate of 8% and depreciation of infrastructures over 20 years, accounted for (Exchange rate: 21,000 VND to 1.00 USD).*
*Source of primary data: Dr. Trinh Quoc Trong, Director of National Breeding Centre for Southern Freshwater Aquaculture, Viet Nam (personal communication)*

## 5. Selection programs in aquaculture

Selective breeding programs contribute significantly to the development of fish farming, not only to achieve the necessary global demand, but also to reduce production costs, improve the resistance to diseases of farmed organisms, improve feed use and product quality (Gjedrem, 1997). Aquaculture is predicted to play a major and ever increasing role in meeting human needs for animal source food (Houston et al., 2020; Nguyen et al., 2022). However, the aquaculture sector is significantly behind plant and farm animals production in applying selective breeding, in spite that it has been suggested that the world

aquaculture production could be doubled in 13 years if selection programs were supplying stocks for farmed species (Gjedrem and Rye, 2016). It is estimated that less than 20% (Mair, 2007) (or even less than 10% according to Gjedrem, (2012)) of the world aquaculture production originates from material developed in selective breeding programs. The low level of genetically improved animals under culture is due to several factors, including the size and maturity of industries, the inability to domesticate or control reproduction in many species, the large number of farmed species, difficulties in retaining pedigree throughout the entire production process, the inability to collect large phenotypic data sets, and a general lack of informative genetic parameters information for traits (Gjedrem, 2012; Jerry et al., 1997).

Despite the current relatively low number of selective programs in aquacultures, fish have advantageous characteristics for breeding programs compared to terrestrial domestic animals. I refer here to the high fertility and, for many species the external fertilization, which guarantees flexibility in the definition of mating with the formation of full-sibs and half-sibling groups. Three main possible reasons may explain the limited genetic improvement programs in the sector: 1) little information about the reproductive cycles of several farmed species; 2) regular capture of wild specimens for breeding, which compromises the domestication of the species; 3) researchers, extensionists and fish farmers with little knowledge about quantitative genetics, theories and breeding programs (Gjedrem, 2005). Besides, some disadvantages of fish species for breeding programs also exist: young fish, such as larvae and fingerlings, are small, preventing early identification, with an increase in the program costs due to the need to grow whole-sibling families in separate structures, up to a size sufficient for marking. This issue could lead to the emergence of biases due to the different environments for the various families, leading to the confusion between genetic and non-genetic effects. Another potential pitfall is that the high prolificity of the animals could easily increase the rate of inbreeding (Farias et al., 2017).

The purpose of selection is to identify and select parents for the next generation who have the highest possible additive genetic merit for a trait or traits of interest (Farias et al., 2017), in our case salinity tolerance (which was measured as survival and growth rate of fish). Selection does not create new genes; its basic effect is to change the allelic frequencies for genes involved in the studied trait. Frequencies of alleles with favourable effects on the phenotype under selection are increased, and the frequency of less favourable alleles are decreased (Gjedrem, 2005). In aquaculture, the traits of economic importance in aquaculture are generally quantitative, governed by a large number of genes. The effect of changing the frequencies of the favourable alleles at these genes can be observed as a change of population mean for the trait under selection (Gjedrem and Rye, 2016; Nguyen, 2016). Available selection methods for fish breeders include individual selection, pedigree selection, family selection, within family selection, and combined

family and within family selection. The efficiency of each method can be predicted by calculating the expected genetic response for a given set of parameters (Lind et al., 2012).

The main goals of breeding programs on fish and shellfish are to increase the profitability and sustainability of production enterprises, while maintaining genetic variability in the cultured stock. The industry aquaculture achieved rates of genetic progress per generation 4-5 times greater than in livestock (Gjerde and Korsvoll, 1999). With several aquaculture species such as Atlantic salmon (*Salmo salar*), Nile tilapia (*Oreochromis niloticus*), and the Pacific white shrimp (*Litopenaeus vannamei*), selection for growth has dramatically increased efficiencies helping establish these species as global commodities. Traditionally, the initial focus has been on growth, which is a moderately heritable trait and relatively easy to select in breeding program (Zenger et al., 2019). While growth is one of the important determinants of aquaculture productivity, other traits such as environmental tolerance, feed conversion efficiency, disease resistance, and carcass or product quality are also economically important. However, carcass or product quality traits are difficult to measure on candidates for selection, can often only be measured late in life, involve destructive sampling, or have low heritability (Yáñez et al., 2015), but have major effects on the production efficiency and profitability of many species in aquaculture.

Mass selection, also known as individual selection, refers to selection solely based on the individual’s phenotype (Gjedrem and Baranski, 2011). The chief objective of mass selection is to increase the frequency of superior genotypes from a genetically variable population with obvious differences for characters of economic interest. The chances of success through mass selection depend upon the extent of genetic variation and strength of relationship between genotype and phenotype i.e heritability of the characters to be selected. Mass selection was considered the simplest way to improve traits such as growth or morphology and was for many years the most commonly used method of selection in aquatic species (Hulata et al., 1986; Lind et al., 2012). However, mass selection is only possible for traits that can be measured or recorded on live animals, since live individuals are naturally required as broodstock for the next generation.

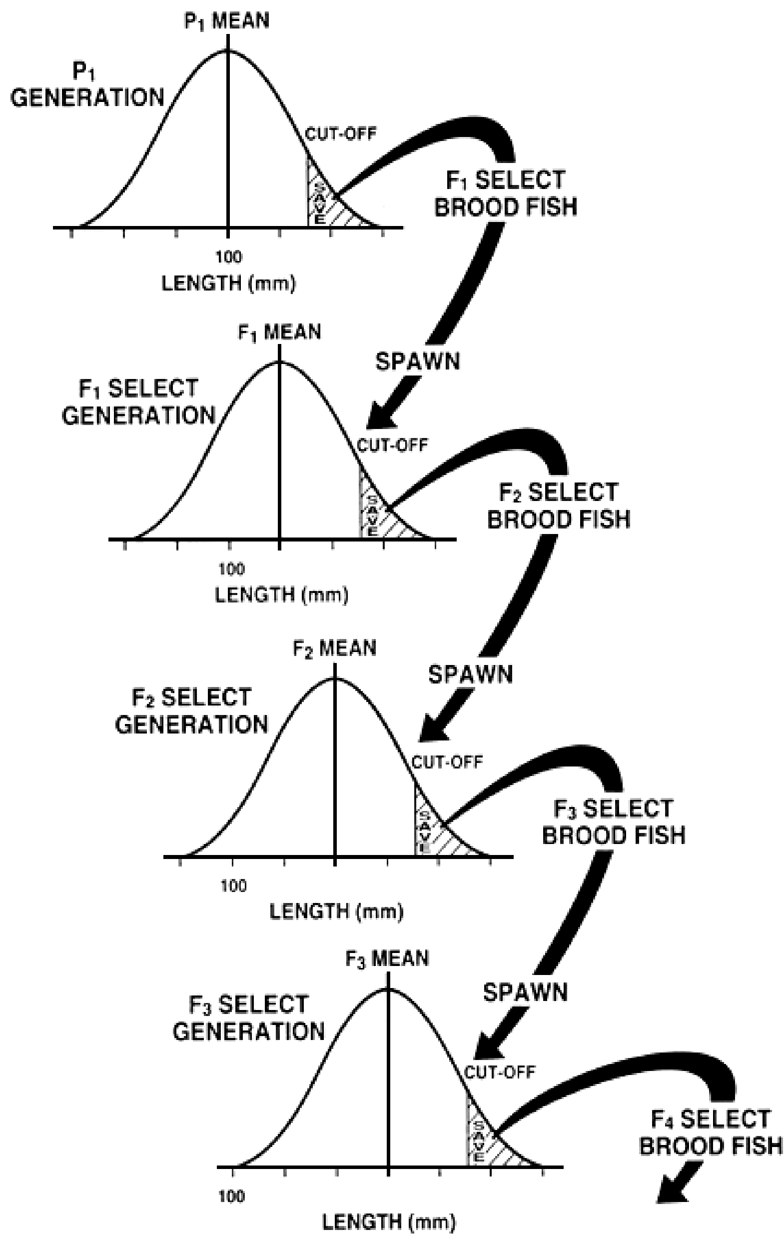


Figure 9: Schematic diagram of individual selection to improve growth rate by selecting for increased length. The figure shows four generations of selection and shows how length can be improved over time.(https://www.fao.org/3/v8720e/V8720E04.htm)

When applying mass selection, the evaluated animals have to be under equal environmental conditions. The influence of external conditions should be minimal and kept the same for all individuals

that are to be compared at any stage of the life cycle. Differences between individuals or groups of individuals for environmental factors like water temperature, salinity, density, light condition, and type of food and feeding regimes, may reduce the accuracy of the selection substantially, due to the confusion between genetic and environmental effects, and in that way, can also reduce the possibility for genetic improvement (Farias et al., 2017). To achieve as equal environmental conditions as possible for all individuals to be compared, they should be hatched at the same day or within a few day period and thereafter reared under identical environmental conditions. For this reason, it is necessary to form groups of animals with similar ages, grown in the same production systems, with similar control of water quality, stocking density, handling and feeding.

Mass selection for growth rate usually does not require individual identification or the maintenance of pedigree records, hence, it may be considered the least expensive method (Lind et al., 2012). However, the lack of pedigree information often leads to a rapid increase in the rate of inbreeding, especially in species where reproduction occurs naturally in growing environment such as tilapia. When intense selection is practiced, with a relatively high heritability of growth rate, there is a significant risk that a large proportion of selected breeders will be close relatives (Gjedrem and Baranski, 2011). As time passes, genetic gain will be reduced and the viability of animals will decrease. Several breeding programs have failed because they used strong selection in each generation without pedigree information and suffered a rapid accumulation of inbreeding which ultimately dramatically reduced productivity (Moav and Wohlfarth, 1976; Teichert-Coddington and Smitherman, 1988). These issues can be largely reduced by using a large number of broodstock in each generation (Bentsen and Olesen, 2002).

Combined selection is a strategy combining the information on individuals as well as information coming from relatives in order to maximize the accuracy of the estimated breeding value (Gjedrem and Baranski, 2011). This method, ideally, combines in an optimal way all available sources of information that can add to our knowledge about the breeding value of an animal: the information recorded on the animal itself, information about full- sibs and/or half-sibs and progenies as well as other relevant pedigree information. This strategy represents a general solution for obtaining the maximum rate of genetic gain, and the other simpler methods are special cases of this method. Combined selection is therefore in principle always the best method (Gjedrem, 2005). However, this method has a higher cost due to the need to mark the evaluated animals and/or to culture the families in separate structures up to a minimum weight for identification. As mentioned, culturing families in separate tanks may lead to the emergence of tank effect that will be confounded with a family effect, a bias that can reduce the accuracy of prediction of genetic values (Dussault and Boulding, 2018; Liu et al., 2016). An alternative is to breed families together, and to

recover parental relationships using DNA (Whalen et al., 2019; Dodds et al., 2019). This is one of the most important reasons to develop novel strategies to identify the parental information based on genomic data.

In several aquaculture species, conventional selective breeding approaches have resulted in significant improvement in productivity, with genetic gains per generation ranging from 8 to 12%. In addition, genetically improved species have shown remarkable vigor and high adaptation to a range of culture conditions (Nguyen, 2016). The selective breeding programs were generally able to increase the growth of 2.3 to 42% per generation, with an overall average of 12.7%. Atlantic salmon has the highest average estimate of genetic gain for growth rate with 17.8% per generation, followed by channel catfish (14.5%) and shrimp species (8.7%) (Gjedrem and Rye, 2016). These results suggest that the growth rate can be doubled in only six generations of selection. For species like tilapia and white leg shrimp (P. *vannamei*), this doubling can be obtained in less than 6 years, while for other species with long generation interval such as grass carp and Atlantic salmon, it may take as long as 24 years to reach this milestone. Doubling of the growth rate during six generations has been obtained for Atlantic salmon (Thodesen et al., 1999) and Nile tilapia (Thodesen et al., 2012).

Moreover, selective breeding has been shown to be effective in improving animal health via the development of more resistant genetic stocks (Suebsong et al., 2019). The improved disease resistance to specific pathogens have been demonstrated in several studies with moderate to high genetic gains per generation, for example increasing 18.4% for survival rate of Taura syndrome in white shrimp (Argue et al., 2002), 18.7% for infectious pancreatic necrosis resistance in Atlantic salmon (Storset et al., 2007), and 9% for streptococcosis resistance in red tilapia (Sukhavachana et al., 2019). Moreover, selection for adaptation to dietary environment is also of great interest in animal breeding so far (Callet et al., 2017; Le Boucher et al., 2012; Yamamoto et al., 2016). In selection for adaptation to dietary shifts of rainbow trout (*Oncorhynchus mykiss*), Le Boucher et al. (2012) demonstrated that survival rate, mean body weight and biomass can be improved after a single generation of selection for ability to adapt to totally plant-based diet (with selection pressure being 3.25%). Similarly, Callet et al. (2017) also succeeded in selecting this species on their ability to grow with a diet completely devoid of fishmeal and fish oil after 3 generations of selection. The final proportion of selected fish ranged from 3.9% (first generation) to 4.9% in third generation. On amago salmon (*Oncorhychus masou ishikawae*), selective breeding for adaptation to a low fish-meal diet presented positive results on growth performance of fish after only one generation of selection (Yamamoto et al., 2014).

In selective breeding programs, heritability is an extremely important parameter. It is the measure of the relative proportion of genetic variance in the total variation of a specific trait (Gjedrem and Baranski, 2011; Ye et al., 2017). The "heritability in narrow sense" describes the proportion of total phenotypic variation due to additive genetic factors, it is important in predicting how a trait will respond to selection. A large number of factors can affect heritability estimates, such as genetic background, genetic variation, sample size, culturing conditions and the number of generations of selection that the population has gone through (Gjedrem, 2005; Visscher et al., 2008). Heritabilities have been estimated for a number of traits in many aquaculture fish species, such as growth in the common carp (*Cyprinus carpio L.*) (Vandeputte et al., 2004), Asian seabass (*Lates calcarifer*) (Domingos et al., 2013; Wang et al., 2008a), and resistance to columnaris disease in the Atlantic salmon (*Salmo salar*) (Evenhuis et al., 2015). Heritability estimates commonly range from 0.1 to 0.4 for economically important traits in aquatic animals (Gjedrem and Baranski, 2011), with some examples in Table 2.

Table 2: Heritabilities for different traits in selected aquatic species

| Species | Trait | Heritability | References |
|---|---|---|---|
| Rainbow trout | Body weight | 0.21 | Gjerde and Korsvoll (1999) |
| Rainbow trout | Survival | 0.16 | Rye et al. (1990) |
| Barramundi | Body weight | 0.22 | Domingos et al. (2013) |
| Atlantic salmon | Body weight | 0.35 | Rye and Refstie (1995) |
| Rohu carp | Body weight | 0.23 | Gjedrem and Baranski (2011) |
| Rohu carp | Survival | 0.16 | Gjedrem and Baranski (2011) |
| Whiteleg shrimp | Body weight | 0.17 | Gitterle et al. (2005) |
| Pacific oyster | Body weight | 0.16 | Evans and Langdon (2006) |

Selective breeding is also one of the methods that is used to improve the salinity tolerance of tilapias (euryhaline fish) to withstand high salinity levels (Jaspe and Caipang, 2011). The results from a few pioneer studies suggested that substantial additive genetic variance exists for growth rate and survival in saline environments that can be exploited through selective breeding programs (Tayamen et al., 2010; Tran et al., 2008). Cnaani and Hulata (2011) stated that the salinity tolerance of the fish is the overall fitness or productivity of fish in a saline environment: a combination of quantitative traits including metabolism, growth, osmoregulation, immunocompetence and fecundity, each of which influenced by multiple genes underlying genetic variation. Salinity tolerance based on growth and survival has also been used in several selection programs on tilapias (Cnaani and Hulata, 2011; Tayamen et al., 2010). The breeding program for salinity tolerance of tilapia (*Sukamadi*) with four generations under high salinity environment resulted in

strong selection on salinity-adaptive regulatory pathways (Xiaofei et al., 2021). In addition, molecular studies of osmoregulatory properties of the gills, kidney, gut, and brain underlying phenotypic differences in salt tolerance have produced a wealth of genomic knowledge that can be used in genetic studies of interspecies variation for salinity tolerance. Genetic polymorphisms can be searched for cultured or natural populations, once relevant genes are identified. Information of quantitative trait loci (QTL) associated with, or genes directly involved in salinity tolerance may facilitate marker-assisted or gene-assisted selection for this trait in fish (Cnaani and Hulata, 2011).

### 6. Selection program on striped catfish in Vietnam

In Vietnam, until now, the genetic selective breeding programs on striped catfish have been limited. Only one selection program for growth of the species has been carried out in freshwater. This program, a combined between and within family selection, was started at the Southern National Breeding Centre for Freshwater Aquaculture, under the supervision of the Research Institute for Aquaculture No.2 in 2001 (Sang et al., 2012, 2010). The base population of this breeding program was made up from stocks collected at three different hatcheries in the Mekong Delta, Vietnam. In the period 1999-2001, each stock was collected from grow-out farms that reared wild fingerlings caught over several seasons and at various locations in the Mekong River. In 2001, the base population animals were mated in single pairs within hatchery to produce the 2001 year-class (G1 generation). These fish were not individually tagged, so the parents of the G2 generation (year-class 2005) were selected amongst the heaviest G1 animals. In 2005, a total of 206 G2 families were produced in May. At the average size and age of 45.9 g and 171 days, respectively, an average of 75 individuals from each full-sib family were randomly sampled and marked by Passive Integrated Transponder tags (PIT- tags, Sokymat, Switzerland). In total, 12,190 fish were tagged, representing 162 families, from 95 sires and 97 dams. In G2 (year-class 2005) one also selected for fillet yield, by combining individual information on body weight and family information for yield. Inclusion of fillet yield as a second trait was motivated by the assumed high economic importance of this character. The G3 families were produced in June and July 2008 with 156 full-sib families from 93 sires and 156 dams (Sang et al., 2012, 2010).

By 2015, after three generations, moderate heritability estimates were obtained for body weight (0.21–0.34) and fillet weight (0.19–0.22), while low heritability estimates were found for fillet yield (0.03–0.05), intestinal fat (0.04), predicted fillet fat percentage (0.04) and fillet colour (0.04) (Sang et al., 2012, 2010). The selection program increased fillet weight and growth survival by 8.5% and 7.4%, respectively.

Direct response to selection for high growth was 9.3% per generation. The improved strain of striped catfish had 13.4% greater growth performance than their wild counterparts of the latest generation (Vu et al., 2019).

No selective breeding program to improve salinity tolerance has been carried out in saline condition to our knowledge. However, many studies have investigated the effects of increased salinity on the growth performance of striped catfish. In generally, this species is well tolerant to moderate increased salinity in water (Huong and Quyen, 2012; Lam et al., 2011; Nguyen et al., 2014; Thao et al., 2013). Synthesized results suggest a substantial additive genetic variance for salinity tolerance traits that could be further exploited through a selection program.

### 7. Parentage assignment in aquaculture

In aquaculture, pedigree information is important in selective breeding programs and hatchery management (Yue and Xia, 2014a). However, maintaining pedigree information in selective breeding programs is one of the challenges in aquaculture (Liu et al., 2016). Obtaining pedigree information in aquatic species is difficult and costly. The basic reason for this issue is that farmed aquatic animals are all too small at hatching (from a few micrograms in mollusks and crustaceans to around 100 mg in salmonids fishes) to be physically tagged (Vandeputte and Haffray, 2014). To solve this problem, a first solution is to rear the fish by family in individual tanks until a size where tagging is possible. This was successfully extended to major aquaculture species such as salmonids, tilapias, oyster, or shrimps (Gjedrem, 2012; Krishna et al., 2011; Thodesen et al., 2012; Zak et al., 2014).

As already explained, a first disadvantage of this method is that it requires a large investment in operational floor space and specially designated tanks and feeding systems. Another mentioned disadvantage is that separate rearing of families create shared environmental effects that are confounded with genetic effects (Dussault and Boulding, 2018). Moreover, the number of families is limited to the number of used family rearing units. Therefore, mating designs are constrained to those where the number of families produced is low for a given number of tested parents. A problem with designs such as single pair mating (where each pair of parent is used for only one family), is that, unlike factorial designs, these simple designs do not allow a proper separation of additive, maternal and dominance effects (Becker, 1968). To address this issue, mixtures of equal-aged progeny from different families can be reared communally to preclude the development of such family-specific environmental effects, and genetic markers can be used subsequently to assign individuals to families after evaluation of individual performance (Doyle and Herbinger, 1994). Thus, the impact of early common environmental effects is considerably decreased if

genetic markers are used for parentage assignment when selecting individuals for early growth rate traits (Herbinger et al., 1999; Norris et al., 2000). Fish can be mixed at any time, even as newly fertilized eggs. If possible, it is even permissible to mix sperm (or eggs) from different fish before fertilization (Raul et al., 2006). Parentage assignment using genetic markers combined with physical tagging provides a new way to conceive breeding programs, which have to be optimized with their specific constraints. One of the most important constraints is to properly define the number of individuals to genotype to limit costs, maximize genetic gain while minimizing inbreeding (Vandeputte and Haffray, 2014).

The principle of molecular analysis for parentage assignment is based on the simple concept that an offspring inherits one of two alleles at each locus from each of its parents. The comparison of genotypes at DNA markers of progeny with those of potential parents allow to determine parentage (Herbinger et al., 1995). From the 1990s, molecular parentage analyses have been used in reconstructing pedigrees in aquaculture species (Yue and Xia, 2014b). Basically, exclusion and likelihood-based methods are two major approaches for parental analysis (Harrison et al., 2013). The principle of exclusion is to check the compatibility of offspring and parental genotypes with Mendelian inheritance, and it is the gold standard of parentage assignment. One drawback of this method is its sensitivity to genotyping errors. However, genotyping errors can be dealt with allowing a limited number of allelic mismatches between an offspring and its parent alleles if error rates are moderate and theoretical assignment power is high (Vandeputte et al., 2006). Exclusion programs used in aquaculture are PROBMAX (Danzmann, 1997), VITASSIGN (Vandeputte et al., 2006), and FAP (Taggart, 2006).

Likelihood-based methods employ Mendel’s law quantitatively to calculate the likelihoods of different candidate relations among a set of individuals and choose relations that have the maximum likelihood as the best inference (Trọng et al., 2013). Likelihood methods generally give higher assignment rates than exclusion for low power marker sets, but sometimes give inconsistent results (Herlin et al., 2007). However, the likelihood based methods were considered as one solution to handle genetic data with high uncertainty (Kalinowski et al., 2007; Riester et al., 2009). Moreover, using sibship information in calculations can greatly improve the efficiency of the likelihood methods (Wang and Santure, 2009). Examples of likelihood programs used in aquaculture are CERVUS (Kalinowski et al., 2007), PAPA (Duchesne et al., 2002), and PARENTE (Cercueil et al., 2002).

Mutations in the genome create genetic variability (or polymorphisms), which is reflected as allelic diversity of molecular markers. Microsatellite markers are simple sequence repeats (SSRs) arranged in tandem arrays scattered throughout the genome, both within known genes and in anonymous regions.

Microsatellite are markers of choice for parentage assignment (Marklund et al., 2009; Vandeputte et al., 2011) in aquaculture species due to their elevated polymorphic information content, codominant mode of expression, Mendelian inheritance, abundance and broad distribution throughout the genome (Liu and Cordes, 2004; Wright and Bentzen, 1994). Microsatellites have been used for the parentage assignment in more than 20 aquaculture species, such as Atlantic salmon (Norris et al., 2000; Sourinejad et al., 2011), Chinese shrimp (Dong et al., 2006), Asian seabass (Liu et al., 2012; Wang et al., 2008b), common carp (Vandeputte et al., 2004), rainbow trout and turbot (Estoup et al., 1998), sea bream (Borrell et al., 2011), and European seabass (Saillant et al., 2006). Tri- and tetra nucleotide microsatellites are relatively easy to genotype and not prone to genotype errors due to less stuttering effects of these markers during PCR. However, they are usually less abundant in genomes and less polymorphic in comparison to dinucleotide microsatellite (Goldstein and Tterer, 1999).

Despite the success of using microsatellites, single nucleotide polymorphisms (SNPs) are becoming increasingly popular in aquaculture research in parentage testing because they are highly abundant, amenable for high throughout genotyping, easier to standardize among laboratories and have lower genotyping error rates (Liu et al., 2016; Strucken et al., 2016). SNPs are single base pair mutations at any given position of DNA. They are one of the most abundant type of the DNA sequence polymorphisms (or variants) and are widely distributed along the genomes (Liu et al., 2011). In the population, SNPs can have four alleles theoretically (i.e. A, C, G, T, corresponding to the four nucleotides of the DNA alphabet), but they are regarded as bi-allelic as, most often, only two alleles are observed at a given SNP (i.e. typically A and G, or C and T). SNPs have several advantages: they are very frequent (approximately 1 every 1000 base-pairs in many genomes, in average), co-dominantly inherited, sequence-tagged markers and highly adaptable to large-scale automated genotyping (Dashrath et al., 2017; Ulloa et al., 2015). Regardless of the fact that individual SNPs are less informative than microsatellites, they are better suited for modern genotyping technologies which are rapidly making them the marker of choice for parentage testing (Strucken et al., 2016). The information for parentage tests provided by 40 to 100 SNPs is approximately equivalent to the information from 14-20 microsatellites on average (Fisher et al., 2009; Glaubitz et al., 2003). According to the study of Lien et al. (2011) on Atlantic salmon, the power of 80 SNPs for parentage analysis is higher than using 11 microsatellites.

The number of DNA markers needed to achieve acceptable levels of correct parentage assignment depends on factors such as the level of polymorphism of the markers, locations of markers in the genome, the number of parents and offspring, and the mating system (Vandeputte et al., 2011; Villanueva et al., 2002). Empirical studies tend to suggest that the adequate number of SNPs for an efficient panel would

number 100-450 (Lapègue et al., 2014; Nguyen et al., 2014; Sellars et al., 2014). In a study on rainbow trout, parentage assignments matched perfectly with the known pedigree when 95 SNPs were used (Liu et al., 2016). According to Anderson and Garza (2006), around 60-100 SNPs with high (0.3-0.5) minor allele frequency are adequate for accurate large scale parentage assignment. The power of a set of markers for parentage analysis is also related to their locations on a genome; linked markers usually show lower power than unlinked markers. Therefore, it is better to select markers on different linkage groups to ensure high power of the set of markers for parentage analysis (L. Hauser et al., 2011; Yue and Xia, 2014a).

**8. Shallow whole genome sequencing for parentage assignment**

Next-generation sequencing (NGS) is a new technology used to determine the order of nucleotides in entire genomes or targeted regions of DNA or RNA automatically. This massively parallel sequencing technology offers ultra-high throughput, scalability, and speed. NGS can sequence hundreds or thousands of genes or even a whole genome in a short period of time. NGS has revolutionized the biological sciences, allowing labs to perform a wide variety of applications and study biological systems at a level not previously possible (Barzon et al., 2011; Qin, 2019). Deep whole genome sequencing (DWGS) refers to sequencing a genomic region multiple times, sometimes hundreds or even thousands of times (Arora et al., 2019; Okada et al., 2018). In short, the targeted DNA is split into small fragments that are subsequently sequenced in parallel, allowing the generation of enormous amounts of sequence data at high speed. Using that approach, almost any genomic piece of the targeted region (possibly a complete genome) gets sequenced multiple times (the average number is called "the sequencing depth", expressed as n X, where n is the average depth), sometimes hundreds or even thousands of times, leading to the name of "deep sequencing" (Kulski, 2012; Okada et al., 2018). However, DWGS at high coverage remains expensive (although the costs are progressively decreasing) (Whalen et al., 2018), requires large amount of high quality input DNA and is associated with large bioinformatics processing time, and the need to store and manipulate large datasets, especially when thousands of individuals are analyzed.

Alternatively, shallow whole genome sequencing (SWGS) follows the same principle, but with a much lower sequencing depth. The objective is to reduce the cost and computer load at the expense of a lower coverage of the targeted region (Raman et al., 2019; Kader et al., 2016). With a new and high-throughput technology, the SWGS can be used to achieve genome-wide genetic variation accurately and cost-effectively with a broad range of species. Based on shotgun sequencing, SWGS can sequence a whole genome at a very low coverage (most frequently between 0.4 x and 1 x). The SWGS is amenable to low quantities of input DNA, processing is compatible with low-output instruments and sequence coverage can

be adjusted flexibly according to the purpose of studies (Parker et al., 2019; Kader et al., 2016; Van Roy et al., 2017). SWGS returns more data, greater statistical power, and new rare variant discovery capabilities than traditional genotyping arrays and genotyping-by-sequencing method. The downstream analysis is simplified by the use of freely available bioinformatics tools and data analysis packages written especially for the analysis of SWGS data (Parker et al., 2019). SWGS can be used in genome-wide association studies (GWAS), evolutionary analyses (Therkildsen and Palumbi, 2017), pharmacogenomics, molecular breeding. In fact, the SWGS has been used in several studies on human, especially on cancer analyses (Beagan et al., 2021; Chin et al., 2018). For example, in a study on the detection of copy number of alterations (somatic changes to chromosome structure that result in gain or loss in copies of sections of DNA, and are prevalent in many types of cancer), Van der Eecken et al. (2022) demonstrated that SWGS is more accurate than fluorescence *in situ* hybridization method.

As the cost of sequencing declines, especially for SWGS, it becomes feasible to use this technology to obtain genomic information for wide range of research and commercial applications (Li et al., 2019). Genetic data generated from SNP array for parentage assignment have been implemented in several species (Holman et al., 2017; Strucken et al., 2016; Tortereau et al., 2017). However, in aquaculture species, few SNPs array have been developed so far (for Atlantic salmon, Channel catfish, rainbow trout) (Dashrath et al., 2017; Palti et al., 2015; Holman et al., 2017). Therefore, in species for which a well-established reference genome is not available, or a proper array has not been developed yet, SWGS data is a flexible alternative to arrays. An attractive feature of using SWGS data is that it does not require up-front costs of developing marker panels (such as SNP panels) and it does not necessitate dedicated equipment. It may also be used without a reference genome sequence, although a draft reference genome is useful for sequence alignment, identifying and ordering SNPs, and for quality control. Using sequence data for parentage assignment was investigated in a few studies (Dodds et al., 2019, 2015; Whalen et al., 2019). In fact, sequencing data may provide a cost-competitive option for parentage testing, especially if other genomic information (such as breed assignment or population structure) is also sough and/or if no parentage marker panel is available for the targeted species.

One of the drawbacks of sequence data is that it is possible to only observe one of the two alleles of a heterozygous genotype in a scarcely covered region, leading to consider a heterozygous genotype as homozygous. This is of course especially the case for SWGS data, where the low coverage could easily lead to such ambiguous genotypes, or even to missing genotypes (Dodds et al., 2015). Because shallow-depth sequencing data does not provide unambiguous genotype calls, the standard parentage analysis

methods no longer apply. Despite a few approach developed for using sequence data in parentage assignment, their performance, to our knowledge, is not well understood and necessitates further studies.

### 9. The role of epigenetic phenomena in the improvement of aquatic species

In theory, these traditional selective breeding methods assume that the influence of a given gene on an animal's phenotype are a functional consequence of the order in which its nucleotide bases are arranged and genetic differences between individuals are due to differences in DNA sequence inherited in a Mendelian manner (Moghadam et al., 2015; Senaldi and Smith-Raska, 2020). However, in recent years, it has been reported from model organisms and from a few non-model species that additional genomic modifications, generally referred to as "epigenetics", can affect the phenotype, with some of these changes being passed on to the next generation (Goddard and Whitelaw, 2014; Lacal and Ventura, 2018). Epigenetics refers to molecular processes that result in the modification of the phenotype without a change to the DNA sequence itself. Epigenetic factors are specific molecules that affect gene expression without changing the DNA sequence (Jablonka and Lamb, 2006; Senaldi and Smith-Raska, 2020). Epigenetic mechanisms (or epigenetic marks) include DNA methylation, histone modification and non-coding RNA activity that can affect gene expression primarily through the local modification of chromatin. Unlike the DNA sequence, the epigenetic marks can be triggered by environmental stimuli, and have therefore been proven to be important mediators of phenotypic responses to environmental signals. They can persist throughout life or across multiple generations (Moghadam et al., 2015; Weaver et al., 2004). For example, in fish, exposure to toxins (tributyltin and triphenyltin) (Wang et al., 2009), freshwater environment (Artemov et al., 2017), changing temperature (Campos et al., 2013), and chemicals (17-ethinylestradiol) (Strömqvist et al., 2010) have all been associated with changes in DNA methylation and concomitant changes in the phenotype.

Although epigenetics has attracted remarkable attention in crop science (Álvarez-Venegas and De-la-Peña, 2016; Ong-Abdullah et al., 2015) and, more recently in livestock (Goddard and Whitelaw, 2014; González-Recio et al., 2015), there is less knowledge on epigenetic mechanisms in farmed aquaculture species. Recently, studies on salmonids, oysters, and mussels have provided the first evidence on epigenetic mechanisms associated to commercially important traits (Gavery and Roberts, 2017; Moghadam et al., 2015). There is some evidence in salmonids that changes in DNA methylation are related to variation in life-history phenotypes including early male maturation (Morán and Pérez-Figueroa, 2011), anadromy (Baerwald et al., 2016), smoltification (Morán et al., 2013), and growth potential (Burgerhout et al., 2017). In European seabass and rainbow trout, the role of epigenetics in mediating phenotypic responses to various

aspects of the diet has been evaluated (Marandel et al., 2016; Panserat et al., 2017; Terova et al., 2016). Moreover, several recent studies have focused on understanding the role of epigenetics in other important phenomena in aquaculture including adaptation to captivity (Le Luyer et al., 2017) and sex control via ploidy manipulation (Covelo-Soto et al., 2015; Jiang et al., 2016; Zhou et al., 2016).

Among epigenetic marks, DNA methylation is the most widely investigated epigenetic mechanism (Moghadam et al., 2015). In mammals, DNA methylation plays a very important role in animal development and is associated with several main processes, including chromosome inactivation, genomic imprinting, transposable element repression and carcinogenesis (Meissner et al., 2008; Seisenberger et al., 2013). DNA methylation refers to the process of the enzymatic addition of a methyl group to a cytosine residue in DNA, which occurs almost exclusively at CpG dinucleotides (Gavery and Roberts, 2017). Approximately 70-80% of such dinucleotides sites remain methylated in non-embryonic cells in mammals (Tucker, 2009). DNA methylation is catalyzed by a family of DNA methyltrasferases (DNMTs), including the maintenance of methyltransferase DNMT1 (responsible for copying pre-existing DNA methylation patterns to the new strand during mitosis) and the *de novo* methyltransferases DNMT3A/3B. Through associations with other DNA-binding proteins or through the physical blocking of transcription factors, DNA methylation is known to be repressive when located in genes promoters (Bell and Felsenfeld, 2000). According to Jones (2012), when methylation affects the promoter regions, it is associated with gene silencing, which is the most well-known function of the methylation process; however, when it involves the transcribed regions, it increase transcriptional activity. In aquatic species, DNA methylation is present in all species examined to date and is the most studied epigenetic mark of fish and shellfish, with both having genes that encode the basic methylation machinery (Gavery and Roberts, 2017). In model fish species such as zebrafish and medaka, patterns and functions of DNA methylation are increasingly understood (Gavery and Roberts, 2017; Metzger and Schulte, 2016). Comparing to mammals, the global DNA methylation levels in fish are remarkably higher, although the significance of this difference remains unclear (Jabbari et al., 1997; Zhang et al., 2016). In oysters, the methylated fraction tends to consist of gene bodies, while other genomic regions exhibit less methylation, especially in transposable elements (Feng et al., 2010; Simmen et al., 1999; Zemach et al., 2010). In shellfish, DNA methylation does appear to be associated with gene regulation. Similarly, there is a positive association between methylation status, both in gene bodies and putative promoter regions, and expression in the Pacific oyster (Gavery and Roberts, 2013; Olson and Roberts, 2014). Riviere et al. (2017) also show that DNA methylation patterns are dynamic across developmental stages of the Pacific oyster, and correlations between DNA methylation and gene expression patterns suggest that DNA methylation plays a role in gene regulation during development.

However, more research is needed to quantify the functional relationship between DNA methylation and gene expression. Such research could have significant implications for improving resilience in aquatic species, particularly if DNA methylation patterns are heritable.

Usually, most epigenetic marks are reprogrammed during the process of the production of the gametes of the parents and the formation of a zygote (Santos et al., 2002). However, some epigenetic marks are not erased and therefore inherited along with the DNA from the parents to offspring (Morgan et al., 1999). According to Van Otterdijk and Michels (2016), maintenance and *de novo* methylation and active and passive demethylation are crucial for embryonic development and epigenetic inheritance. After fertilization, gametes are completely demethylated and are remethylated to erase all epigenetic marks that an individual accumulated over his lifespan. However, this resetting process is hindered during early development, perhaps accounting for transgenerational transmission of these epigenetic footprint. In mammals, the meiotic transmission of DNA methylation patterns, thus the opportunity for transgenerational epigenetic inheritance through DNA methylation, is rare, because individuals undergo extensive DNA methylation reprogramming in the early embryo stage (Daxinger and Whitelaw, 2012). However, transgenerational epigenetic inheritance is more common in plants, which do not exhibit extensive resetting of DNA methylation between generations (Hauser et al., 2011). In fish, there was clear evidence of transgenerational inheritance of environmentally-induced DNA methylation patterns, suggesting that at least some part of the genome putatively escapes resetting between generations (Shao et al., 2014).

Considering that environmental factors can alter the epigenetic marks of animals'genome, that these marks can have an impact on gene expression and on phenotypes of interest, and that such changes can be inherited across generations, this phenomenon should be considered for exploitation in selection programs (Goddard and Whitelaw, 2014). But caution is needed. For some traits, all offspring will inherit epigenetic marks from their parents, so that some offspring groups that appear to have inherited superior additive genetic variation (additive genetic variation causing permanent and accumulative effects on the phenotype of subsequent generations), might actually have inherited temporary epigenetic marks that have a strong influence on traits (Bjornsson et al., 2004; Moghadam et al., 2015). The influence of epigenetic factors will depend on the structure of the population, the size of the epigenetic component relevant to additive genetic component and on the persistence of the epigenetic marks. If the effects of epigenetics on traits are stable over many generations, then we could consider them as part of the true additive genetic effect (Moghadam et al., 2015). In striped catfish, several studies investigated the effects of environmental stimuli such as salinity and temperature on biological factors as well as growth performance (Hieu et al., 2021; Nguyen et al., 2021; Thao et al., 2013). However, no studies evaluated the reprogramming and the effects of external

stimuli on epigenetic marks and on the DNA methylation process of striped catfish. Moreover, selection programs based on growth rate in freshwater have proven an improvement of growth performance of striped catfish in subsequent generations (Sang et al., 2012; Vu et al., 2019), but no study investigated the possibility of intergenerational epigenetic inheritance and its importance in the improvement observed in these selection programs. With the growing attention of the application of epigenetic knowledge for improvement of the productivity and sustainability of aquaculture industry, the investigation on epigenomics of striped catfish may facilitate greater understanding about the interactions between phenotypes and environmental factors (e.g feeds, water quality, global warning, salinity intrusion), thus increasing the sustainability and profitability of this species.

## 10. The role of high-quality reference genomes in aquaculture

A reference genome assembly can serve as a foundation for studies on genotypic and phenotypic variation, on genome structure, on the evolution of the target organism and for breeding programs (Lischer and Shimizu, 2017; Whibley et al., 2021). It has become hugely important for understanding the biology of entire populations of a species in relation to key aspects such as its ability to adapt to its environment or its economic value. Evaluation of the quality of a reference genome can be done with some parameters including completeness, continuity, number of gaps, misassemblies, rates of base errors, fragment coverage distribution, and low-quality regions. A high quality reference genome sequence can be used to facilitate basic research, positional cloning, DNA marker discovery, genome wide association study (GWAS) and genomic selection (GS), resolution of difficult regions and assessment of novel genetic variation (Yue and Wang, 2017). Reference genomes provide a view of the architecture of the genome, including both genic and intergenic regions. They include repetitive regions, some of which are challenging to assemble, such as segmental duplications, centromeres and telomeres, satellites, and mobile elements. Moreover, using reference genomes as guide is most useful for the identification of classical genetic variants, such as SNPs and copy number variants (CNVs), as well as structural variants that are particularly difficult to detect in fragmented and incomplete reference genomes alone, but are potentially important in adaptation to environmental change (Mérot et al., 2020). The power of population genomics guided by reference genomes to identify runs of homozygosity as a means to estimate inbreeding has also been reported, as well as to reveal the dynamics and fate of deleterious variation in several species (Dussex et al., 2021; Kardos et al., 2016). In addition, according to Flanagan et al. (2018), the use of reference genomes in population genomics facilitates the identification of traits under natural selection that form the basis and architecture of local adaptations. Reference genomes provide the functional and genomic contexts for regions influenced by selection, thereby enabling association of such loci with phenotypes important to adaptation and resilience.

In recent years, with the rapid advance of sequencing technologies and bioinformatics analysis of large volumes of sequencing data, the development of whole genome sequencing in aquaculture is happening very fast. Currently, the draft sequence of whole genomes of over 563 aquatic species have been published, including more than 24 genomes of species with the highest production value (e.g Atlantic cod, Pacific oyster, common carp, European seabass, Nile tilapia), and serve as reference genomes for downstream studies on important issues related to aquaculture (Ronco et al., 2021; Houston et al., 2020; Yue and Wang, 2017). These reference genomes have been used to identify a large number of SNPs and microsatellites, which have been and are being in a broad context, including linkage mapping, QTL mapping, GWAS, comparative genomics and functional analysis for economically important traits. For example, in channel catfish (*Ictalurus punctatus*, Ictaluridae), a 250 K SNP microarray was developed. The chip was used to conduct GWAS for columnaris disease resistance (Geng et al., 2015) and low oxygen tolerance (Wang, 2016). In Atlantic salmon, 5650 SNPs covering the whole genome were used in GWAS to identify SNPs related to fillet fat content and fillet firmness (Sodeland et al., 2013). In rainbow trout, to identify DNA markers associated with fillet yield, a GWAS for fillet yield was conducted in the synthetic population using a SNP array with 57 K SNPs (Gonzalez-Pena et al., 2016).

Stress tolerance is a critically important factor in aquaculture and increasing tolerance to stress is one of the important breeding goals. Wang et al. (2012) sequenced the whole genome of the Pacific oyster in combination with transcriptome analysis, and identified an extensive set of genes responding to environmental stress. The European seabass, a temperate-zone euryhaline teleost and important aquatic species, is subdivided into two naturally hybridizing lineages. One lineage lives in the north-eastern Atlantic Ocean and the other the Mediterran and Black Sea. Comparative genome analysis revealed a high degree of synteny with more derived teleosts and the expansion of gene families, which were specifically associated with ion and water regulation. This finding highlights the adaptation to salinity variation (Tine et al., 2014). Selection, including natural selection in the wild and selective breeding for genetic improvement in farmed aquatic species, leaves signatures within the genome of a given species. In aquaculture, identifying these signatures is of importance in finding genes that facilitate the acceleration of genetic improvement. In a study on tilapia, sequencing the genome of 47 individuals resulted in a total of 1.43 million high quality SNPs and over a hundred putative selective sweep regions in each line of tilapia were identified. Most detected selection signatures were located in non-coding regions of the tilapia genome. Moreover, integrin signaling pathway and the Wnt signaling, gonadotropin-releasing hormone receptors were under positive selection in all genetically improved farm tilapia line. In tilapia aquaculture,

the genome-wide map of genetic variation and selection footprints are important for genetic studies and for accelerating genetic improvement (Hong Xia et al., 2015).

The DNA sequence of the full genome is used to detect causative genes in the QTL for important traits in aquaculture species. With the availability of reference genomes, DNA sequences covering significant QTL for important traits can be easily identified by blasting DNA sequences of marker flanking the QTL against the reference genome (Yue and Wang, 2017). Although growth is one of the economically important traits in aquaculture, the molecular mechanisms underlying growth are nevertheless not yet fully understood. Liu et al. (2014), used a GWAS to analysis growth in tilapia. They identified significant QTLs for growth traits on several chromosomal regions. Using the reference genome, the *GHR2* gene was identified as important, affecting the difference in growth among tilapia species. Comparative analysis of genome sequence in combination with RNA-seq between carnivorous and herbivorous fish species may detect genes, pathways and gene networks related to dietary preference. Discovering these genes could contribute to the development of strategies to replace fish meals and oils with plant-based proteins and oils. Likewise, using comparative genomics and transcriptomics between species (e.g Nile and Mozambique tilapia) may identify genes, pathways and gene networks related to adaptation to different environment, which is potentially useful to obtain a novel strain adapted to diverse environment (Wang et al., 2009; Yue and Wang, 2017). Using reference genomes, the aquaculture industry will be able to develop improved broodstock sources using MAS and GS, and bring better varieties to farmers faster than with the conventional selective breeding methods.

However, in comparison to terrestrial livestock, genomic resources in aquaculture generally lag far behind, in particular for sequencing and the assembly of reference genomes; most published genomes of aquaculture species are still in a draft state (Houston et al., 2020; Yue and Wang, 2017). Several economically important aquaculture species remain without a publicly available high-quality reference genome and have limited genomic resources. The first reason may be attributed to fewer funding opportunities for aquaculture species. The second issue may reflect partly the traditionally challenging nature of genome assembly in non-mammalian and non-avian species, particularly for aquatic species with complex genomic features. These include the widespread presence of duplicated loci due to genome duplications of aquatic species, for instance, in salmonids (Lien et al., 2016), common carp (Xu et al., 2014), and sturgeons (Ludwig et al., 2001), and the exceptionally high heterozygosity observed, for example, in Pacific white shrimp (Zhang et al., 2019) and bivalves (Hollenbeck and Johnston, 2018; Plough, 2016). These issues seriously hinder assembly algorithms using short-read sequence data. As a result, many genome assemblies are very fragmented. Ironically, these genome regions can underlie adaptive capacity

and phenotypic plasticity in production environment, and might contribute to genetic regulation of production-relevant traits (Lien et al., 2016; Macqueen et al., 2017).

Therefore, in the future, research on whole genome sequencing in aquaculture should focus on improved reference genomes by filling in the gaps and correcting errors in the genome assembly, and improving functional annotation of these genomes. With the current advances in sequencing technologies, including long range sequencing and high-throughput chromatin conformation capture approaches, mapping technologies, and sequence assembly algorithms are expected to improve the available reference genomes of aquaculture species more effectively (Houston et al., 2020; Yue and Wang, 2017). The updated reference genomes will offer a necessary tool for addressing important issues related to sustainable and profitable aquaculture.

# Chapter 2

# Objectives

# Chapter 2: Objectives

## 1.General objectives

Global change, and particularly climate change, is a real and growing challenge worldwide. Vietnam, more specifically the Mekong Delta, is a region likely to be significantly affected by climate change, especially by saline intrusion from the sea water. Given the importance of striped catfish to the prosperity of the Mekong Delta region, and predictions of increasing saline intrusion into the region, feasible solutions have to be considered. Among the solutions, selecting a new strain of striped catfish with higher salinity tolerance is a promising approach. Although there is no published data available demonstrating that salinity tolerance could be increased by selection on striped catfish, differential growth and survival rates observed when fish were exposed to similar levels of salinity (Huong and Quyen, 2012; Nguyen et al., 2014) suggest the presence of genetic components involved in salinity tolerance traits. If confirmed, these traits could be further improved through a selection program. Therefore, the general objective of my project was to conduct and provide complementary scientific support to a selection program to develop a saline tolerant strain of striped catfish for brackish water environments based on the present freshwater population living in the Mekong Delta, Vietnam.

## 2.Specific objectives

1) The first objective was to develop a new algorithm for parentage assignment using shallow whole genome sequencing data. This new approach was tested on simulated and empirical data. From the method developed in this study will be a precious tool for the process of selective breeding program for salinity tolerance of striped catfish.
2) The second objective was to implement a selection program targeting fish displaying improved tolerance to salinity. After selection, several studies were carried out to evaluate the effectiveness of selection in a next generation. Important genetic parameters such as heritability and response to selection of growth under saline condition were estimated.
3) The third objective was to construct a high-quality reference genome for striped catfish. This work was done by combining HiFi long reads and Hi-C data. The quality of the new version of striped catfish genome was evaluated by several approaches including comparison with a previous draft.
4) The last objective was to uncover the DNA reprogramming in striped catfish during early development through modifications in the DNA methylation patterns, as well as the effects

of salinity and salinity tolerance-based selection program on genes involved in DNA methylation.

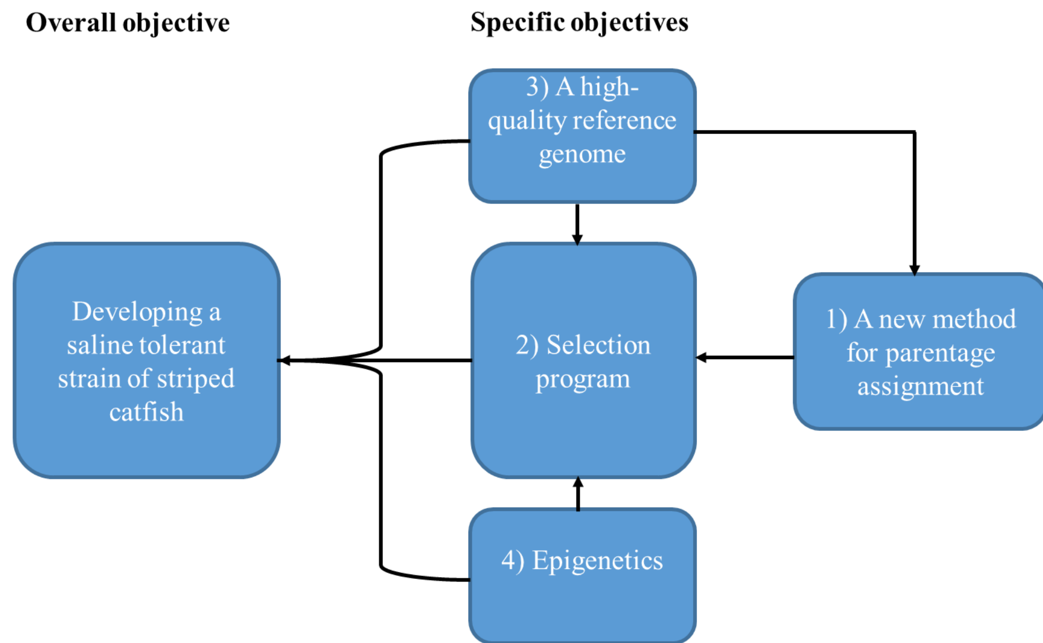


Figure 1: An overview scheme of the four studies in the current project

# Chapter 3

# Parentage assignment using shallow whole genome sequencing data

---

****: Accepted for publication pending satisfactory revision on Molecular Ecology Resources*

Dao Minh Hai [1,2], Duong Thuy Yen[2], Pham Thanh Liem[2], Bui Minh Tam[2], Do Thi Thanh Huong[2], Bui Thi Bich Hang[2], Vo Nam Son[2], Dang Quang Hieu[2], Wouter Coppieters[5], Mutien-Marie Garigliany[3], Patrick Kestemont[4], Nicolas Antoine-Moussiaux[1], Nguyen Thanh Phuong[2], Frédéric Farnir[1].

*[1.] FARAH/Sustainable Animal Production, Faculty of Veterinary Medicine, University of Liege (B43), B-4000 Liege, Belgium.*

*[2.] College of Aquaculture and Fisheries, Can Tho University, Viet Nam.*

*[3.] FARAH/Veterinary Public Health, Faculty of Veterinary Medicine, University of Liege (B43), B-4000 Liege, Belgium.*

*[4.] Research Unit in Environmental and Evolutionary Biology, Institute of Life, Earth & Environnment, University of Namur, Rue de Bruxelles 61, B-5000 Namur, Belgium.*

*[5.] Genomics Platform, GIGA, University of Liege, B-4000 Liege, Belgium*

# Chapter 3: Parentage assignment using shallow whole genome sequencing data

## Preamble

Accurate parentage assignment is essential for the implementation of selective breeding strategies as well as for the estimation of genetic parameters and breeding values in aquaculture, allowing the reconstruction of pedigree information from mixed families, the control of inbreeding and, more generally, of the genetic diversity. Previous studies on parentage assignment have focused on using genetic data generated from microsatellite markers and more recently from SNP arrays. However, in aquaculture species, few SNP arrays have been developed so far. Therefore, in species for which a well-established reference genome is not available, or a proper array has not been developed yet (as is, the case for striped catfish), applying traditional approach for parentage assignment is a challenge. For this reason, in this study, we have developed a new parentage assignment algorithm based on a likelihood approach to identify the most suitable (i.e. likely) set of parents for each offspring using shallow whole genome sequencing data. In order to test the algorithm, we have simulated data and used the likelihood approach to reconstruct the families.

**Abstract**

Pedigree information is important in estimating genetic parameters in selective programs and hatchery management of aquaculture. In this study, we evaluated the use of shallow full genome sequencing (SWGS) data to analyze parentage assignment instead of deep sequencing data or SNP arrays. The use of SWGS data raises two challenges. First, confirmed genotypes are not available for low-coverage (e.g., < 2X). Second, read errors are common in next generation sequencing and more harmful when few reads per marker are available. To address these issues, we have extended a parentage assignment algorithm based on a likelihood approach allowing identification of the most suitable (i.e. likely) pair of relatives for each tested offspring. It also allows to test whether the identified individuals are the parents, relatives, or unrelated individuals. In order to test the algorithm, we have simulated data and used the approach to reconstruct families. The results show that near-perfect assignment can be obtained even for very low coverages if a sufficiently large set of markers is used. A full factorial design involving 60 parents (30 males and 30 females) of striped catfish (*Pangasianodon hypophthalmus*) provides a proof of concept in a realistic situation, with significant attributions of the parents in most cases. Our results show that using SWGS data enables the generation of highly accurate pedigree information using an appropriate algorithm. We make the algorithm available in a standalone program (called *Shallowped*), with a companion script (the generation of simulation data) and a manual.

Keywords: deep full genome sequencing, parentage assignment, shallow full genome sequencing, striped catfish

## 1.Introduction

Accurate parentage assignment is important in many applications, including ecological and evolutionary studies (Kruuk & Hill, 2008). It is also essential for the implementation of selective breeding strategies as well as for the estimation of genetic parameters and breeding values in domestic species. Additionally, in species where parentage relationships are not directly observed, parental assignment procedures allow the reconstruction of pedigree information from mixed families (Hai et al., 2022), the control of inbreeding (Holman, Garcia, Onoufriou, Hillestad, & Johnston, 2017) and, more generally, of the genetic diversity (Coetzer et al., 2017). Examples of such cases are aquaculture species, insects and wild animals: either the size of the offspring at birth, the non-observation of the mating or both prevent a reliable assignment of the parents. Two computational strategies are commonly used for parentage assignment, namely, exclusion-based methods and likelihood-based methods. In short, exclusion-based methods use Mendel's rules to exclude parent-offspring couples for example when the two individuals carry different homozygous genotypes at some loci. Likelihood-based methods provide a solution to deal with genetic data with high uncertainty and can include strategies for determining confidence in the parentage assignments (Jones, Small, Paczolt, & Ratterman, 2010; Vandeputte & Haffray, 2014). This family of methods employs Mendel's rules quantitatively to calculate the likelihoods of different candidate relations among a set of individuals and choose relations that have the maximum likelihood as the best inference (Wang, 2012).

Genomic information comes from genetic markers. Any DNA polymorphism can serve as a genetic marker. For example, single nucleotide polymorphisms (SNPs) are single base pair mutations at any given position of DNA. They are the most abundant DNA sequence polymorphisms (or variants) and are widely distributed along the genomes (Liu et al., 2011). SNPs have several advantages: they are very frequent (approximately 1 every 1000 base-pairs in many genomes, on average), co-dominantly inherited, sequence-tagged markers and highly adaptable to large-scale automated genotyping. According to Anderson and Garza (2006), genotyping around 60 -100 SNPs with high (0.3 - 0.5) minor allele frequency (MAF) is sufficient for accurate parentage assignment.

In recent years, DNA sequencing has become extensively used to deliver genomic information (Lee et al., 2013). Deep whole (or full) genome sequencing (DWGS) refers to techniques of DNA sequencing directly from DNA fragments without the need for cloning in vectors. Using that approach, targeted regions (possibly a complete genome) are sequenced multiple times (the average number is called "the sequencing depth", expressed as nX, where n is the average depth), sometimes hundreds or even thousands of times, leading to the name of "deep sequencing" (Kulski, 2012; Okada et al., 2018). However, DWGS at high

coverage remains expensive (although progressively more affordable) (Whalen et al., 2018), requires a large amount of high quality input DNA. It is associated with large bioinformatics processing time, and the need to store and manipulate large datasets, especially when thousands of individuals are analyzed. Alternatively, shallow whole (or full) genome sequencing (SWGS) follows the same principle, but with a much lower sequencing depth, commonly below 1X. The objective is to reduce the cost and the computer load at the expense of a lower coverage of the targeted region (Kader et al., 2016; Raman et al., 2019). SWGS is amenable to low quantities of input DNA, processing is compatible with low-output instruments, and sequence coverage can be adjusted flexibly according to the purpose of studies (Kader et al., 2016; Parker et al., 2019; Van Roy et al., 2017).

Using genetic data generated from a SNP array for parentage assignment has been implemented on many species (e.g. Holman et al., 2017; Strucken et al., 2016; Tortereau, Moreno, Tosser-Klopp, Servin, & Raoul, 2017). However, in species for which a well-established reference genome is not available, or a proper array has not been developed yet, SWGS data is a flexible alternative to arrays (Pasaniuc et al, 2012). In addition, sequencing is nowadays available in many laboratories while SNP arrays require specialized equipment, longer development times and are species-specific. Since sequencing costs are continuously decreasing (SWGS was estimated to be less than half the price of using a tailored 5K SNP array in our experiment (Illumina, personal communication)), all these factors make genotyping by sequencing a competitive approach. The biggest challenge in using SWGS data is the potentially high uncertainty in the true genotype of an individual: while SNP arrays produce called genotypes, SWGS only generates read counts for the reference and alternative alleles. Calling the genotypes from the read counts is easy when sequence data coverage is high. However, with low sequencing depth used in SWGS data, few reads (often 0, 1 or 2, rarely more) are available at any given locus for each individual. Consequently, only a probabilistic distribution of the genotypes can be obtained, which does not provide unambiguous genotypes calls (Nielsen et al, 2011). For this reason, discriminating between heterozygous and homozygous loci is also particularity challenging (heterozygous individuals are sometime called homozygous). This issue makes parentage assignment more difficult because many parentage assignment algorithms depend on finding opposing homozygous loci to filter out putative parents (Meagher & Thompson, 1986; Whalen et al., 2019). In these situations, the genotypes have to be inferred probabilistically based on the observed reads, and likelihood-based methods become very helpful (Boichard et al., 2014).

In this study, we evaluated the use of SWGS data to perform and test parentage assignment using simulated and empirical data from an aquaculture species, striped catfish *Pangansianodon hypophthalmus* (Sauvage, 1878); Pangasiidae; Siluriformes). To perform parentage assignment, we modified and extended an algorithm able to extract information from such data. This algorithm is based on a likelihood approach

to identify, for each tested individual, the most related male and female within a set of potential relatives and infer the corresponding relationships. This method extends the approach of Whalen et al (2019) and differs from the method of Dodds et al (2019), two studies that also performed parentage assignment using genotyping-by-sequencing (GBS) data. In the study of Whalen et al (2019), likelihood-based methods were developed for relationship classification (including parentage). Our method extends this approach: we introduce a new algorithm to infer allelic frequencies and we develop a hypothesis testing rationale for assessing the identified relationships significance. In the case of Dodds et al (2019), assigning parentage was based on relatedness estimates and expected mismatch rates. In many previous GBS analyses, data were generated using deep sequencing at the same set of loci determined by restriction enzymes across all samples (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Baird et al., 2008; Elshire et al., 2011) with approximately only 1.2% of the genome sequenced (Dodds et al., 2015). To generalize the approach, we have assumed in our study that random loci had been sequenced across the genome with very low sequence depth to obtain the SWGS data (Kader et al., 2016; Parker et al., 2019; Van Roy et al., 2017). Although targeting SWGS data, our algorithm is of course able to use DWGS data as well. In short, our algorithm allows obtaining genotype probabilities distributions based on the observed reads in situations where sets of relatives and offspring are available, as is the case, for example, in selection experiments. Based on these distributions, the algorithm calculates the posterior probability of a set of potential relationships for each offspring and keeps the couple with the highest posterior probability as the most related individuals. In an additional step, we propose to test various hypotheses on the exact relationship between the offspring and the "best" couple. This approach can be used for parentage assignment, but also allows investigating other relationships. It also provides a quantitative assessment of the validity of the estimated relationships using classical p-values. In the next section, we develop the mathematical expression of the likelihood. A computer program - named *Shallowped* - implementing the approach has been written and is freely available.

## 2. Materials and methods

### 2.1 Relationship assignment method

We assume that SWGS data on one (or several) offspring and on $N_M$ males and $N_F$ females (adult animals, or breeders in a selection experiment for example) is available. For each offspring, we aim at finding the most related male and female within this set of potential relatives. In the following description of the method, we consider that SWGS data consists of the number of reads for each allele of each used marker for each individual. Since we are focusing on shallow sequencing, these numbers of reads will generally be very small (for low coverages, they are very often equal to 0). Although we could easily extend

the approach to markers with more alleles, we will consider biallelic markers (such as SNPs or InDels) in the description, because these markers are the most abundant in the genome and allow for expanding the number of used markers as much as needed. In addition to obtaining the most related individuals within the proposed set of relatives, we will obtain an assessment of the nature of these relations. The next section describes a likelihood approach that allows reaching these goals. Applications of the method on simulated and empirical data are described next.

**2.1.1 Likelihood approach**

Our approach is similar to the approach developed in Whalen's (2018) paper, but with differences and extensions that we will detail. We use the following notation:

- $d_i(s)$ is the available genomic information of individual i for marker s. With (shallow) sequencing data, $d_i(s)$ is the number of reads for all alleles of marker s. $\boldsymbol{d_i}$ is the complete genomic information for individual i (i.e. $\boldsymbol{d_i} = [d_i(1), d_i(2), \dots, d_i(M)]$ where M is the number of markers).
- $g_i(s)$ is the (unknown) genotype of individual i for marker s and $\boldsymbol{g_i} = [g_i(1), g_i(2), \dots, g_i(M)]$ is the complete (unknown) genotype of individual i.
- $R(o, m, f)$ represents a relationship between an offspring *o*, a male *m* and a female f. Although many relations are possible, we have limited the number of relations to few possibilities, similarly to Whalen (2018). In this study, the tested male (female) is either the father FAT (mother MOT), the paternal grandfather PGF (maternal grandmother MGM), a full-sib of the father FSF (mother FSM), a half-sib of the father HSF (mother HSM), or unrelated to the father UNF (mother UNM). We could extend the approach proposed below to other relationships if needed but, for parental assignment, these five relationships should be sufficient in most cases. Note also that, although 25 combinations of these relationships are possible, we have limited the development (and the computations) to the combinations where at least one of the tested male or female is the father or mother. We will thus work with the following combinations:

$$\boldsymbol{R} = [(MOT|FAT),\ (MGM|FAT),\ (FSM|FAT),\ (HSM|FAT),\ (UNM|FAT), (FAT|MOT), (PGF|MOT), (FSF|MOT), (HSF|MOT), (UNF|MOT)]$$

For example, $R = (FAT|MOT)$ means that we are assuming that the tested male is the father, given that the tested female is the mother.

We will make the reasoning for the 5 relationships which assume that the tested female is actually the mother (the relationships with $MOT$ in $\boldsymbol{R}$), but the reasoning is identical *mutatis mutandis* for the

situations assuming that the tested male is actually the father (the relationships with $FAT$). Since we want to identify the best relationship for a trio using genomic information, we want to maximize the posterior probability of a relationship given the genomic information of the trio:

$$\max_R p(R|\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f})$$

Using Bayes' theorem, we write this probability as:

$$p(R|\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}) = p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}|R) * \frac{p(R)}{p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f})}$$

We "demarginalize" the denominator to obtain:

$$p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}) = \sum_R p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}, R) = \sum_R p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}|R) * p(R)$$

The posterior probability of the relationship is thus:

$$p(R|\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}) = \frac{p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}|R) * p(R)}{\sum_R p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}|R) * p(R)}$$

In the absence of other information, we could consider the prior probability of the relationships as uniform, making the posterior probability equal to:

$$p(R|\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}) = \frac{p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}|R)}{\sum_R p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}|R)}$$

We next need to derive the conditional probabilities $p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}|R)$ for the various tested relationships. We start with $R_{FATMOT} = (FAT|MOT)$.

As mentioned earlier, we do not know the exact genotypes of the individuals: we only know sequencing information. Consequently, these genotypes ($\boldsymbol{g_o}, \boldsymbol{g_m}, \boldsymbol{g_f}$) have to be inferred probabilistically. We use the same trick as above to obtain:

$$p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}|R_{FATMOT}) = \sum_{g_m} \sum_{g_f} \sum_{g_o} p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}, \boldsymbol{g_o}, \boldsymbol{g_m}, \boldsymbol{g_f}|R)$$

We can easily condition the genomic data on the genotypes:

$$p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f} | R_{FATMOT}) = \sum_{g_m} \sum_{g_f} \sum_{g_o} p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f} | R_{FATMOT}, \boldsymbol{g_o}, \boldsymbol{g_m}, \boldsymbol{g_f}) * p(\boldsymbol{g_o}, \boldsymbol{g_m}, \boldsymbol{g_f} | R_{FATMOT})$$

We simplify the first factor, because the genomic data from one individual only depends on this individual's genotype, and are conditionally independent of each other and of the assumed relationship:

$$p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f} | R_{FATMOT}, \boldsymbol{g_o}, \boldsymbol{g_m}, \boldsymbol{g_f}) = p(\boldsymbol{d_o} | \boldsymbol{g_o}) * p(\boldsymbol{d_m} | \boldsymbol{g_m}) * p(\boldsymbol{d_f} | \boldsymbol{g_f})$$

The second factor can also be rewritten:

$$\begin{aligned} p(\boldsymbol{g_o}, \boldsymbol{g_m}, \boldsymbol{g_f} | R_{FATMOT}) &= p(\boldsymbol{g_o} | R_{FATMOT}, \boldsymbol{g_m}, \boldsymbol{g_f}) * p(\boldsymbol{g_m}, \boldsymbol{g_f}) \\ &= p(\boldsymbol{g_o} | R_{FATMOT}, \boldsymbol{g_m}, \boldsymbol{g_f}) * p(\boldsymbol{g_m}) * p(\boldsymbol{g_f}) \end{aligned}$$

The last equality is true if we assume that these two parents are not related.

The probabilities $p(\boldsymbol{g_o} | R_{FATMOT}, \boldsymbol{g_m}, \boldsymbol{g_f})$ are easy to compute using Mendel's rules, since we assume that m and f are o's parents (Table 1A).

We can thus rewrite the posterior probability as follows:

$$\begin{aligned} & p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f} | R_{FATMOT}) \\ & \quad = \sum_{g_m} \Bigg\{ p(\boldsymbol{d_m} | \boldsymbol{g_m}) * p(\boldsymbol{g_m}) \\ & \quad * \sum_{g_f} \Bigg\{ p(\boldsymbol{d_f} | \boldsymbol{g_f}) * p(\boldsymbol{g_f}) * \sum_{g_o} \{ p(\boldsymbol{d_o} | \boldsymbol{g_o}) * p(\boldsymbol{g_o} | R_{FATMOT}, \boldsymbol{g_m}, \boldsymbol{g_f}) \} \Bigg\} \Bigg\} \end{aligned}$$

If we write $F(\boldsymbol{g_m}) = \sum_{g_f} \{ p(\boldsymbol{d_f} | \boldsymbol{g_f}) * p(\boldsymbol{g_f}) * \sum_{g_o} \{ p(\boldsymbol{d_o} | \boldsymbol{g_o}) * p(\boldsymbol{g_o} | R_{FATMOT}, \boldsymbol{g_m}, \boldsymbol{g_f}) \} \}$, the posterior probability becomes:

$$p(\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f} | R_{FATMOT}) = \sum_{g_m} \{ p(\boldsymbol{d_m} | \boldsymbol{g_m}) * p(\boldsymbol{g_m}) * F(\boldsymbol{g_m}) \}$$

To simplify the computations, we will compute these probabilities assuming that the markers are independent, which is only approximately true when large number of markers are considered, due to physical linkage between the markers (Whalen, 2018). Accordingly, the final probability will be the product over all markers of the single marker probabilities. We will therefore make the derivation for one marker s, and we will omit the reference to the marker in most of the following equations to lighten the writing.

To evaluate this probability, we still need to compute probabilities $p(\boldsymbol{d}|\boldsymbol{g})$ and $p(\boldsymbol{g})$. We start with the first one. For each marker s, $d(s)$ only depends on $g(s)$. If we assume a biallelic marker (such as a SNP), $d(s) = [n_A(s), n_B(s)]$, where $n_A(s)$ stands for the number of reads bearing the A allele and $n_B(s)$ stands for the number of reads bearing the B allele. We will assume that each allele is equally likely to be transmitted to an offspring. In addition, we postulate that each allele will be sequenced correctly with a probability $(1-\varepsilon)$ and incorrectly (A allele read as a B allele, or B allele read as a A allele) with a probability $\varepsilon$, where $\varepsilon$ represents a (assumed constant across all markers) sequencing error rate. We can use a binomial distribution to obtain the needed conditional probabilities (Van Raden, 2015):

$$\begin{cases} p(n_A, n_B|g = AA) = C_{n_A+n_B}^{n_A} * (1-\varepsilon)^{n_A} * \varepsilon^{n_B} \\ p(n_A, n_B|g = AB) = C_{n_A+n_B}^{n_A} * 0.5^{n_A+n_B} \\ p(n_A, n_B|g = BB) = C_{n_A+n_B}^{n_A} * (1-\varepsilon)^{n_B} * \varepsilon^{n_A} \end{cases}$$

Finally, we have to evaluate prior genotypes probabilities $p(\boldsymbol{g})$. If we assume that the parental population is in Hardy-Weinberg equilibrium, we can obtain these probabilities based on the allelic frequencies $f_A$ and $f_B = 1 - f_A$. When the genotypes are available for the parents, these frequencies can be obtained by counting the number of alleles. Unfortunately, with shallow sequencing data, we don't have the full information, and an alternative solution might be necessary. An idea is the following: for each marker, we would like to compute $f_A$ as:

$$f_A = \frac{\sum_{p=1}^{p=N}\left[p\left(g_p = AA \middle| d_p\right) * 2 + p\left(g_p = AB \middle| d_p\right) * 1\right]}{2 * N}$$

where the sum is over all N parents. We can again use:

$$p(g|d) = p(d|g) * \frac{p(g)}{p(d)} = \frac{p(d|g) * p(g)}{\sum_g p(d|g) * p(g)}$$

Replacing the conditional probabilities in the denominator by these expressions and using the $p(\boldsymbol{d}|\boldsymbol{g})$ derived above, we note that $f_A$ depends on $p(g)$, and $p(g)$ depends on $f_A$. This suggests an iterative procedure, where new values of $f_A$ are computed using previous values of $f_A$:

$$f_A^{(0)} \to p(g) \to p(g|d) \to f_A^{(1)}$$

We start with a random initial value for $f_A^{(0)}$ (for example 0.5) and iterate the process until $\left|f_A^{(i+1)} - f_A^{(i)}\right| < \tau$, where $\tau$ is a small value (for example, $\tau = 0.001$).

This completes the computation of the probabilities for $R_{FATMOT} = (FAT|MOT)$.

The next relationship probability to compute is for $R_{PGFMOT} = (PGF|MOT)$ (i.e. assuming that $\boldsymbol{d_m}$ corresponds to sequencing data from *o*'s paternal grandfather and $\boldsymbol{d_f}$ corresponds to sequencing data from o's mother). The derivation is similar to the previous one, except that we need to include a summation over the possible father's genotypes $g_{fat}$:

$$p\left(d_o, d_m, d_f | R_{PGFMOT}\right) = \sum_{g_m} \sum_{g_{fat}} \sum_{g_f} \sum_{g_o} p\left(d_o, d_m, d_f, g_o, g_m, g_f, g_{fat} | R_{PGFMOT}\right)$$

This leads to:

$$p\left(d_o, d_m, d_f | R_{PGFMOT}\right) = \sum_{g_m} \left\{ p(d_m|g_m) * p(g_m) * \sum_{g_{fat}} \left\{ p\left(g_{fat}|g_m\right) * F\left(g_{fat}\right) \right\} \right\}$$

Since *m* is the paternal grandfather, probabilities $p\left(g_{fat}|g_m\right)$ can be derived easily (Table 1B).

For the next probability, the relationship is $R_{FSFMOT} = (FSF|MOT)$. To obtain the corresponding probability, we have to sum over the paternal grandparents genotypes $g_{pgf}$ and $g_{pgm}$, and obtain:

$$p(d_o, d_m, d_f | R_{FSFMOT}) = \sum_{g_{pgf}} \left\{ p(g_{pgf}) * \sum_{g_{pgm}} \left\{ p(g_{pgm}) * \left( \sum_{g_m} \{ p(g_m | g_{pgf}, g_{pgm}) * p(d_m | g_m) \} \right) * \left( \sum_{g_{fat}} \{ p(g_{fat} | g_{pgf}, g_{pgm}) * F(g_{fat}) \} \right) \right\} \right\}$$

For $R_{HSFMOT} = (HSF|MOT)$, the probability is:

$$p(d_o, d_m, d_f | R_{HSFMOT}) = \sum_{g_{pgf}} \left\{ p(g_{pgf}) * \left( \sum_{g_m} \{ p(g_m | g_{pgf}) * p(d_m | g_m) \} \right) * \left( \sum_{g_{fat}} \{ p(g_{fat} | g_{pgf}) * F(g_{fat}) \} \right) \right\}$$

Finally, for the relationship $R_{UNF} = (UNF|MOT)$ where the tested male is not related to the real father, we obtain:

$$p(d_o, d_m, d_f | R_{UNFMOT}) = \left( \sum_{g_m} \{ p(g_m) * p(d_m | g_m) \} \right) * \left( \sum_{g_m} \{ p(g_{fat}) * F(g_{fat}) \} \right)$$

Using the five conditional probabilities derived so far, we can now obtain posterior probabilities for each of the assumed relationships. For example, the posterior probability that the tested male is the father, assuming that the tested female is the mother, is:

$$p(R_{FATMOT} | \boldsymbol{D}(s) = [d_o(s), d_m(s), d_f(s)]) = \frac{p(\boldsymbol{D}(s) | R_{FATMOT})}{p(\boldsymbol{D}(s) | R_{FATMOT}) + p(\boldsymbol{D}(s) | R_{PGFMOT}) + p(\boldsymbol{D}(s) | R_{FSFMOT}) + p(\boldsymbol{D}(s) | R_{HSFMOT}) + p(\boldsymbol{D}(s) | R_{UNFMOT})}$$

Considering all markers, we can calculate:

$$p\left(R_{FATMOT} \middle| \boldsymbol{D}[\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}]\right)$$
$$= \frac{\prod_s p(\boldsymbol{D}(s)|R_{FATMOT})}{\prod_s p(\boldsymbol{D}(s)|R_{FATMOT}) + \prod_s p(\boldsymbol{D}(s)|R_{PGFMOT}) + \prod_s p(\boldsymbol{D}(s)|R_{FSFMOT}) + \prod_s p(\boldsymbol{D}(s)|R_{HSFMOT}) + \prod_s p(\boldsymbol{D}(s)|R_{UNFMOT})}$$

Alternatively, although the meaning of this expression is slightly different from the meaning of the previous one, we could compute the probability as a geometric mean over the markers to ease the computations:

$$p\left(R_{FATMOT} \middle| \boldsymbol{D} = [\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}]\right) = \sqrt[M]{\prod_{s=1}^{M} p\left(R_{FATMOT} \middle| \boldsymbol{D} = [d_o(s), d_m(s), d_f(s)]\right)}$$

Using logarithms of this expression leads to the following formula, easier to compute when the number of markers gets large (potentially leading to underflows in the first formula):

$$log\left[p\left(R_{FATMOT} \middle| \boldsymbol{D} = [\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}]\right)\right] = \frac{\sum_{s=1}^{s=M} log\left[p\left(R_{FATMOT} \middle| \boldsymbol{D} = [d_o(s), d_m(s), d_f(s)]\right)\right]}{M}$$

After scanning all the males from the potential male parents set and computing the corresponding probabilities, we can identify the most likely father (we could also test other relationships) as the one with the highest posterior probability $p\left(R_{FATMOT} \middle| \boldsymbol{D} = [\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}]\right)$. A similar procedure for the females leads to the most likely couple.

Table 1: Probabilities of the offspring genotypes given both parents (*f, m*) genotypes (*A*) or one parent (*p*) genotype (*B*).

**(A)**

| $P(G_o = AA \mid G_f, G_m)$ | | $G_f$ | | |
|---|---|---|---|---|
| | | $AA$ | $AB$ | $BB$ |
| $G_m$ | $AA$ | 1.00 | 0.50 | 0.00 |
| | $AB$ | 0.50 | 0.25 | 0.00 |
| | $BB$ | 0.00 | 0.00 | 0.00 |
| $P(G_o = AB \mid G_f, G_m)$ | | $G_f$ | | |
| | | $AA$ | $AB$ | $BB$ |
| $G_m$ | $AA$ | 0.00 | 0.50 | 1.00 |
| | $AB$ | 0.50 | 0.50 | 0.50 |
| | $BB$ | 1.00 | 0.50 | 0.00 |
| $P(G_o = BB \mid G_f, G_m)$ | | $G_f$ | | |
| | | $AA$ | $AB$ | $BB$ |
| $G_m$ | $AA$ | 0.00 | 0.00 | 0.00 |
| | $AB$ | 0.00 | 0.25 | 0.50 |
| | $BB$ | 0.00 | 0.50 | 1.00 |

**(B)**

| $P(G_o \mid G_p)$ | | $G_p$ | | |
|---|---|---|---|---|
| | | $AA$ | $AB$ | $BB$ |
| $G_o$ | $AA$ | $f_A$ | $0.5 * f_A$ | 0 |
| | $AB$ | $f_B$ | 0.5 | $f_A$ |
| | $BB$ | 0 | $0.5 * f_B$ | $f_B$ |

### 2.1.2 Significance testing

Obtaining the most likely couple of relatives for a given offspring does not tell much about the exact relationships between these individuals. We have therefore developed strategies to test various hypotheses about the exact relationships with the following rationale.

Assume we want to know whether the best (most likely) male is the father, given that the mother has been correctly identified (as in the previous paragraph, we could easily extend the approach to other possibilities for the female). We start by deriving the father’s genotype distribution:

$$p(g_m|d_o, d_f, R_{FATMOT}) = \sum_{g_f}\sum_{g_o} p(g_f, g_m, g_o|d_o, d_m, R_{FATMOT})$$

$$= \sum_{g_f}\sum_{g_o} p(d_o, d_m|g_f, g_m, g_o, R_{FATMOT}) * \frac{p(g_f, g_m, g_o, R_{FATMOT})}{p(d_o, d_m)}$$

Using the same derivation as above, we obtain the following expression:

$$p(g_m|d_o, d_f, R_{FATMOT}) = \mathrm{C} * p(g_m) * F(g_m)$$

where $\mathrm{C} = \frac{1}{p(d_o, d_m)}$ is a constant, common to all genotypes. For example:

$$p(g_m = AA|d_o, d_f, R_{FATMOT}) = \frac{p(g_m = AA) * F(g_m = AA)}{\sum_{g_m} p(g_m) * F(g_m)}$$

Using this result, we can sample from this distribution to obtain a father's genotype $\boldsymbol{g_m}$, use this genotype to simulate reads $\boldsymbol{d_m}$ with a coverage similar to the putative father, and use these genomic data to compute $p(R_{FATMOT}|\boldsymbol{D} = [\boldsymbol{d_o}, \boldsymbol{d_m}, \boldsymbol{d_f}])$. Repeating this random sampling a large number of times will generate a distribution of these posterior probabilities. We name this distribution $p_{FATMOT}$. Comparing the real posterior probability to this distribution provides a p-value, allowing a test of the null hypothesis that the best male is actually the father, assuming that the best female is the mother.

Other hypotheses on the best male can be tested similarly: we need to obtain the corresponding probabilities for the male genotype assuming the tested relationships, and use the same procedure as above.

Since the derivation of these conditional genotypes probabilities is similar to the previous, we only give the results:

$$p(g_m|d_o, d_f, R_{PGFMOT}) = \mathrm{C} * p(g_m) * \sum_{g_{fat}} \{p(g_{fat}|g_{pgf}) * F(g_{fat})\}$$

$$p(g_m|d_o, d_f, R_{FSFMOT}) = C * \sum_{g_{pgf}} \left\{ p(g_{pgf}) * \sum_{g_{pgm}} \left\{ p(g_{pgm}) * p(g_m|g_{pgf}, g_{pgm}) * \sum_{g_{fat}} \{ p(g_{fat}|g_{pgf}, g_{pgm}) * F(g_{fat}) \} \right\} \right\}$$

$$p(g_m|d_o, d_f, R_{HSFMOT}) = C * \sum_{g_{pgf}} \left\{ p(g_{pgf}) * p(g_m|g_{pgf}) * \sum_{g_{fat}} \{ p(g_{fat}|g_{pgf}) * F(g_{fat}) \} \right\}$$

$$p(g_m|d_o, d_f, R_{UNFMOT}) = p(g_m)$$

### 2.2 Simulations

#### 2.2.1 Accuracy, power and false positive rate

We have used simulated data to evaluate the accuracy and significance of the relationships assignments, and to compare the results provided by the program AlphaAssign, implementing the approach of Whalen (2018), to our results. Since both approaches are similar except for the significance assessment, we restricted our simulations to situations where coverages allow for a comparison of the two approaches: for high coverages, both methods lead to similar correct results with high power, and for very low coverages, the power is very low and the differences become very small. Consequently, we used coverages of 1 and 0.2 for the parents and the offspring, respectively, and we used pedigree with 30 males, 30 females and 100 offspring in each configuration. Since we did not observe significant effects when changing the sequencing error rate from 1% to 5% (data not shown), we used the default 1% sequencing error rate in all simulations. We have simulated several situations:

- Parents were both present in the putative parents list (noted F_M).
- The father was absent from the list, while the mother was present (f_M).
- The father and a full-sib of the father were present in the list, while the mother was present (FSF_F_M).

- The father was absent from the list while a full-sib of the father and the mother were present (FSF_f_M).
- The last three scenarios, but with the mother absent (f_m, FSF_F_m, FSF_f_m).

In a first set of simulations, we varied the number of markers from low (200) to average (1000) and to high (5000), and we generated the markers genotypes independently (i.e. assuming no linkage between the markers), which might be overoptimistic in real situations when the number of markers becomes large (i.e. 1000 and 5000 markers). Therefore, we also simulated situations with linked markers to see the impact of linkage and/or linkage disequilibrium on the performances. For linked markers, we generated the parents' genotypes randomly, but we considered a 1% recombination rate between successive markers for the transmission from the parents to the offspring. To generate linkage disequilibrium, in addition to using linked or unlinked markers, we used drift in a fixed size population to generate the genotypes of the parents. We first generated randomly the marker genotypes in the first generation, and generated each individual of the next generation by randomly sampling in the previous generation. After 20 generations, we used the individuals from the last generation to generate the parents. We measured linkage disequilibrium using D' values computed between successive markers (Lewontin, 1964).

In all our simulations, we estimated accuracy, power and false positive rates as follows:

- The accuracy is the proportion of correctly identified relationships, even when the assignment was not significant. The meaning of "correctly identified relationships" is the following:
    - When one parent was present, the method had to identify that parent.
    - When a parent was not present while a full-sib of that parent was, the method had to identify that full-sib.
- The power (for $\alpha = 5\%$) is the proportion of correctly and significantly assigned relationships. We use the following definitions of "significance":
    - In *Shallowped*, a father assignment was significant when the p-value in the $p_{FATMOT}$ distribution was above 5%. For a full-sib assignment, the p-value in the $p_{FSFMOT}$ distribution had to be between 2.5% and 97.5 while the p-value in the $p_{UNFMOT}$ distribution was below 5%. For an unrelated father, the p-value in the $p_{UNFMOT}$ distribution was above 5% while the p-value in the $p_{FSFMOT}$ distribution was below 5% (see Figure 1 for a visual interpretation of these rules). Similar rules hold for the mother.
    - In AlphaAssign, we used the closest reported value from the file listing the expected score value for a parent, a full-sib of the parent or an unrelated individual. Note that this definition

of significance is less stringent than the definition used in AlphaAssign to assign parents. Our definition might lead to an increased power, but also to an increased false positive rate.

- The false positive rate is the proportion of significantly assigned, but incorrect parents.

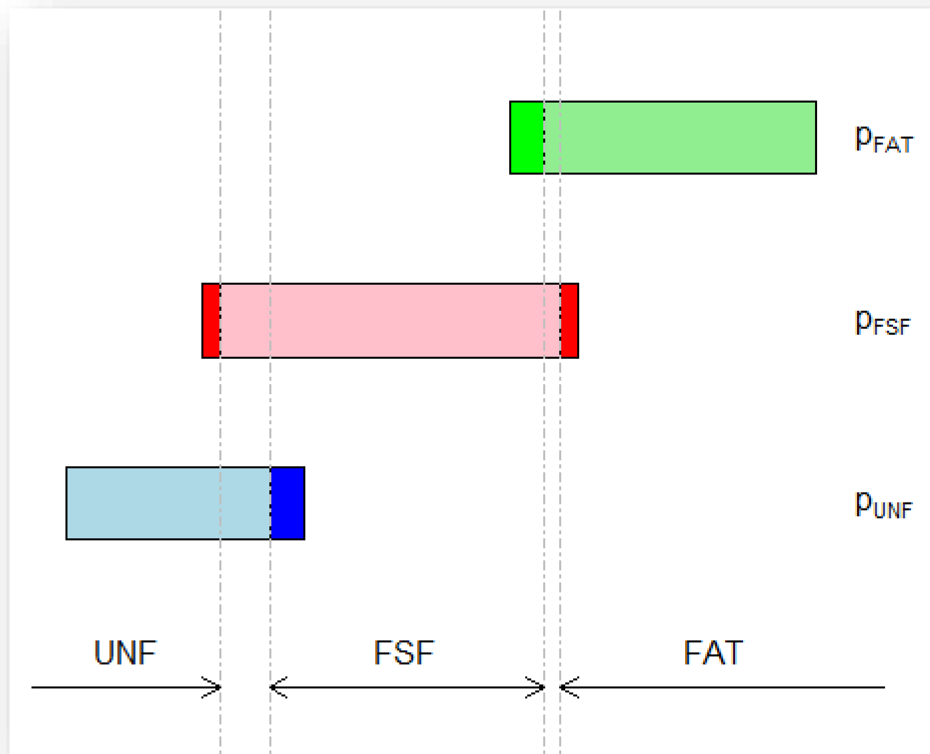


Figure 1: graphical representation of the criteria used to take statistical decisions. The horizontal axis represents likelihoods values. The colored rectangles represent the spans of the distributions identified on the right. The regions containing the extremities of the distributions (unilateral for $p_{UNF}$ and $p_{FAT}$, bilateral for $p_{FSF}$) are shown using darker colors. The arrows at the bottom of the graph show the decision (UNF (unrelated to father), FSF (full-sib of the father) or FAT (father)) given the likelihood value for the tested male.

### 2.2.2 Bias due to differing coverages among the parents

To check the behavior of the method in the presence of different amount of information among the parents, we have also simulated situations with different coverages among the parents. In our simulations, we used a 5-fold difference between the "high" and "low" coverages in two scenarios: one with relatively high coverages ("high" = 5x and "low" = 1x) and one with low coverages ("high" = 1x and "low" = 0.2x).

We simulated half of the male parents and half of the female parents with high coverages and the other halves with low coverages. We also simulated offspring with either high or low coverages. In all these simulations, we used 2000 markers. We selected these coverages and number of markers because higher genomic information (higher coverages and/or more markers) would lead to situations similar to DWGS, and lower genomic information would lead to very scarce data, low power and consequently, difficulties to assess the exact effect of the various densities among the parents.

### 2.3 Empirical data

#### 2.3.1 Experimental fish and DNA sampling

Empirical data originated from a striped catfish selection experiment (details provided in selection part of Chapter 4). In short, this work has been conducted in collaboration with three striped catfish hatcheries in the Mekong Delta, Viet Nam from July/2017 to September/2018. Firstly, broodstock of striped catfish were selected from three different sources (An Giang, Vinh Long and Can Tho province, with 10 males and 10 females in each hatchery). A piece of caudal-fin (~1 $cm^2$) from each fish (*i.e.* for a total of 60 broodstock) was collected and preserved in 95% ethanol for genetic analyses.

After that, spawning was induced by injection with human chorionic gonadotrophin. Twenty-nine females (~ 3,000 eggs of each female) were fully crossed with 30 males to create 870 families (one female from the Can Tho hatchery could not be used due to the low egg quality). From each family, 2,000 good quality larvae were selected for nursing (leading to 1,740,000 larvae in total). After one year, 500 individuals selected from the remaining offspring were used for parentage assignment. A piece of caudal-fin (~1 $cm^2$) from each fish was collected and preserved in 95% ethanol for genetic analyses.

#### 2.3.2 DNA extraction

In a preliminary step, due to the lack of genomic information for this species at the beginning of our experiment, we sampled a striped catfish individual (blood sample) in order to obtain a *de novo* draft reference sequence. Genomic DNA was extracted from blood of this fish using the Maxwell® 16 Blood DNA Purification Kit (Promega, USA).

For the 60 broodstocks and 500 offprings, genomic DNA were extracted using Wizard® SV Genomic DNA Purification kit (Promega, USA). DNA quality and quantity were evaluated using electrophoresis on a 1% agarose gel, and using Quant-iT™ PicoGreen ® dsDNA Reagent and Kits (Thermo

Fisher Scientific, USA). All DNA samples were sent to the Interdisciplinary Center for Biomedical Research (GIGA) at the University of Liege, Belgium for sequencing.

### 2.3.3 Deep and shallow whole genome sequencing

Deep sequencing was applied for only one striped catfish to build a *de novo* draft reference genome. Paired-end (PE) libraries with 350 and 550 bp fragments were constructed using a TruSeq DNA PCR-Free Kit (Illumina) according to manufacturer protocols. PE libraries were sequenced using Illumina NextSeq 500 sequencing platform with Illumina protocols for whole-genome shotgun sequencing with read lengths (~ 150 bp) and fold coverage (~ 144 X). A set of 663.7 million of reads has been obtained, leading to a total of 99.7 Gbp. After filtering out low quality reads using usual quality controls, the remaining reads have been assembled *de novo* using CLC Genomics Workbench 11. In the current state, we have a *de novo* draft genome (~ 691 Mb), assembled into 35,401 contigs with size range from 1.000 to 268.333 bp.

For shallow sequencing, PE libraries with average fragments (~ 570 bp) were constructed using iGenomX RipTide High Throughput Rapid DNA Library Prep Kit based on manufacturer protocols. The targeted fold coverage was from 1 to 2X for the 60 parents and ~ 0.5X for the 500 selected offspring. PE libraries were sequenced using the Illumina NovaSeq 6000 sequencing platform with Illumina protocols with read lengths (~ 150 bp). A total of 121.65 Gbp with 811 million of reads and 864.45 Gbp with 5,763 million of reads were generated for 60 parents and 500 offspring, respectively. Quality of raw sequencing reads was assessed using FastQC v.0.11.8 (Andrews, 2010) before using for downstream analysis.

### 2.3.4 SNP identification and filtering

Although a completed sequence is not available yet for striped catfish, we used the 60 parents to infer SNPs in the current sequence. We used the remaining reads after the quality control steps and mapped them to our draft genome using BWA v.0.7.17 (Li & Durbin, 2009). Then, SAMtools and BCFtools v.0.1.19 (Danecek et al., 2011; Li et al., 2009) were used to identify genomic variants, including SNPs. This procedure led to the discovery of more than 5,900,000 variants.

High-quality SNP selection from these variants was performed in multiple steps using different criteria with the aims of achieving high mapping quality, informative SNPs and adequate spacing across the entire genome. Self-made scripts were used to perform the selection according to the following criteria. Firstly, in order to minimize the effect of sequencing errors and bioinformatics artefacts, only SNPs with mapping quality equal to 60 and coverage depth (across all samples) between 80 and 150 were selected.

Secondly, to increase the power of parentage assignment, SNPs with a MAF between 0.4 and 0.5 were preferentially selected. Thirdly, to reduce the effects of physical linkage between SNPs, SNPs were chosen on different contigs (35,401 contigs make up the used reference sequence). After filtering, 26,726 SNPs were used for downstream analyses for the 500 offspring.

#### 2.3.5 Obtaining reads for these SNPs from the parents and offspring

Another self-made program was implemented for these analyses. Firstly, the program read the 26,726 selected SNPs and the draft reference sequence to identify the reference allele for each SNP using this reference. Next, the SAM files of each parent and offspring were read to extract the reads fully matching the reference sequence and harboring a SNP allele. Finally, the program counted the number of (reference, alternates) alleles in the selected reads.

#### 2.3.6 Obtaining the parents from the genotypes

At this stage, all read counts for the 60 parents and 500 offspring were available. To avoid sequencing errors, we also removed markers for which we had only observed one read carrying one of the two alleles, or when more than two alleles were present. Since this could differ for each offspring, we performed this last filter individually, leading to 20,822 to 20,959 informative SNPs (out of 26,726) conserved for parentage analysis.

We have used this dataset to compare the results obtained using *Shallowped* and *AlphaAssign* (Whalen, 2018) in a more realistic context. For *Shallowped*, we considered father assignments as significant when the p-value for the FAT-MOT hypothesis was larger than 5% while the p-values for the other tested hypotheses (FSF-MOT, HSF-MOT and UNF-MOT) were smaller than 5% (and similarly for the mothers). For *AlphaAssign*, we either used the default (parents reported in the pedigree file) or we chose the hypothesis (either the tested individual is the parent, a full-sib of the parent or an unrelated individual) with the estimated score closest to the tested individual's score.

To check for the effect of the varying coverages in the parents on the allocation in this specific dataset, we have performed a second analysis where the coverage for all parents was set to the same value, equal to the lowest coverage among the parents. To that end, we have randomly discarded reads from all parents to end up with the same number of reads for each parent, and then rerun the allocation program and compared the result to the results with the complete set of reads.

## 3. Results

### 3.1 Implementation of the method

A fortran 95 program (named *Shallowped*) implements the method described above. The program obtains the best male and female relatives and the corresponding p-values using hypotheses specified in a parameters file. A manual (Supplementary data in appendices) describes the installation and the use of the programs and gives examples. We also provide companion (perl) script (named “mksimul.pl”) to generate simulated datasets. The program and script are freely available at the following link: https://doi.org/10.5281/zenodo.6033536

For our analyses, we have run the analyses on a Ubuntu 20.04 LTS implementation of Linux on an Intel(R) Core(TM) i7-8665U CPU @ 1.90GHz 2.11 GHz. Running time scales linearly with the number of markers, the number of offspring and the size of $F \times M$. In an example with 2,000 markers, coverages of ~2, 30 males × 30 females (900 likelihoods evaluations), 100 tested individuals, tests of four hypotheses for both males and females using 200 samples each, the program took 504.9 s to complete the analyses (and obtain correct and significant assignments), which averages to 5 s per individual.

### 3.2 Simulation results

#### 3.2.1 Accuracy, power and false positive rates in the parentage assignments

Table 2: Results of the simulations on accuracy (ACC), power (POW, α=5%) and false positive rates (FPR, α=5%) as functions of the tested scenarios, the marker numbers and the program (AA = *AlphaAssign*, SH = *Shallowped*). Markers were biallelic and generated independently. Definitions of ACC, POW and FPR are provided in the Material and Methods section. In scenarios, F (f) represent situations where the father was present (absent) among the males, FS represent situations where a full-sib of the father was present among the males, and M (m) situations where the mother was present (absent) among the females. Used coverages were 1.0 for the parental cohorts, and 0.2 for the offspring.

| | | Males | | | | | | Females | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | # markers | 200 | | 1000 | | 5000 | | 200 | | 1000 | | 5000 | |
| Scenario | Program | AA | SH | AA | SH | AA | SH | AA | SH | AA | SH | AA | SH |
| F_M | ACC | 0,02 | 0,02 | 0,02 | 0,01 | 1,00 | 1,00 | 0,01 | 0,01 | 0,01 | 0,02 | 1,00 | 1,00 |
| | POW | 0,01 | 0,00 | 0,00 | 0,00 | 0,07 | 0,84 | 0,01 | 0,00 | 0,00 | 0,00 | 0,08 | 0,87 |
| | FPR | 0,92 | 0,41 | 0,20 | 0,36 | 0,00 | 0,00 | 0,92 | 0,43 | 0,18 | 0,36 | 0,00 | 0,00 |
| f_M | ACC | | | | | | | 0,19 | 0,18 | 0,81 | 0,79 | 1,00 | 1,00 |
| | POW | | | | | | | 0,17 | 0,07 | 0,27 | 0,47 | 0,07 | 0,86 |
| | FPR | 1,00 | 0,41 | 1,00 | 0,03 | 1,00 | 0,98 | 0,70 | 0,32 | 0,00 | 0,03 | 0,00 | 0,00 |
| FS+F_M | ACC | 0,23 | 0,23 | 0,68 | 0,66 | 1,00 | 1,00 | 0,31 | 0,27 | 0,69 | 0,70 | 1,00 | 1,00 |
| | POW | 0,23 | 0,14 | 0,19 | 0,34 | 0,06 | 0,86 | 0,27 | 0,17 | 0,19 | 0,34 | 0,02 | 0,90 |
| | FPR | 0,69 | 0,38 | 0,02 | 0,06 | 0,00 | 0,00 | 0,61 | 0,30 | 0,02 | 0,05 | 0,00 | 0,00 |
| FS+f_M | ACC | 0,07 | 0,09 | 0,22 | 0,22 | 0,81 | 0,83 | 0,20 | 0,24 | 0,74 | 0,73 | 1,00 | 1,00 |
| | POW | 0,00 | 0,00 | 0,21 | 0,00 | 0,81 | 0,78 | 0,18 | 0,09 | 0,25 | 0,40 | 0,06 | 0,93 |
| | FPR | 0,93 | 0,36 | 0,78 | 0,00 | 0,19 | 0,17 | 0,69 | 0,26 | 0,00 | 0,03 | 0,00 | 0,00 |
| f_m | ACC | | | | | | | | | | | | |
| | POW | | | | | | | | | | | | |
| | FPR | 1,00 | 0,41 | 1,00 | 0,08 | 1,00 | 0,97 | 1,00 | 0,39 | 1,00 | 0,04 | 1,00 | 0,99 |
| FS+F_m | ACC | 0,17 | 0,18 | 0,78 | 0,77 | 1,00 | 1,00 | | | | | | |
| | POW | 0,16 | 0,08 | 0,25 | 0,38 | 0,01 | 0,88 | | | | | | |
| | FPR | 0,72 | 0,28 | 0,01 | 0,03 | 0,00 | 0,00 | 1,00 | 0,34 | 1,00 | 0,07 | 1,00 | 0,99 |
| FS+f_m | ACC | 0,05 | 0,05 | 0,23 | 0,24 | 0,87 | 0,88 | | | | | | |
| | POW | 0,00 | 0,00 | 0,20 | 0,00 | 0,87 | 0,86 | | | | | | |
| | FPR | 0,95 | 0,33 | 0,77 | 0,05 | 0,13 | 0,12 | 1,00 | 0,34 | 1,00 | 0,06 | 1,00 | 0,96 |

Results of the simulations are reported in Table 2. Quasi-perfect assignment can be obtained for most scenarios in the conditions of our simulations (coverages of 1.0 and 0.2 for the parents and the offspring, respectively, and unlinked markers) if at least 5000 SNPs are used. Simulations also show that the effect of potential sequencing errors (assuming an error rate of 0.01) becomes very small when the number of markers increases. As expected, increasing the number of markers and/or the coverage increases the available information, and consequently assignment accuracy (Table 2, Figure 2). Note that, unlike SNP panels, the number and identity of informative markers (i.e. markers with reads) will vary slightly from individual to individual in SWGS.

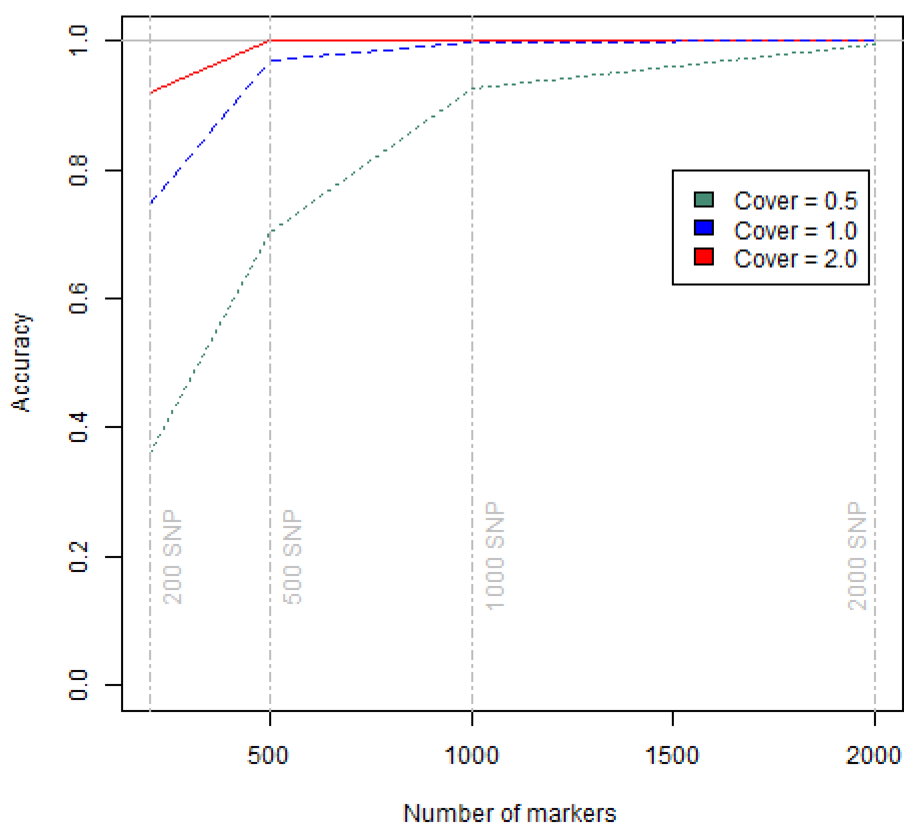


Figure 2: Accuracy as a function of the number of markers and the coverage. Accuracy is defined as the proportion of correctly assigned parents in the simulations. In these simulations, the coverage was similar for the potential parents (30 males and 30 females) and the offspring. Sequencing errors probability (as defined in the text) has been set to 0.01, and used markers were independent.

Table 3 shows the results of the simulations where we have introduced linkage and linkage disequilibrium between the selected markers. D' values, measuring markers information redundancy when we included linkage disequilibrium (Lewontin, 1964), were 0.149 and 0.483 for unlinked and linked markers, respectively. Linkage between markers did not significantly affect the results, while linkage disequilibrium reduced the power and increased the false positive rates.

Table 3: Results of the simulations showing the effect of linkage and linkage disequilibrium on the accuracy (ACC), the power (POW) and the false positive rates (FPR). Detailed definitions of ACC, POW and FPR are provided in the Material and Methods section. Recombination fraction between linked markers was 0.01, and we used a 1% error rate for alleles calls. Coverages were set to 1.0 for the parents and 0.5 for the offspring. We measured ACC, POW and FPR on 100 offspring (200 parents, chosen among 30 potential fathers and 30 potential mothers). Linkage disequilibrium was generated using 20 generations of random drift in a population of size 100, and measured using D'. D' values were computed between successive markers and averaged. We used 1000 markers in these simulations.

| Linkage | LD | D' | ACC | POW | FPR |
|---|---|---|---|---|---|
| Yes | Yes | 0.483 | 0.885 | 0.670 | 0.055 |
| No | Yes | 0.149 | 0.885 | 0.665 | 0.040 |
| Yes | No | - | 0.995 | 0.725 | 0.000 |
| No | No | - | 0.980 | 0.750 | 0.000 |

### 3.2.2 Bias due to differing coverages among the parents

Table 4 shows the results of the simulations using various coverages among the parents. In all scenarios using the highest coverages (5 and 1X), both *Shallowped* and *AlphaAssign* provide accurate results, although the power is (sometimes much) higher for *Shallowped*, despite using a more permissive definition of the assignments in *AlphaAssign*, as described earlier. We observed no false positive results in these simulations.

For the low coverages (i.e. 1X and 0.2X) scenarios, the accuracy and the power decrease for the 2 methods due to the lower information content of the genomic data. Although reduced, the power is still much higher for *Shallowped* than for *AlphaAssign*. We observed no false positives for *AlphaAssign* using the assignment criterion defined earlier. In *Shallowped*, we obtained false positive rates compatible with the used type 1 error rate ($\alpha = 5\%$) except when the offspring and both parents had low coverages. Among

these false positives, we observed no trend to select preferentially parents with higher or lower coverages. This suggests that the potential bias towards individuals with more available information is at most quite limited.

Table 4: Results of the parent assignments when the coverages vary among the parents. LL, LH and HH correspond to the coverages of the simulated parents of the tested offspring (although in all simulations, half of the tested parents were sequenced using L coverage and the other half, using H coverage). Results are proportions obtained after testing 100 offspring in each scenario. SH corresponds to the results obtained using *Shallowped*, and AA to those obtained using *AlphaAssign*. Accuracy, power and FPR (False Positive Rate) are defined in the text, and significance threshold was $\alpha = 5\%$.

| **"High" coverages**: L = 1x, H = 5x | | | | | | | |
|---|---|---|---|---|---|---|---|
| Parents | Offspring | Accuracy | | Power | | FPR | |
| | | SH | AA | SH | AA | SH | AA |
| LL | L | 1.000 | 1.000 | 0.985 | 0.040 | 0.000 | 0.000 |
| LH | L | 1.000 | 1.000 | 0.985 | 0.505 | 0.000 | 0.000 |
| HH | L | 1.000 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| LL | H | 1.000 | 1.000 | 1.000 | 0.475 | 0.000 | 0.000 |
| LH | H | 1.000 | 1.000 | 0.985 | 0.730 | 0.000 | 0.000 |
| HH | H | 1.000 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| **"Low" coverages**: L = 0.2x, H = 1x | | | | | | | |
| Parents | Offspring | Accuracy | | Power | | FPR | |
| | | SH | AA | SH | AA | SH | AA |
| LL | L | 0.570 | 0.575 | 0.240 | 0.000 | 0.090 | 0.000 |
| LH | L | 0.705 | 0.780 | 0.460 | 0.100 | 0.045 | 0.000 |
| HH | L | 0.880 | 0.895 | 0.730 | 0.175 | 0.025 | 0.000 |
| LL | H | 0.970 | 0.970 | 0.605 | 0.000 | 0.005 | 0.000 |
| LH | H | 0.990 | 0.995 | 0.820 | 0.030 | 0.000 | 0.000 |
| HH | H | 0.995 | 1.000 | 0.995 | 0.035 | 0.000 | 0.000 |

## 3.3 Empirical data

### 3.3.1 Coverage level of parents and offspring

We could have obtained information on polymorphisms using SWGS data only, since 60 parents with an average coverage of 1x lead to an average coverage of 60x, which is sufficient to identify variants such as SNPs. We used an alternative option: a preliminary sequence (obtained in our lab from another experiment involving deep sequencing of one striped catfish) was available, and we have used it for read alignment and variant detection. After initially obtaining more than 5,900,000 genomic variants and performing marker selection to obtain an informative set of markers, we ended up with between 20,822 and 20,959 SNPs for parentage assignment of each offspring. In general, coverage was not markedly different between the 60 parents and the 500 offspring: coverage ranged from ~ 0.32 to ~ 2.83 in the fathers (average: 1.46), from ~ 0.48 to ~ 2.97 in the mothers (average: 1.65) and from ~ 0.006 to ~ 3.69 (average: 1.99) in the offspring.

### 3.3.2 Parental assignments

We have computed the likelihoods and determined the most likely parental couple for each offspring. Figure 3 shows typical results for one offspring, plotting the 900 log-likelihoods corresponding to the 30*30 possible parental couples. Using *AlphaAssign*, we obtained no significant assignment using the default criterion, while all parent assignments were significant using the hypothesis with the closest score to the parent's score. After running this analysis, 875 of the 1000 parents were identical and significant ($p < 0.05$) using both *Shallowped* and *AlphaAssign* programs. Forty-one additional parents were identical in both programs, although not significantly allocated by shallowped. Eighty parents differed between the two programs, but were not significantly assigned by shallowped. The last four parents, although significant, differed for the two programs, indicating false positive results for one or the other program. The offspring with the lowest coverage (0.0069) was one of these 4 individuals with one parent incorrectly assigned by one (or both) program(s). All 30 males had at least one offspring among all tested individuals, while this was only the case for 20 females. The set of missing females included one female that we included as a negative control (her eggs did not hatch in our field experiment).

Using the reduced set of reads to obtain equal coverages for all parents, we obtained coverages ranging between 0.556 and 0.562 for all offspring (except for 3 offspring, with very low coverages). The results from the comparison between the results obtained with the full and the reduced datasets show that 950 of the 1000 parents were identical for the two runs. Both parents differed for 7 individuals and one parent changed for 36 individuals. Twenty-seven of the fifty changed parents were not significantly assigned before the reduction of the dataset.

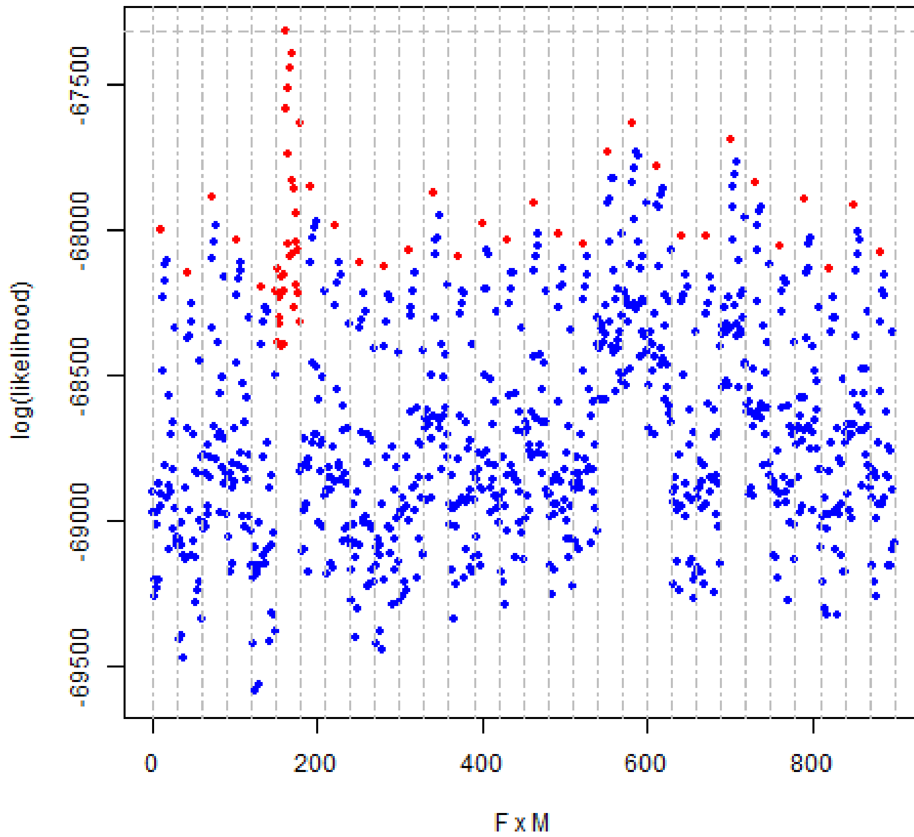


Figure 3: Plot of the log-likelihoods for an offspring and a set of 30 potential fathers and 30 potential mothers. Each dot corresponds to one triplet. The horizontal dashed line corresponds to the best likelihood, and the corresponding parents are the putative parents of the tested offspring. The red dots correspond to likelihoods computed with the putative father and/or the putative mother. The dash-dotted line splits the likelihoods into sets of likelihoods coming from the same father.

## 4. Discussion

In this study, we have shown that shallow sequencing is an affordable technique allowing accurate and powerful assignment of the relatives of tested individuals. The devised method, although similar to other methods such as *AlphaAssign* (Whalen, 2019), allows testing several competing hypotheses, and is able to discriminate parents from close relatives if sufficient genomic information is provided. In addition, we have found that the inferences we made were rather insensitive to confounding factors such as

sequencing errors or differential coverages among the tested individuals. We next discuss several aspects of the results presented in the previous sections.

### 4.1 Parentage assignment accuracy with SWGS data

One of the main goals of the simulations was to quantify the amount of SWGS data required to accurately perform parentage assignment. We found that the total amount of data required is relatively low. For example, only 200 SNPs are required to accurately assign parents when using 4X coverage SWGS data. Our results were in agreement with previous estimates for SNP array data (Anderson & Garza, 2006; Liu, Palti, Gao, & Rexroad, 2016; Strucken et al., 2016), where between 60 and 700 SNPs were required to accurately assign parentage. The difference in the exact number of markers required (200 compared to 60-700) is due to the difference of information between called genotypes from array data and read information from SWGS data on one hand, and also potentially to the structure of the underlying genetic data (*i.e*, number of chromosomes, MAF of the markers) (Whalen et al., 2019) on the other hand. In addition, we found that low-coverage SWGS (for example, 0.5X or even less) data can also be used to assign parents, although necessitating a larger number of SNPs (5000 independent markers were sufficient with 0.5X coverages). The required number of markers has to be increased to counterbalance the lower information content at an individual locus from low-coverage SWGS data, which is generally not a problem with such data where the number of polymorphisms is potentially huge. These findings are particularly important given the emerging availability and affordability of SWGS as an alternative to SNP array data for assigning parents, especially for species where SNPs panels are not available yet.

It has been recognized for long that genotyping problems have a large impact on parentage analysis (Douglas, Skol & Boehnke, 2002; Sobel, Papp & Lange, 2002), and, if ignored, may cause extremely erroneous inference. In various situations, due to many factors such as genotyping errors, mutations, genotyping technology, missing data (Bonin et al., 2004; Pompanon, Bonin, Bellemain & Taberlet, 2005) and sequencing errors, the quality of some genotypes may be poor. According to Trọng et al (2013), detecting an error can be done when comparing a given set of genotypes against a reference that was obtained from high-quality genotype or from multiple repeats. Incorrect genotypes calls can cause problems for parentage assignment even for error rates lower than 1%, as shown in a simulation study by Marshall et al (1998), or more recently in studies using SNP (Liu, 2007; Bresadola, 2020). In a study of Hoffman and Amos (2005), a genotype error rate of 1% per locus resulted in more than 20% of incorrect paternity exclusion on Antarctic fur seals (*Arctocephalus gazelle*). The effects of genotyping errors have increasingly raised concerns and several efforts have been made to accommodate errors in parental analysis (Jones et al., 2010; Kalinowski, Taper & Marshall, 2007; Pompanon et al., 2005). These studies just focused on high

quality data such as SNPs genotyped from microarrays or from high-coverage NGS technologies. In our case, using low-coverage sequencing data, the probability of miscalling is higher: sequencing errors are still possible, low coverage does not allow an easy detection of sequencing errors, and lack of sufficient information at each locus might easily lead to call a heterozygous locus as homozygous (Wang, 2019). However, likelihood approaches such as the one used in the present study have been considered as the best model for reducing the negative effects from genotyping problems in parentage inference, especially with large number of markers (Tortereau et al., 2017; Wang, 2019). Moreover, with the development of NGS techniques (we used sequencing-by-synthesis in our work), the real average error rate is reported to be only 0.1% per nucleotide, most of which are single nucleotide substitutions (Fox, Reid-Bayliss, Emond & Loeb, 2014). In the current study, we assumed an error rate of 1%, i.e. ten times higher than the average expected one, but this did not greatly affect the accuracy of parental assignment (0.976 instead of 0.985) in situations where large numbers of SNPs were used.

Although the quality of the reads is important to decrease the probability of sequencing errors, MAF is another factor affecting the parental assignment performances. Although not mentioned in our results, our simulations confirm that high MAF should be preferred, similar to other studies. Based on data simulation, Anderson & Garza (2006) and Fisher et al (2009) predicted that SNPs with MAF from 0.4 to 0.5 provided the most power for parentage inference. Liu et al. (2016) selected SNPs with MAF greater than 0.45 for parentage assignment in rainbow trout. As MAF increases, more opposing homozygotes become available for parentage assignment. With low MAF, the probability of observing opposing homozygotes declines due to the overrepresentation of one homozygous genotype (Strucken et al., 2016). Consequently, filtering markers on MAF could allow reducing the number of markers if needed.

Although coverage, error rates and marker MAF are important factors, our results nevertheless show that the effects of these factors can be largely balanced by the large number of markers. This is a clear benefit from SWGS data, where potentially millions of variants could become available for the subsequent analyses. In our work, 4,673,445 raw SNPs (out of 5,900,000 genomic variants) were generated in the first stage of analysis. Although quality controls and filters can largely reduce these numbers, the numbers of markers to be used in the analyses could be significantly higher than for example in the studies of (Dodds et al., 2015; Dodds et al., 2019) where 78,042 and 30,923 SNPs, respectively, were produced from GBS data for parentage assignment.

### 4.2 Effect of the coverage on parentage assignment accuracy of striped catfish

As mentioned above, the coverage is one of the main factors influencing the accuracy of parentage analysis, and the number of markers needed to obtain the desired accuracy for a given coverage needs to be

estimated. Our simulations, with an expected 0.5X coverage given the number of tested individuals, the capacity of the sequencing equipment and the available money, have shown that 5,000 SNPs were required to obtain 100% of accuracy. Increasing the coverage to 1X allows decreasing the required number of SNPs to approximately 2,000. Note that the level of required coverage for obtaining 100% of accuracy in our study is lower than that of Whalen et al (2019) on GBS data with 5,000 markers. In Whalen et al (2019), the required coverage was 2X instead of only 0.5X in our study. The difference might originate from different hypotheses in the design of the simulations. For example, we have considered independent markers, which might be overoptimistic when the number of markers becomes large. In real situations, the needed number of markers will also depend on other factors, such as the size of the genome and the history of the population.

Based on the simulation result, the coverage in our empirical data experiment should be sufficient for accurate parentage assignment, although the SNP were not guaranteed to be independent (they were chosen on different contigs of the current sequence, but with no information on the relative positions of these contigs). Our coverage was significantly lower than that of Dodds et al (2019) with 3.6 and of Dodds et al (2015) with 7.9, where both studies used GBS data for parentage assignment on deer and Atlantic Salmon, respectively. In the case of Dodds et al (2015), the mean depths ranged from 0.38 to 38.10 for each individual, while the range in our study was from 0.003 to 2.97 (actually, only 3 offspring had a coverage below 0.3). The range of average depths in our study could be due to stochastic sampling in the sequencing process and/or to slight differences in DNA concentrations in the libraries preparation. However, no parent appeared to be an outlier, and all parents were thus retained for parentage assignment. On the other hand, as expected the offspring with a very low coverage (0.003) turned out to be difficult to assign significantly to a couple of parents, as could be expected.

Sequencing individuals to low depth will generate many missing genotypes. For example, a coverage of 1X should generate about 37% of markers with no read at all. When the same amount of genetic information (*e.g* the same number of SNPs and the same sequencing coverage) is available for all parents, the parentage inference with a likelihood approach works well, but this method may break down when individuals are genotyped at different number of markers or at different coverage level (Whalen et al., 2019). The common approaches for dealing with missing genotype problems are to filter genotypes based on read depth (Thrasher, Butcher, Campagna, Webster & Lovette, 2018) (which may drastically decrease the amount of genomic information available) or replace them with estimated values. In our study, the use of the first strategy led to a drastic reduction in the number of reads but did not dramatically change the picture in terms of the identities of the parents of the selected offspring. The simulations involving parents

with different coverages led to the same conclusions. This seems a good indication that the paternities obtained using this approach are rather robust with respect to a possible heterogeneity in the parents' available information.

Alternative methods could be used to equalize the amount of information available for each potential parent, such as, for example, develop an algorithm to infer an optimal set of markers with respect to the information available among the parents, or using linkage disequilibrium to infer missing alleles.

Altogether, our results indicate that, in studies similar to ours, problems like low sequencing depth and random distribution of the coverages across samples and libraries (which is one of the potential pitfalls with SWGS data) can be circumvented by taking a larger number of markers. In the study of Whalen et al (2019), the GBS data with low-coverage may have caused false positive parentage assignment due to the fact that they were not able to obtain a sufficient number of loci with precisely inferred genotypes to find opposing homozygous loci.

**4.3 Statistical significance**

Our main achievement is the definition of a rationale that allows testing statistically various hypotheses about potential relationships in the data. Improvements remain possible like adding more hypotheses in the set of potential relationships. For example, we could infer the distribution of a father's genotype assuming that the best female is actually a full-sib of the mother. However, we think that this feature provides a clear method to decide on the significance of a relationship assignment, without resorting to parameters difficult to define. Other programs use confidence levels, which are tolerance thresholds for false positive assignment. For example, the likelihood-based method (*Cervus* program) calculates separately different levels of confidence: relaxed confidence as 80%, strict confidence as 95% (Marshall et al., 1998) and 99% confidence (Kalinowski et al., 2007). In the study of Trọng et al (2013), the *Cervus* program was used for parentage assignment on Nile tilapia (*Oreochromis niloticus*). Using 12 microsatellites and 122 SNPs, the mean assignment rate was 18% and 39% at 95% confidence, respectively. *AlphaAssign* (Whalen, 2018) also uses a threshold corresponding to a 99% posterior probability of the assignments. As visible from our results, even relaxing this constraint in our simulations, this made significance difficult to reach in many cases, leading to a reduced power compared to the results obtained using *Shallowped*.

## 5. Conclusions

This paper presents a new parentage assignment method using genomic information from SWGS based on a newly developed algorithm. Using directly low coverage sequence data allows reducing the sequencing costs considerably. Additionally, the costs and labors of developing marker panels can be avoided, which is another advantage. Based on the analysis from simulated and empirical data, the method has demonstrated its effectiveness for parentage assignment, with assignment performances that can be improved by increasing the number of markers. These results suggest that SWGS data can be used for accurate parentage assignment using the appropriate algorithm and we believe that it can be a useful tool in parentage assignment not only for aquaculture but also for a wide range of animals.

### Data accessibility

The *Shallowped* program is a Fortran 95 program implementing the likelihood approach and the significance testing procedure. The script *mksimul.pl* is a perl script used to generate simulation data in a convenient format for use in *Shallowped*. Both the program and the script are freely available at the following link: https://doi.org/10.5281/zenodo.6033536

### Acknowledgements

This work was part of the PANGAGEN project, supported by the Académie de Recherche et d'Enseignment Supérieur - ARES (Begium) and funded by the Belgian Development Cooperation - (DRP/TPS 2017). We would like to thank staffs of Genomics Platform, GIGA, University of Liege for sequencing all DNA samples. We are also grateful to PANGAGEN team at Can Tho University and numerous volunteer students for their help with fish sampling.

# Chapter 4

# Selective breeding of saline-tolerant striped catfish (*Pangasianodon hypophthalmus*) for sustainable catfish farming in climate vulnerable Mekong Delta, Vietnam

*** *Published in Aquaculture Report: https://doi.org/10.1016/j.aqrep.2022.101263*

Dao Minh Hai[1,2], Duong Thuy Yen[2], Pham Thanh Liem[2], Bui Minh Tam[2], Vo Nam Son[2], Do Thi Thanh Huong[2], Bui Thi Bich Hang[2], Nguyen Thi Ngoc Tran[2], Dang Quang Hieu[3], Kestemont Patrick[3], Nguyen Thanh Phuong[2], Farnir Frédéric [1].

*1. FARAH/Sustainable Animal Production, Faculty of Veterinary Medicine, University of Liege (B43), B-4000-Liege, Belgium*

*2. College of Aquaculture and Fisheries, Can Tho University, Viet Nam.*

*3. Research Unit in Environmental and Evolutionary Biology, Institute of Life, Earth & Environnment, University of Namur, rue de Bruxelles 61, B-5000 Namur, Belgium.*

# Chapter 4: Selective breeding of saline-tolerant striped catfish (*Pangasianodon hypophthalmus*) for sustainable catfish farming in climate vulnerable Mekong Delta, Vietnam

## Preamble

In Chapter 3, we succeeded to develop a new parentage assignment method using genomic information from SWGS. In this study, we applied it to identify pedigree information in the selection program. To prepare for a genetic selection program targeting fish displaying improved adaptation to salinity, a base population made up of 900 half and full-sib families was produced from a full factorial cross of 30 females and 30 males from 3 different provinces in the Mekong Delta to serve as genetic diversity resource for selection. The selection program was implemented through three sequential upward mass selections under the targeted salinity of 10 ppt. After selection, several tests were carried out on the next generation of fish to evaluate the salinity tolerance of striped catfish and the effectiveness of selection after one generation in saline conditions.

## Abstract

Striped catfish (*Pangasianodon hypophthalmus*), a freshwater species cultured mainly in the Mekong Delta region in Southern Vietnam, is facing a significant challenge due to salinity intrusion as a result of climatic changes. Given these evolving environmental conditions, selecting new strains with a higher salinity tolerance could make production of striped catfish economically feasible in brackish environments. In this study, we carried out a selection program aimed at developing a striped catfish strain able to survive and grow fast in a saline environment. To implement the selection program, we first collected males and females from different provinces in the Mekong Delta. We next performed a factorial cross of these breeders to produce half- and full-sib families. When fish reached fry stage (47-day post hatching (dph)), we put them in a saline environment (10 ppt) and subsequently kept 50% of the fastest-growing fish after 143 dph. We repeated this mass selection procedure after 237 dph and 340 dph. We maintained in parallel a randomly selected group in saline conditions and a group of fish reared in freshwater to serve as controls. After crossing the selected individuals, we performed several tests on the next generation of fish to evaluate the effectiveness of selection after one generation in saline conditions. Average direct responses to selection were 18.0% for growth and 11.4% for survival rate after one generation of selection. We estimated a moderate realized heritability (0.29) for body weight. The genetic gains obtained in our study for body weight and survival rate after one generation of selection under saline conditions suggest that selection can be effective to improve ability of striped catfish to cope with saline stress. We conclude that our selection program has succeeded in developing a productive strain of striped catfish with better tolerance to salinity.

Keywords: striped catfish, heritability, genetic adaptation, selection program, climatic change

## 1. Introduction

Striped catfish (*Pangasianodon hypophthalmus*) has become one of the most important farmed freshwater indigenous species in the aquaculture sector in the Mekong Delta since artificial breeding was developed successfully in the mid-1990s (Nguyen et al., 2013). By the year 2018, catfish production reached nearly 1.42 million tons, (for a total production area of approximately 5,400 ha) and the export value was 2.26 billion USD (contributing about 1% GDP of Vietnam), with exports to more than 140 countries and territories (VASEP, 2019). De Silva and Nguyen (2011) estimated that this sector employs more than 180,000 Vietnamese, primarily women, in the processing sector.

However, the Mekong Delta is projected to be heavily affected by salinity intrusion as a result of climate change impacts (IPCC, 2007). This region is one of three extensive low-lying regions in Vietnam, with a maximum elevation of less than 4.0 m above mean sea level, and current models predict a 1 m rise in sea level in this century (IPCC, 2007; Thi et al., 2015). If these predictions are realized, approximately 1,000 $km^2$ of cultivated land and farming area in Vietnam will become salt marshland. In total, seawater is likely to inundate 15,000 to 20,000 $km^2$ of the Mekong Delta, with a loss of 76% of arable land (De Silva and Soto, 2009).

Striped catfish culture is largely dependent on the availability of an abundant supply of fresh water from the Mekong river (Nguyen and Dang, 2010). Therefore, the increasing salinization of freshwater areas will negatively impact striped catfish culture (Sebesvari et al., 2011). The Mekong Delta has already faced serious drought and saline water intrusion into inland areas (up to 55-100 km from the coastline) in 2016 and 2020, causing damage to aquaculture, rice and fruit production and important economic losses (Hai et al., 2020; VAWR, 2017). These major environmental changes have a significant direct impact the growth rate of catfish (Hossain et al., 2021). Selectively breeding a salinity-tolerant strain of striped catfish is a promising approach to maintaining sustainable catfish aquaculture in the region in the current changing environmental context.

Selective breeding has been successful in improving the salinity tolerance of tilapias (euryhaline fish) (Jaspe and Caipang, 2011), and a few pioneer studies suggest that substantial additive genetic variance exists for growth rate and survival in saline environments (Tayamen et al., 2010; Tran et al., 2008). Cnaani and Hulata (2011) argued that the salinity tolerance of the fish is the overall fitness or productivity of fish in a saline environment resulting from a combination of quantitative traits including metabolism, growth, osmoregulation, immunocompetence and fecundity, each of which is influenced by multiple genes. Several

selection programs on tilapias have all also evaluated salinity tolerance based on growth and survival data (Cnaani and Hulata, 2011; Tayamen et al., 2010).

Only one published study has described a selection program for striped catfish, and it focused growth in freshwater (Sang et al., 2012, 2010; Vu et al., 2019). To our knowledge, no such selection program has been carried out in saline conditions. On the other hand, several studies have investigated the effects of increased salinity on growth performance of striped catfish ((Hossain et al., 2021; Nguyen et al., 2014).

Although there is no published data demonstrating that salinity tolerance could be increased by selection on striped catfish, differential growth and survival within groups facing the same salinity conditions observed in previous studies (Huong and Quyen, 2012; Nguyen et al., 2014) might suggest the presence of genetic components involved in the salinity tolerance traits. To test this genetic hypothesis, we have conducted a selection experiment in saline conditions. Therefore, the main objectives of the present study were to assess the efficiency of selection on the salinity tolerance of striped catfish in terms of growth and survival after one generation of selection in 10 ppt brackish water. In addition, as a first step towards a better understanding of the involved genetic mechanisms, we have investigated physiological parameters differences between the selected and control strains.

## 2. Materials and Methods

The fieldwork of the study was conducted in the College of Aquaculture and Fisheries (CAF), Can Tho University and three striped catfish hatcheries in Viet Nam from July/2017 to May/2021. The analysis of the genetic data was performed at Liege University, Belgium.

### 2.1 The base population (G0) and production of the first generation (G1)

We obtained striped catfish broodstock (base population G0) from three different hatcheries located in An Giang, Vinh Long, and Can Tho provinces (freshwater areas), with 10 unrelated males and 10 unrelated females from each hatchery (*i.e.* 60 broodstock total), weighing between 5 and 7 kg on average. Principally, the base population individuals had to be healthy without visible injury or deformities. We collected a piece of fin tissue (~1 $cm^2$) from each fish and preserved these biopsies in 95% ethanol to extract DNA for parentage assignment.

We produced the first generation (G1) using a full factorial cross of the G0 broodstock. Each of 30 females (~3,000 eggs of each female) being individually crossed with each of 30 males in 1:1 crosses to create 900 full-sib families. Spawning was induced by injection with human chorionic gonadotrophin with

a total dose of 5,500 UI/kg weight of female and 1,000 UI/kg for male. A dry fertilization process (Tam et al., 2010) was used, where eggs and milt were mixed gently. The fertilization solution (3 g urea and 4 g salt in 1 L of water) was added to the mixture of eggs and milt to trigger fertilization after 4 min. The fertilized eggs were then transferred into family-specific plastic boxes for incubation. The fertilized eggs started to hatch 24 h after fertilization. Fertilization rates varied from 83-90% and the hatching rates were 61-73% (except for one female whose eggs did not hatch). From each family, we selected 2,000 good quality larvae (no deformities, uniform size, swimming actively and responding to external stimuli quickly) for nursing. Larvae from all mixed families were transferred to two rearing earthen ponds within 15 h after hatching (leading to 1,740,000 larvae in total) under freshwater condition. After 47 days post-hatching (dph), 21,550 randomly chosen fry were transferred from the nursing ponds to CAF, Can Tho University. Twenty-one thousand fry were put into a recirculating aquaculture system (RAS - 700 $m^3$ water volume) to start the selection process under saline condition. Simultaneously, the remaining fish (550 fry) were transferred to another RAS system (50 $m^3$ water volume) to be reared in freshwater until maturation. This group (referred to as “freshwater group”) represents normal broodstock sources currently used in Mekong Delta (Figure 1).

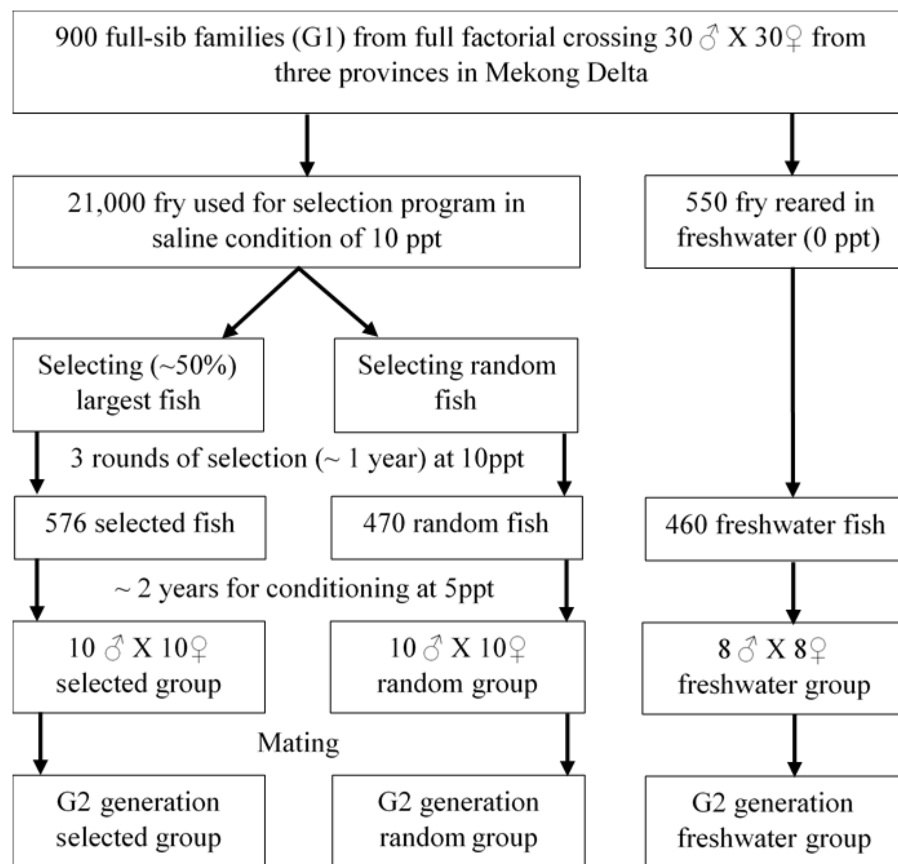


Figure 1: Schematic representation of mating design for obtaining the experimental fish

### 2.2 Selection rounds

In RAS (700 $m^3$ water volume), we progressively (1 ppt/day) raised the salinity to 10 ppt and maintained at this level during the whole selection experiment. Fish dying during the acclimation or later were removed from the protocol, meaning that survival was the first selected trait. When fish reached 148 dph, they were randomly divided into two groups of equal sizes. The fish in each group were distributed equally into three hapas (12 x 4 x 4 m). At this time, a first selection round was carried out. In the first group (referred to as the "selected group"), the fish were individually weighed and approximately 50% of the heaviest fish were kept for the next rounds of selection. In parallel, the second group (referred to as "random group") underwent a random selection of the fish until reaching the same biomass as in the selected group. These fish served as a control population for the selected group (to see the effect of selecting heaviest fish in saline conditions on the progeny) and for the freshwater group (to see the effect of the saline conditions on various traits). The fish not selected in either group were removed from the RAS system. We monitored all fish from the selected and random groups to keep them in the same conditions (feed, biomass).

Two additional rounds of selection took place 3 months later (237 dph) and 6 months later (340 dph) along the same lines as for the first round (Figure 1). The growth rate of striped catfish is fast during the first year of age, and then slows down. In practice, grow-out farmers often harvest striped catfish when they are around one-year old and reach a weight of around 1 kg (De Silva and Nguyen, 2011; Nguyen and Dang, 2010; Phan et al., 2009). Since growth profiles might vary among fish during the first year, and because we wanted to maintain biomass conditions compatible with the size of our RAS system, we have used this 3 stages selection based on biomass (Fig.1). Similar protocols have been used in various other studies such as on brown trout (*Salmotrutta fario*) (Bernard et al., 2004) and on rainbow trout (*Oncorhynchus mykiss*) (Le Boucher et al., 2012). After the third selection round, all remaining fish were tagged using Passive Integrated Transponders (PIT) tags. These PIT–tags were injected into the dorsal muscle of fish and corresponding DNA samples were collected.

In order to prevent any reduction in gamete quality potentially associated with the saline stress and differences in gamete quality between selected, random and freshwater breeders and growth performance of their offspring, we then decreased the salinity was to 5 ppt and maintained this level until the fish matured. After two years in these conditions, the three-year-old fish from each of the three groups had matured and could be used as parents for the next generation (G2).

### 2.3 Parentage assignment and estimation of heritability in G1 generation

We implemented and used a new algorithm allowing parentage assignment in the three groups based on very low sequencing depth (SD) from the offspring (SD between 0.2 and 0.5) and the 60 possible parents (SD =1.0). Briefly, this method is based on a likelihood approach to identify the most suitable (*i.e.* likely) set of parents for each offspring: the method computes the likelihood for each possible parental couple, keeps the most likely one and checks whether this best couple is significantly more likely than any other possible couple. Significant couples are considered as the putative parents for this offspring. The parentage information was used to identify couples of unrelated individuals (to avoid inbreeding) and to be mate them to produce the next generation (G2) (the detail of this method was presented in Chapter 3).

**2.4 Production of the G2 generation**

We wanted to assess the effects of the selection in the previous generation on the performances of the offspring in the next generation on one hand, and of the rearing in saline conditions on the other hand. To that end, we produced a G2 generation using random couples from each of the three G1 groups (*i.e.* from the selected group (10♂ X 10♀), the random group (10♂ X 10♀), and the freshwater group (8♂ X 8♀) (Fig. 1)). We pooled the same number of fertilized eggs per family within each group and incubated them together. We subsequently reared the three groups in identical freshwater environments under optimal conditions for the development of striped catfish. At each developmental stage, we subsampled a set of individuals from the three groups to challenge them with various salinity levels. Larvae hatched after 29-34 h and were transferred to three 2 $m^3$ outdoor composite tanks for nursing. The larvae from each group were nursed separately in composite tanks with the same density of 5 individuals/L. The larvae were fed Rotifer, *Artemia* nauplii, and Moina in the first 5 days. After that period, a high protein flake diet (40% protein) was used in combination with *Moina* until 4-5 weeks of age. Subsequently, fries (0.6-0.8 g of body weight) from each group were transferred to 10 $m^3$ outdoor composite tanks, where they were raised until fingerling stage with an average body weight of 20 g. Rearing experiments were divided into four distinct periods according to each developmental stage of striped catfish: embryonic stage, larva to fry stage, fry to fingerling stage and fingerling to adult stage, as described in the next sections.

**2.5 Evaluation of salinity tolerance of striped catfish after selection**

To assess the salinity tolerance of striped catfish, the G2 generation from three groups were challenged under different levels of salinity for the various developmental stages of striped catfish

**2.5.1 Experiment on the embryonic stage**

In this stage, we measured salinity tolerance using gastrulation rate (i.e. percentage of eggs reaching the gastrula stage) and hatching rate (higher rates meaning higher tolerance), deformity rate (lower rate

meaning higher tolerance), and hatching time (longer times corresponding to lower tolerance). To evaluate these traits, fertilized eggs from the three groups were transferred from freshwater to saline water and incubated in five different salinities (0, 2.5, 5, 7.5 and 10 ppt). Three hundred fertilized eggs from each group were randomly allocated to 60 plastic boxes (five salinity levels x three groups x four replicates). The salinity in all boxes was maintained using RAS systems where well-aerated water was supplied continuously for all boxes. Saline water was produced by mixing deep sea water (90 ppt seawater, from Bac Lieu province, a coastal province in Vietnam) with tap water in 1000 L composite tanks to correspond to the desired salinity treatments. Eight hours after incubation, transparent living eggs were separated from the opaque undeveloped ones and counted. The undeveloped eggs were removed immediately to avoid fungal infection. In each box, the gastrulation rate was calculated as the ratio of the number of fertilized eggs reaching the gastrula stage to the total number of eggs, multiplied by 100. The hatching rate was calculated as the ratio of the number of hatched larvae to the number of fertilized eggs, multiplied by 100. The deformity rate was calculated as the ratio of the number of deformed larvae to the total number of larvae, multiplied by 100. The hatching time was defined as the total time from the fertilization to the time when 50 % of larvae had hatched out.

We monitored water quality parameters such as dissolved oxygen (DO), pH, temperature, and salinity every five hours using a multiple-parameter water quality meter (DKK-TOA, WQC-24, Japan). During the incubating period, water temperatures were similar among RAS systems and fluctuated from 26.3 to 28.4 $^{0}$C. Values of pH were stable in the range from 8.2 to 8.6. DO was 5.1 to 5.6 mg $L^{-1}$.

**2.5.2 Experiment on larval to fry stages**

In this and the subsequence experiments, we considered survival and growth as indirect indicators of the salinity tolerance, with higher values corresponding to higher tolerance. Two hundred larvae (2 dph) from each group were transferred from freshwater and distributed randomly into each of five tanks with different salinities (0, 5, 10, 15 and 20 ppt), and we made 4 replicates of each condition. This design led to 60 composite tanks (20 L water volume), each tank containing 200 larvae. To avoid potential biases linked to the relative positions of the tanks, we allocated these positions randomly. Tanks with saline water were brought to their targeted salinity by progressively replacing water in each tank with saline such that all treatments reached their target concentration 10 days after the start of the experiment. Aeration was provided in each tank by an air-stone. We fed the larvae in each tank on rotifers, nauplii of *Artemia*, and *Moina* during the first 5 days and only nauplii of *Artemia*, and *Moina* from day 6 to day 21. The larvae were manually fed o apparent satiation four times per day (at 06:00 h, 10:00 h, 14:00 h, 18:00 h). We siphoned

off the waste accumulated in the bottoms of the tanks daily to avoid negative effects on larvae. After 21 days of experiment, all fish were harvested to evaluate the survival rate and growth performance.

The same water quality parameters as for the embryonic stage were monitored daily in this period. During the experiment, temperature varied over time from 27.2 $^0$C to 28.7 $^0$C, levels of pH ranged from 7.2 to 8.1 and DO from 4.8 to 5.5 mg $L^{-1}$, although these parameters were similar among tanks at any given time.

### 2.5.3 Experiment on fry to fingerling stage

The experimental design was similar to the one described for the larvae, with 200 fry (28 dph) from each group (initial body weights: 0.8 ± 0.1 g for the selected group, 0.7 ± 0.1 g for the random group and 0.6 ± 0.1 g for the freshwater group) distributed randomly in 60 300 L composite tanks. Individuals in each tank were fed to apparent satiation twice daily at 08:00 h and 15:00 h using a commercial pelleted feed with protein content of 35%. We maintained a constant salinity in all tanks using a RAS where well-aerated water was supplied continuously. Around 5% of the culture water in each tank was exchanged every 3 days in addition to siphoning off the accumulated wastes each of the at tank bottoms. After a 70-day experimental period, we harvested all survivors for growth rate assessment, and analyses of osmoregulatory parameters. Tank water samples were also collected and stored at time of sampling to measure osmotic pressure.

We monitored water quality daily using the same parameters as for the previous stages. During the experiment, temperatures ranged from 27.1 $^0$C to 30.5 $^0$C among RAS systems. The values of pH ranged from 7.1 to 8.0, and DO ranged from 5.0 to 6.5 mg $L^{-1}$. Additionally, total ammonia nitrogen (TAN) and $N\text{-}NO_2^-$ were measured weekly using a Multiparameter Photometer (Hana Hi 83300). The concentrations of TAN and $N\text{-}NO_2^-$ were lower than 0.5 mg $L^{-1}$ and 0.3 mg $L^{-1}$, respectively.

We estimated growth performance for each salinity treatment as weight gain (WG, g), daily weight gain (DWG, g $day^{-1}$), specific growth rate (SGR, % $day^{-1}$) and feed conversion ratio (FCR). The estimated values were based on the following standard formulae used routinely as performance indicators in aquaculture studies (Bandyopadhyay and Das Mohapatra, 2009; Fagbenro and Arowosoge, 1991):

$$SGR(\%/day) = [(\ln Wf - \ln Wi)/T] \times 100$$

DWG (g/day) = (Wf - Wi)/T, where Wf and Wi refer to the mean final and initial weights, respectively, and T is the experimental period in days.

Feed conversion ratio (FCR) = total weight of feed consumed (g)/total fish weight gain (g)

We collected blood samples from 3 fish per tank with a heparinized syringe and transferred the individual blood samples to 1.5-ml labelled tubes. Blood samples were then centrifuged for 15 min at 4500 rpm at 4 $^0$C. Next, we separated plasma, and stored it frozen at – 20 $^0$C for later analysis. Plasma osmolality was measured using Advanced Instrument Osmometer Model 3320. Na+ and $Cl^-$ ions concentrations were measured using Flame Photometer 420 and MKII Chloride Analyzer 926s, respectively.

**2.6 Estimation of realized heritability and the response to selection in G2 generation**

In this experiment, the three fish groups were stocked together in a common garden design. The environment was identical for all groups to minimize the error caused by confounding between-tank variation in separate replicates and genetic effects (Dussault and Boulding, 2018; Liu et al., 2016). The fingerlings from the three groups were distinguished using visible implant elastomer (VIE) tags with three different colors: red for the selected, yellow for the random and green for the freshwater fish. The method for tagging VIE on fish was presented in detail in GEV (2020). It was reported that VIE tags in Zebrafish (*Danio rerio*) could be retained for one year and that tagging did not interfere with long-term growth and survival (Hohn and Petrie-Hanson, 2013).

We used a RAS system with four 10 $m^3$ composite culture tanks (each tank corresponding to one replicate) in this experiment. From each group, 130 fish (100 dph, initial body weight of selected fish: 22.0 ± 1.4 g; random fish: 20.2 ± 0.7 g, freshwater fish: 19.3 ± 2.6 g) were randomly allocated to a culture tank (total of 390 fish per tank). We then brought the tanks gradually to the targeted saline condition by progressively replacing water in each tank with saline water over 10 days (1 ppt per day) until reaching the desired salinity of 10 ppt (the same level as in the selection experiment). We fed each tank to apparent satiation twice daily at 09:00 h and 16:00 h using a commercial pelleted feed with protein content of 28%. Approximately 3% of the culture water in each tank was exchanged every 2 days in addition to siphoning off the accumulated waste at the tank bottom. The entire experiment lasted for 8 months when the fish reached commercial sizes and an age (11 months) similar to the age of the fish at the end of the selection experiment. At harvest, all survivors were weighted individually for growth rate assessment. The fish from the three groups were classified according to their VIE color.

In this stage, the same water quality parameters as above were monitored daily in the RAS system. During this experiment, temperatures ranged from 27.3 $^0$C to 30.8 $^0$C, the level of pH ranged from 7.1 to 8.5 and DO ranged from 3.9 to 5.2 mg $L^{-1}$. The concentration of TAN and N-$NO_2^-$ was smaller than 1.5 mg $L^{-1}$ and 0.8 mg $L^{-1}$, respectively.

At this stage, the direct genetic response to selection of the growth in saline condition in the G2 generation was estimated as the relative weight superiority of the selected group on the random group, i.e. R = Ws - Wr, or R% =100*[(Ws –Wr)/Wr], where Ws and Wr were the mean weights at harvest of offspring from the selected and random groups, respectively, and assuming that the average Wr estimated the corresponding average in the preceding generation. The realized heritability for growth in saline condition was calculated as: $h^2 = R/S$, where S was the selection differential, computed as the average difference in trait values between the selected individuals and the overall population in the preceding generation. Coefficients of variation for body weight for each group were estimated as: (standard deviation/mean) *100.

**2.7 Statistical analysis**

Since most investigated parameters were not normally distributed, we used permutations procedures (Manly, 2007) applied to our factorial designs in order to evaluate the effects of salinities, fish groups and their interaction. Differences between factor least square means were tested using Duncan's multiple-range tests. For all tests, the level of significance was set at $P < 0.05$. Statistical analyses were conducted using SAS 9.3 (SAS, 2009)

**3. Results**

After 3 selection stages (11 months of age), the average body weight and SEM of the fish in the selected group and the random group in the G1 population were 1,380 ± 7.3 g (576 fish) and 793 ± 10.6 g (470 fish), respectively. As described above, these G1 fish were the parents of the measured G2 fish (Figure 1).

**3.1 Embryonic development of fish under saline conditions**

Increased salinity resulted in a decrease in gastrulation and hatching rates and an increase in deformity rate and hatching time. The statistical analyses showed that salinity, fish group and their interaction had significant effects on gastrula, hatching, and deformity rates ($P < 0.05$) (Figure 2). For salinity level set to 0 ppt, these parameters were similar for the three groups of fish ($P > 0.05$). The parameters started to differ significantly between the three groups above 2.5 ppt: the gastrulation and hatching rates of the selected and random groups were significantly higher and their hatching time and deformity rate were significantly lower than for the freshwater group. Although not always significantly, the values from the selected group were consistently better than those from the random group.

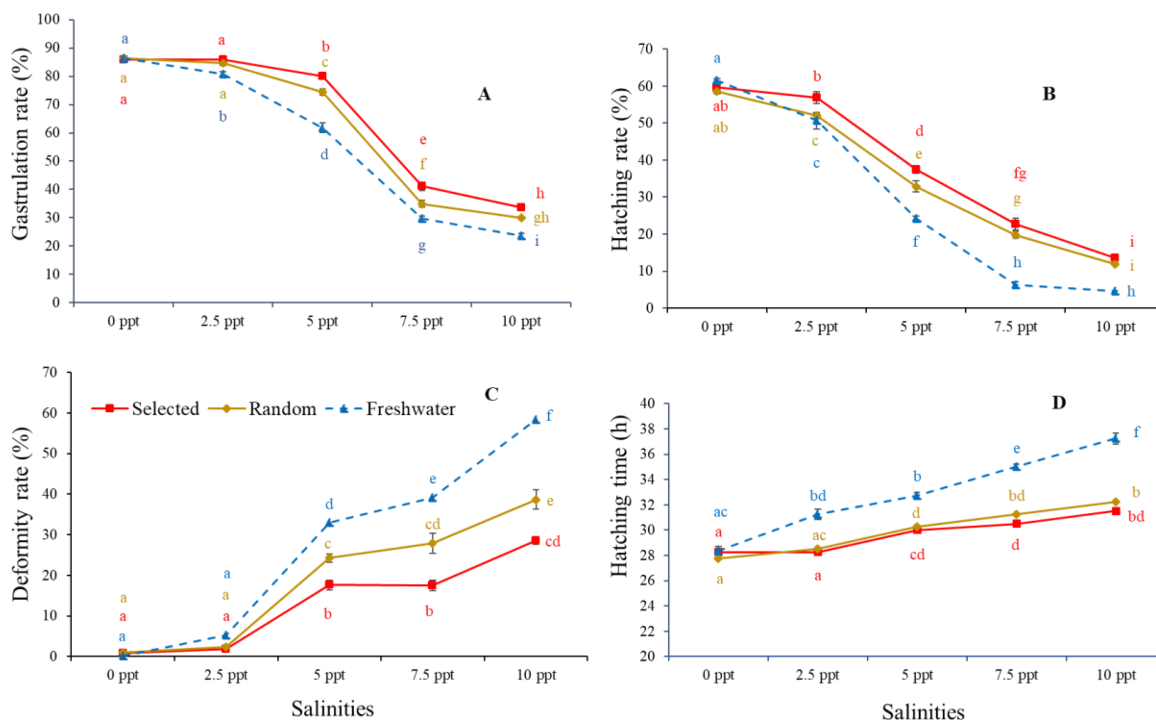


Figure 2: Effects of salinity, fish group on embryonic development of three fish groups under saline conditions; Gastrulation rate (A), Hatching rate (B), Deformity rate (C) and Hatching time (D). Values are given least square means (LSM) ± SEM. Different letters indicate significant difference (P < 0.05).

### 3.2 Growth and survival of fish under saline conditions on larvae to fry stage

Results showed that the survival rates and the final weights tended to decrease with increasing salinities. Salinity levels and fish group had significant effects (P < 0.05) on the survival rates and final weights, with no indication of significant interactions (Figure 3). No fish survived when the water salinity was 20 ppt. At 15 ppt, only few fish from the selected and random groups survived. Although the survival rates were systematically higher in the selected group, the differences were not significant at 0 ppt, but became significant for 5 and 10 ppt. Similarly, the final weights in the selected group tended to be higher than in the random and freshwater groups, with significant differences from 10 ppt (P < 0.05).

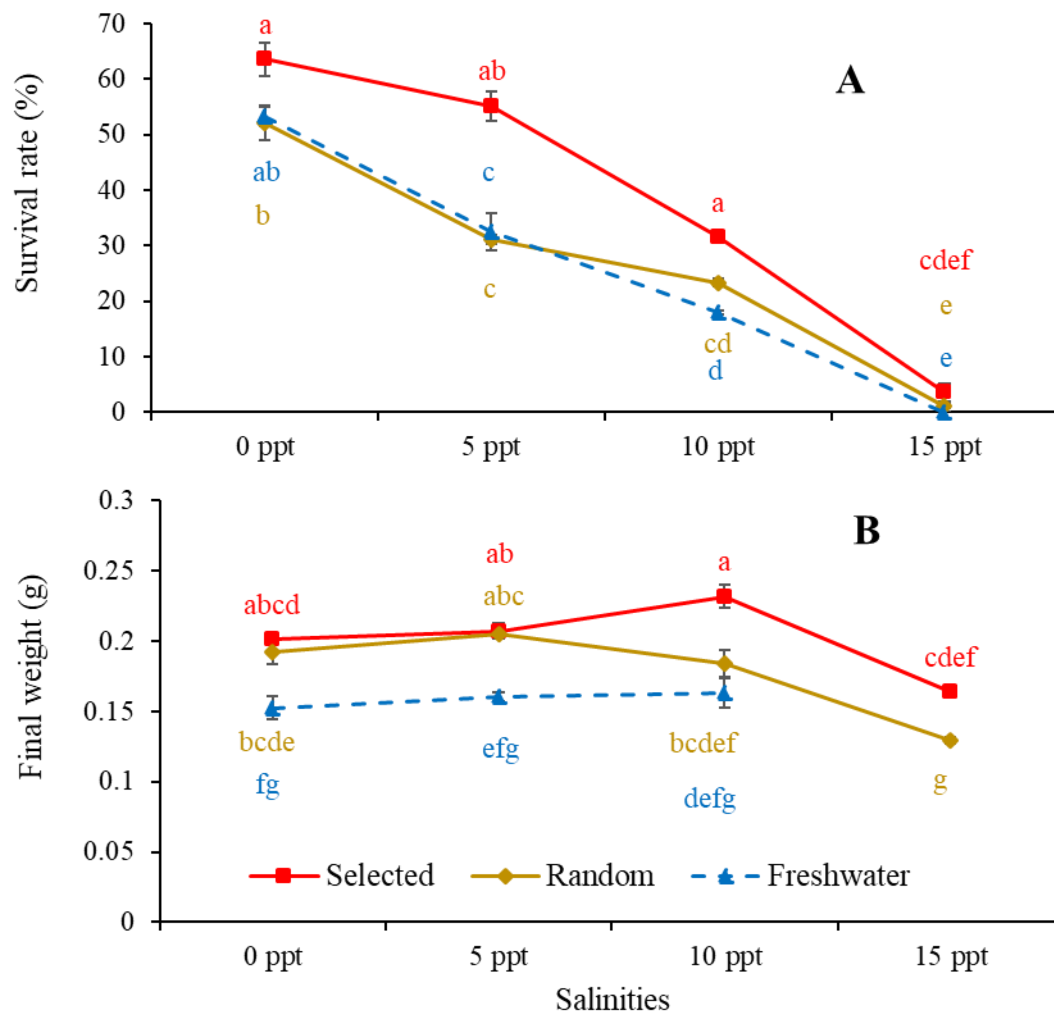


Figure 3: Growth and survival of three fish groups under saline conditions on larvae to fry stage; Survival rate (A) and Final weight (B). Values are given least-square means (LSM) ± SEM. Different letters indicate significant difference ($P < 0.05$).

**3.3 Growth performance and changes in osmoregulatory parameters of fish under saline conditions on fry to fingerling stage**

**3.3.1 Fish survival, growth rate, and FCR**

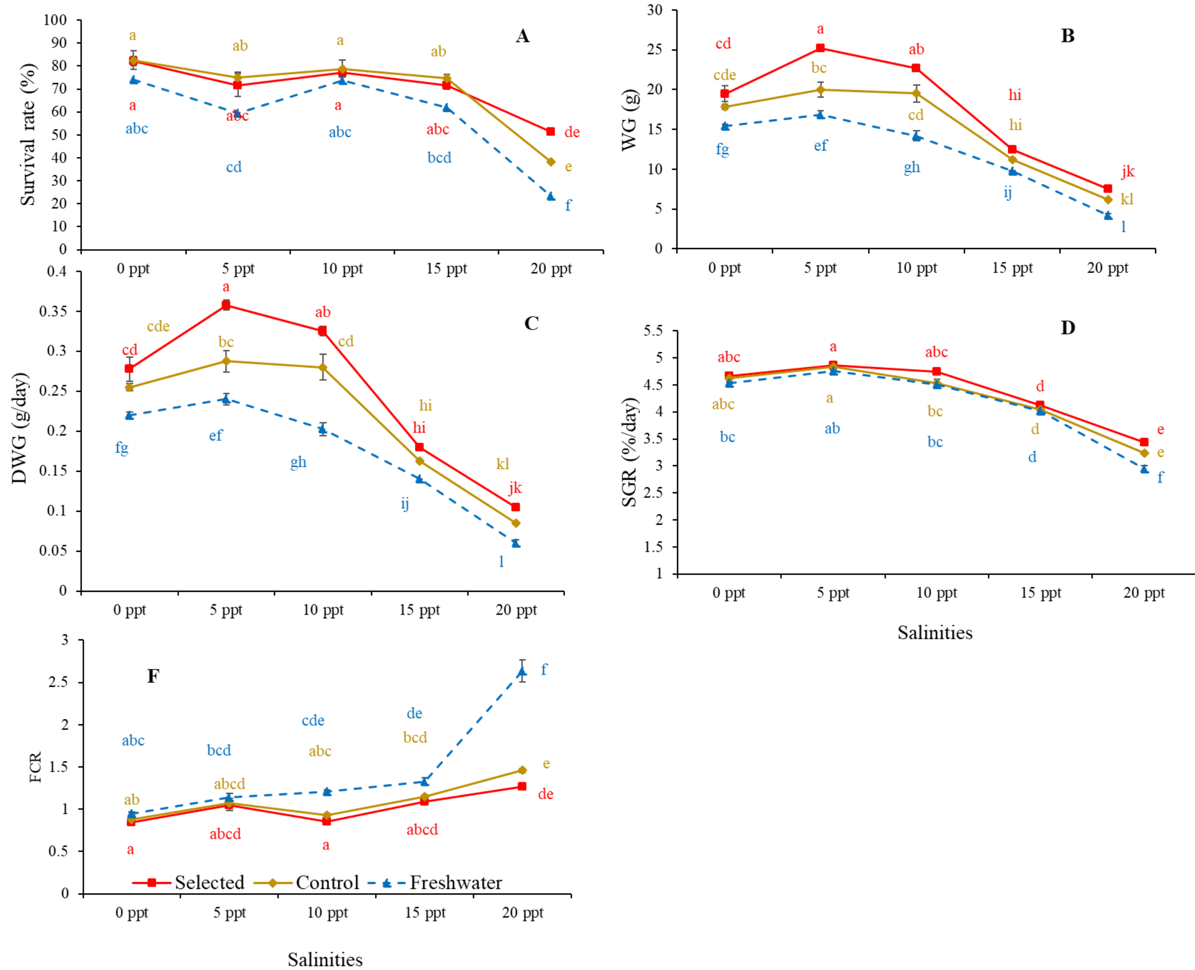


Figure 4: Growth performances of three fish groups under saline conditions on fry to fingerling stage; Survival rate (A), WG: weight gain (B), DWG: daily weight gain (C), SGR: specific growth rate (D), FCR: Feed Conversion Ratio (F). Values are given least-square means (LSM) ± SEM. Different letters indicate significant difference ($P < 0.05$).

For all tested salinity levels, the fish from the selected group grew faster, followed by those from the random, and the freshwater group (Figure 4). Fish growth rates (WG, DWG, and SGR) were highest in 5 ppt ($P < 0.05$) and did not differ between 0 and 10 ppt ($P < 0.05$). In terms of feed utilization efficiency, FCR values were higher for the freshwater group than for the random group, and higher for the random group than for the selected group, with the differences tending to become significant as the salinity increased (Figure 4). The survival rates were similar for the selected and the random groups from 0 ppt to 15 ppt, and lower, although not significantly, in the freshwater group. At 20 ppt, the survival rates dropped

significantly, with a more marked decrease for the random and the freshwater groups than for the selected group (Figure 4).

#### 3.3.2 Osmolality and ionic concentration ($Na^+$ and $Cl^-$)

Generally, raising salinity led to an increase in these parameters (Figure 5). Plasma osmolality values from the three groups were similar up to salinity of 15ppt. The differences became significant at 20ppt, with osmolalities in the selected (417.3 ± 17.9 mOsm/kg) and random (414.9 ± 8.2 mOsm/kg) groups significantly higher than in the freshwater group (384.5 ± 10.1 mOsm/kg). At 10 ppt, there were no significant difference between external osmotic pressure (288.8 ± 3.2 mOsm/kg) and plasma osmotic pressure of fish in the three groups. At this level, fish were in conditions that were essentially isosmotic to their internal environment. Similar trends were observed for $Na^+$ and $Cl^-$ concentrations in our experiment.

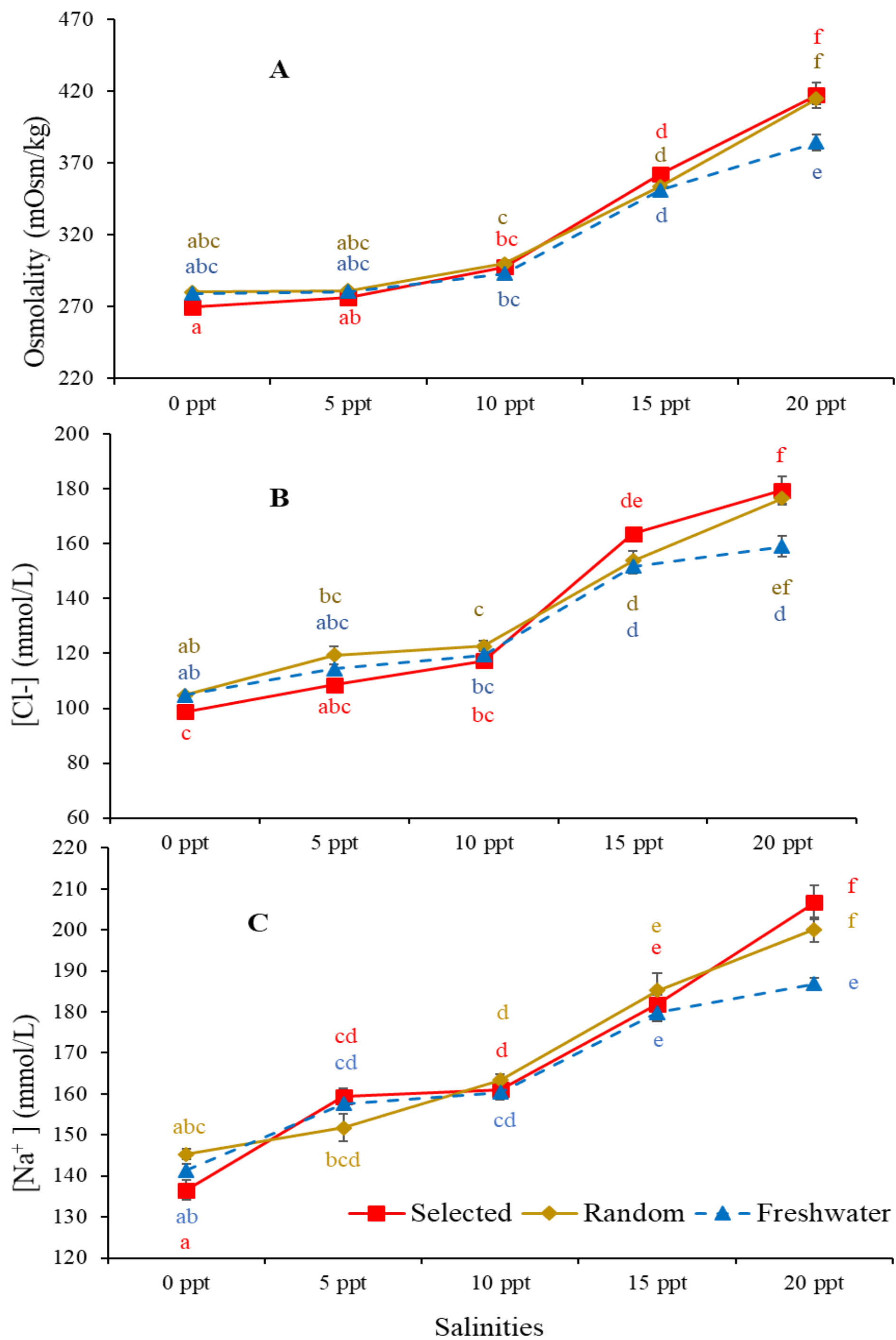


Figure 5: The change of osmolality level (A), ion $Cl^-$ (B) and ion $Na^+$ (C) of three fish groups under saline stress. Values are given least-square means (LSM) ± SEM. Different letters indicate significant difference ($P < 0.05$).

### 3.4 Heritability and selection response

After one generation of mass selection under saline condition, the realized heritability for the body weight of striped catfish was 0.29 (Table 1). Interestingly, our selection procedure led to increases in both harvest weight (18%) and survival rate (11.4%). For harvest weight, we found lower values of the coefficient of variation in the selected group in both generations, with 12.7% in G1 generation and 21.8% in G2 generation, while the values for the random (with 28.8% in G1 and 25.5% in G2) and freshwater (25.1% in G2) groups were similar.

Table 1: Heritability and response to selection for body weight (given same age -11 months at harvest time) after one generation of selection under salinity of 10ppt.

| Parameters | G1 generation | | G2 generation | | |
|---|---|---|---|---|---|
| | Selected | Random | Selected | Random | Freshwater |
| BW (g) | $1,380 \pm 7.3^{a}$ | $793 \pm 10.6^{b}$ | $1,106 \pm 12.8^{c}$ | $935 \pm 13.6^{d}$ | $827 \pm 12.3^{e}$ |
| SR (%) | - | - | $76.9 \pm 4.8^{a}$ | $69.0 \pm 1.5^{a}$ | $57.1 \pm 1.9^{b}$ |
| S for BW (g) | 587 | | - | - | - |
| R for BW (g,%) | - | - | 171 (18.0 ) | | - |
| R for SR (%) | - | - | 7.9 (11.4) | | - |
| $h^2$ (BW) | 0.29 | | | | - |

Values are given as least square means (LSM) ± SEM, BW = body weight, SR = survival rate, S = selection differential, R = selection response, $h^2$ = realized heritability. Different letters indicate significant differences within the same row ($P < 0.05$).

## 4. Discussion

### 4.1 Effect of the salinity on the survival and growth performance of striped catfish

Salinity is one of the critical factors affecting the growth rate, survival rate, geographical distribution and metabolism of fish (Hossanin et al., 2021; Gilles and Patrick, 2001). The analyses in our experiment confirm that salinity can affect the survival and the growth performances of striped catfish in various developmental stages. These impacts have already been shown in previous works: increasing the salinity reduced the probability of eggs reaching an observable developmental stage and hatching rates, delayed the hatching time and development of fertilized eggs, and increased the deformity rate of the larvae (Borode et al., 2002; Huong and Quyen, 2012; Hossanin et al., 2021). Nevertheless, for some freshwater fish, slightly raised salinities could lead to faster growth compared to freshwater condition (Gilles and

Patrick, 2001). A hypothesis potentially explaining this observation is that the energetic cost of osmoregulation is lower in an isosmotic medium (for most freshwater fish, the iso-osmotic salinity normally ranges from around 10 to 12 ppt (Varsamos et al., 2005)), where the gradients between blood and water are minimal, and that energy savings are substantial enough to increase the growth (Gilles and Patrick, 2001). In the current study, we obtained the best growth performances from larva to fingerling stages for salinities of 5 and 10 ppt, and the performances started to decline for higher salinities. This decline of the performances for higher salinities have been observed in other studies. Nguyen et al. (2014) reported a poor survival rate (38.2%) for striped catfish in 18 ppt and Amornsakun et al. (2017) observed a poor survival rate (16.7%) for snake head fish (*Channa striatus*) at 14 ppt. Consequently, our results support the statement that striped catfish can tolerate low to moderate saline conditions, with minor and sometimes positive effects on their survival and growth performances.

### 4.2 Improvement of the salinity tolerance of striped catfish using selection

The results presented in the preceding section show the effect of the salinity on the measured parameters (Figures 2 to 4). In addition, we also observed differences in salinity tolerance of striped catfish in our selection program. In general, the group with the offspring of the fish selected on size ("selected" group) tended to perform better in terms of growth and survival than the offspring of the fish from the randomly chosen fish ("random" group). We interpret this difference as a response to the selection in the previous generation (see also next paragraph). Furthermore, the fish originating from parents raised in saline conditions performed better than those coming from parents kept in freshwater did. Several hypotheses might support these observations. A first hypothesis is that fish raised in a saline environment might have a better tolerance to stressful osmotic conditions in comparison to those raised in fresh water. Indeed, stress due to the environmental salinity is known to induce a costly adjustment of the osmoregulation, which consumes up to 50% of total energy (Gilles and Patrick, 2001; Hasan et al., 2017). This energetic expense has a negative impact on the growth rate. As a first step to confirm this hypothesis, we have observed in our study that striped catfish are able to regulate internal osmolality to adapt to the salinity fluctuations. More importantly, we have observed significantly higher osmolalities in the selected and the random groups than in the freshwater group at high level of salinity (20 ppt) (Figure 5), indicating a better tolerance to stressful osmotic conditions for the fish raised in a saline environment and a lower energetic expense for osmoregulation. We can propose another hypothesis to explain the differences in growth between the offspring of the fish reared in freshwater and those coming from the randomly selected sample in saline conditions: although not selected on growth, fish from the random selection sample are selected on survival. This selection might contribute to select favorable alleles for growth at the same time if growth and survival

are correlated traits. Further studies are needed to potentially support this hypothesis, such as a genome-wide association study (GWAS) designed to identify quantitative trait loci (QTL) underlying the difference between the two sets of fish. Alternatively, we might consider using candidate genes approach (maybe targeting genes involved in osmoregulation) to explain part of the genetic difference between rearing groups.

Coming back to the comparison between the selected and the random groups in saline conditions, our hypothesis is that mass selection in the previous generation has increased the frequency of favorable alleles for salinity tolerance in the selected group. The transmission of these variants to the next generation should lead to the observed better performances for the offspring of the selected fish than for those of the randomly selected fish. These results are therefore a confirmation that selection in saline conditions could be a first step towards genetic adaptation to the new environmental conditions (Donelson et al., 2019). Several previous studies provide additional support to this hypothesis. A genetic component of the salinity tolerance has been observed in several studies and species, such as Arctic charr (*Salvelinus alpinus*) (Norman et al., 2011) and Nile tilapia (*Oreochromis niloticus*) (Gu et al., 2018; Rengmark et al., 2007). In striped catfish, a number of putative genes potentially related to salinity tolerance have been identified using RNA sequencing data (Thanh et al., 2014; Viet et al., 2016). The increased salinity tolerance obtained in selection programs has been reported for tilapia in several previous studies (Jaspe and Caipang, 2011; Tran et al., 2008). Similarly, Afonso et al. (1998) have provided an extensive review of successful selection programs for stress tolerance on rainbow trout and common carp. Progressive genetic adaptation to salinity has been documented in laboratory strains of guppy (*Poecilia reticulate*) (Shikano et al., 2001; Shikano and Fujio, 1998a, 1998b). Purcell et al. (2008) also showed that mosquitofish *(Gambusia affinis)* living in saline environment required genetic adaptation through natural selection for higher individual salinity tolerance than fish from freshwater conditions where no such selection for salinity tolerance exists. This process can sometimes occur remarkably rapidly in a few generations if such favorable variants exist in the challenged population, (Barrick and Lenski, 2013), as seems to be the case in our experiment. Furthermore, gametes, embryos and larvae might undergo selection for alleles that provide advantage in the parental environment, particularly in highly fecund species as fish (Torda et al., 2017). High phenotypic variance coupled with a reasonable heritability indicate that the high genetic diversity underlying a trait could be efficiently used in a selection program (Tahapari et al., 2018). In our study, the large phenotypic variation observed for growth (Table 1) and the corresponding reasonable heritability (0.29) suggest opportunities for a successful selection program targeting an improved tolerance and a progressive adaptation to the emerging saline conditions.

### 4.3 Heritability and response to selection

Our study demonstrated that genetic selection effectively improved the performance of striped catfish in brackish water of moderate salinity (10 ppt). The gain achieved for body weight in this population of striped catfish averaged 18.0% after one generation of selection. This achievement was higher than those reported in striped catfish from a selection program conducted under favorable freshwater pond environment with only 9.3% per generation (Vu et al., 2019). However, this is in line with genetic gain found in other fish species, such as tilapias (Bentsen et al., 2017), common carp (Ninh et al., 2013), channel catfish (Rezk et al., 2003) and Atlantic salmon (Thodesen et al., 1999). In general, the genetic gain achieved after one generation in our breeding program under brackish water was comparable to those reported in aquatic animals, which are about 12.7% on average (Gjedrem and Rye, 2018).

Furthermore, our results also showed that selection for increased growth performance in saline conditions seemed to enhance the survival rate of striped catfish during the grow-out phase (11.4%), although this increase was not significant. The survival gain obtained in the present study was also remarkably higher than that reported for striped catfish selection program carried out under freshwater condition of only 7.4% (Vu et al., 2019). In the present study, selection in G1 generation was obviously made among the surviving fish under saline stress, making survival part of the tolerance to salinity and, as such, a selected trait. Survival is also an economically important trait for striped catfish in brackish water system because it affects the number of fish harvested and marketed, and hence the production yield per unit of culture and the farmers' income. In other species, responses to selection for both body weight and grow-out survival were positive in giant freshwater prawn (Vu et al., 2017), negative in Pacific white shrimp (Zhang et al., 2017) and non-significant in common carp (Dong et al., 2015; Ninh et al., 2014). In Nile tilapia, the long term-selection program for high growth in the Genetically Improved Farmed Tilapia (GIFT) strain over 10 generations did not cause any changes in grow-out survival in pond (Hamzah et al., 2017).

The genetic gains of body weight and survival obtained in our selection program under saline condition suggest that selection to improve ability of striped catfish to adapt to saline stress can be achieved. These results also demonstrate for the first time that at least two major production traits (survival and growth) associated with the ability to adapt to salinity were positively modified after only a single generation of selection, indicating again that this population of striped catfish have the genetic potential to adapt to salinity changes. Similarly, improved tolerance of striped catfish observed in this study has been found for disease resistance to specific pathogens for salmonids (Leeds et al., 2010; Storset et al., 2007) and tilapia (Shoemaker et al., 2017; Sukhavachana et al., 2019). Among these studies, several have

demonstrated moderate to high genetic gains per generation, for example 18.7% for infectious pancreatic necrosis resistance in Atlantic salmon (Storset et al., 2007), and 9% for streptococcosis resistance in red tilapia (Sukhavachana et al., 2019).

Although culture of this improved strain of striped catfish in current hatcheries conditions (*i.e.* in freshwater) could be an objective, it is likely that saline-tolerant strains of striped catfish such as the one developed in our program could also perform well in either low salinity or freshwater environment. Specifically, growth performance of fish from the selected group was slightly higher than those from the random and freshwater groups in 5 ppt and 0 ppt treatments although selection program was implemented in 10 ppt. These findings seem to be in agreement with the hypothesis proposed by Falconer (1990) based on a review of 24 studies in both farmed animals and model species that antagonistic selection (*i.e.* selection under unfavorable environments as saline stress in our study) may produce genotypes that can perform well across a range of production system environments. In contrast, selection under favorable environmental conditions (*i.e.* synergistic selection) may result in sensitive genotypes that do not thrive under stressed conditions. This hypothesis was supported in the study of Thoa et al. (2016): a genetic line of Nile tilapia having undergone five generations of selection under a moderate saline water environment could be cultured successfully in freshwater systems. It suggests that our breeding program to improve growth performance of striped catfish under saline conditions can provide catfish seeds for both freshwater and saline water production systems.

The moderate heritability estimates obtained for body weight under saline condition in our study is generally in line with those reported in other selection programs for striped catfish in favorable freshwater environment with a range from 0.21 to 0.34 after three generations of selection (Sang et al., 2012; Vu et al., 2019). In a selection program on channel catfish (*Ictalurus punctatus*) grown in earthen ponds, realized heritability for three generations of selection were 0.17 and 0.19 for Kansas and Marion strains, respectively (Rezk et al., 2003). In a breeding program on Nile Tilapia, Tran et al. (2008) reported heritability estimates of 0.24 and 0.19 for harvest weight after three generations with selection in freshwater and in brackish water respectively. The heritability achieved for body weight after one generation of selection in the current study suggests that the population will continue responding to future selection rounds under saline conditions.

## 5. Conclusion

The results from this study demonstrate that a selective breeding for increased saline tolerance of striped catfish can already be effective after one generation of selection under saline environment. The improved salinity tolerance of striped catfish developed in the present study is of practical significance in

the context of Mekong Delta aquaculture, where the striped catfish breeding activities are projected to be greatly affected by salinity intrusion due to climate change. The selected animals originating from this study also constitute a valuable genetic resource for subsequent studies targeting a better understanding of the physiology and genomic characteristics of this improved strain as well as the mechanisms of osmoregulatory adaptation to salinity in striped catfish.

**Acknowledgements**

This work was part of PANGAGEN project, supported by ARES-CCD (Académie de Recherche et d'Enseignement Supérieur – Commission de la Coopération au Développement and funded by the Belgian Development Cooperation (DRP/TPS 2017). We would like to thank the staff of Genomics Platform, GIGA, University of Liege for sequencing all DNA samples. We are also grateful to PANGAGEN team at Can Tho University and numerous volunteer students for their works in catching and sampling fish during the experiment.

**Data accessibility**

Additional information about the results of parentage assignment, the husbandry of the various stages, and a summary of water quality parameters are presented in Supplementary material of Chapter 4 in appendices.

# Chapter 5

# A high-quality genome assembly of striped catfish (*Pangasianodon hypophthalmus*) based on highly accurate long-read HiFi sequencing data

---

*** *Publised in Genes: https://doi.org/10.3390/genes13050923*

Dao Minh Hai [1,2], Duong Thuy Yen[2], Pham Thanh Liem[2], Bui Minh Tam[2], Do Thi Thanh Huong[2], Bui Thi Bich Hang[2], Dang Quang Hieu[2], Mutien-Marie Garigliany[3], Wouter Coppieters[5], Patrick Kestemont[4], Nguyen Thanh Phuong[2], Frédéric Farnir[1].

[1.] *FARAH/Sustainable Animal Production, Faculty of Veterinary Medicine, University of Liege (B43), 4000-Liege, Belgium*

[2.] *College of Aquaculture and Fisheries, Can Tho University, Viet Nam.*

[3.] *FARAH/Veterinary Public Health, Faculty of Veterinary Medicine, University of Liege (B43), 4000-Liege, Belgium*

[4.] *Research Unit in Environmental and Evolutionary Biology, Institute of Life, Earth & Environnment, University of Namur, rue de Bruxelles 61, B-5000 Namur, Belgium.*

[5.] *Genomics Platform, GIGA, University of Liege, 4000-Liege, Belgium*

# Chapter 5: A high-quality genome assembly of striped catfish (Pangasianodon hypophthalmus) based on highly accurate long-read HiFi sequencing data

## Preamble

In the Chapter 4, the genetic gains for body weight and survival rate after one generation of selection under saline conditions suggest that selection can be effective to improve ability of striped catfish to cope with saline stress. To to increase the genetic gains, selection process should be continued with more generations, especially focusing on modern approaches such as MAS or GS. However, the genomic resources of striped catfish are still limited. For example, the genome sequence of this species is still in a draft state due to the high contents of repeat sequences in this genome. Recently, PacBio introduced the HiFi long reads with lengths up to 25 kb and with high read accuracy of > 99.9 %. This technology allows rapid, efficient, contiguous and complete assemblies of even the most complex large genome with highly repetitive genomic regions. In the present study, we have constructed the high-quality genome of striped catfish based on HiFi sequencing data. The improved version of striped catfish genome is expected to provide a foundation that can facilitate basic research such as DNA marker discovery, resolution of difficult regions and access to novel genetic variation. This new knowledge should be useful for future work, especially for marker-assisted selection and/or genomic selection to improve salinity tolerance, but also other economically traits of striped catfish.

## Abstract

The HiFi sequencing technology yields highly accurate long-read data with accuracies greater than 99.9% that can be used to improve results for complex applications such as genome assembly. Our study presents a high-quality chromosome-scale genome assembly of striped catfish (Pangasianodon hypophthalmus), a commercially important species cultured mainly in Vietnam, integrating HiFi reads and Hi-C data. A 788.4 Mb genome containing 381 scaffolds with an N50 length of 21.8 Mb has been obtained from HiFi reads. These scaffolds have been further ordered and clustered into 30 chromosome groups, ranging from 1.4 to 57.6 Mb, based on Hi-C data. The present updated assembly has a contig N50 of 14.7 Mb, representing a 245-fold and 4.2-fold improvement over the previous Illumina and Illumina-Nanopore-Hi-C based version, respectively. In addition, the proportion of repeat elements and BUSCO genes identified in our genome is remarkably higher than in the two previously released striped catfish genomes. These results highlight the power of using HiFi reads to assemble the highly repetitive regions and to improve the quality of genome assembly. The updated, high-quality genome assembled in this work will provide a valuable genomic resource for future population genetics, conservation biology and selective breeding studies of striped catfish.

Keywords: striped catfish; chromosome-scale genome assembly; selective breeding; HiFi reads

## 1. Introduction

The striped catfish (*Pangasianodon hypophthalmus)* is a freshwater species that is widely cultured in the Mekong River Delta, Vietnam (Nguyen and Dang, 2010; Phan et al., 2009). In 2021, striped catfish farming in the Mekong Delta produced 1.48 million tons, being cultured on a farming area of approximately 5400 ha, with an export value of 1.61 billion USD (VASEP, 2021). Catfish aquaculture is however facing several challenges, including the lack of genetically selected broodstocks to respond to the high mortality caused by diseases and increasing salinity intrusion in the culture area due to climate change (De Silva and Nguyen, 2011; Hai et al., 2020; Hoa et al., 2021). Use of the genomic information in selection programs might contribute by supplying necessary tools to tackle these challenges in striped catfish aquaculture and thus ensure its sustainability and profitability.

Genomic information provides powerful tools to enhance the fundamental research and applications for genetic improvement programs across many aquaculture species (Abdelrahman et al., 2017; Yue and Wang, 2017). The whole genome sequence is now available for several economically important fish species such as channel catfish (*Ictalurus punctatus*) (Liu et al., 2016), Atlantic salmon (*Salmo salar*) (Lien et al., 2016), Nile tilapia (*Oreochromis niloticus*) (Brawand et al., 2015), rainbow trout (*Oncorhynchus myskiss*) (Berthelot et al., 2014), and Atlantic cod (*Gadus morhua*) (Star et al., 2011). Based on these reference genomes, genomic markers such as single nucleotide polymorphisms (SNPs) or microsatellites can be detected and used in parentage assignment tools (Griot et al., 2020), linkage maps can be obtained, allowing for example QTL/GWAS (Quantitative Trait Loci/Genome-Wide Association Study) to be performed. These genomic tools can then be used in breeding programs, including marker-assisted selection (MAS), genomic selection (GS), and genome editing. For instance, genomic resources have been used for mapping QTL for feed conversion efficiency in crucian carp (*Carassius auratus*) (Pang et al., 2017), and GWAS for growth rate and MAS for disease resistance in Atlantic Salmon (*Salmo salar*) (Gutierrez et al., 2015; Moen et al., 2015).

For striped catfish, two version of genome assemblies have been reported to date, with one based only on Illumina short reads (Kim et al., 2018), and another one based on an hybrid approach combining Illumina short reads, Nanopore long reads and Hi-C data (Gao et al., 2021). Illumina reads are accurate and have a low cost per base, but their read length, typically less than 600 bases, are shorter than many repeat elements in the genomes (Amarasinghe et al., 2020; Derakhshani et al., 2020). Consequently, de novo

assembly using short reads often fails to resolve a full genome due especially to difficulties in the majority of repetitive regions. In contrast, long reads sequencing technologies (or third-generation sequencing) can generate reads longer than 10 kb that can span highly repetitive regions of genomes, and bridge previously intractable gaps in assemblies to improve overall contiguity (Nath et al., 2021; Pollard et al., 2018). The main drawback of long reads is a lower accuracy (an average accuracy of 90% for Nanopore and PacBio Sequel) (Lin et al., 2016) compared to Illumina short reads (99.9%) (Shendure et al., 2017). As a result, the draft genome produced only by long reads were more contiguous, but they also contained more errors (Chaisson et al., 2015; Jain et al., 2018). Combination of the advantages of the two different sequencing technologies in so-called hybrid approaches can overcome their drawbacks (Nowak et al., 2019; Tan et al., 2018). In the hybrid approach, long-read assemblies are polished with Illumina short reads to improve accuracy (Hu et al., 2020; Walker et al., 2014). On the negative side, most polishing tools may be unable to fix errors in repetitive regions that have ambiguous short-read mappings (Jain et al., 2018; Wenger et al., 2019a) and combining sequencing technologies may lead to increased costs and to more complexity in the projects.

As an alternative approach to eliminate dependencies on short-read data polishing, PacBio recently introduced high-fidelity reads (HiFi reads), with an average length of 13.5 kb, and a substantial increase in the base accuracy compared to previous long reads (greater than 99.9%) (Logsdon et al., 2020; Wenger et al., 2019a). The highly accurate long-read HiFi sequencing data have been proven to significantly improve the continuity and completeness of genome assembly in many studies (Vollger et al., 2020; Yang et al., 2022). Several new assemblers have been developed to fully exploit HiFi reads, that do not require a final polishing phase (Di Genova et al., 2021; Nurk et al., 2020). The assembly exclusively based on long highly accurate PacBio HiFi reads outperforms Illumina-Nanopore hybrid and Nanopore assembly. In addition, *de novo* assemblers using HiFi reads required less data compared to other strategies. Using HiFi data, a 16x genome coverage is sufficient in order to produce high quality assemblies (Gavrielatos et al., 2021). For the human genome, the high accuracy of PacBio HiFi sequence data has improved variant discovery, reduced the assembly time and provided access to even more complex regions of repetitive DNA (Huddleston et al., 2017; Logsdon et al., 2020; Wenger et al., 2019b). With HiFi reads, more than 50% of the regions previously inaccessible with Illumina short-read sequence data in GRCh37 human reference genome are now accessible (Wenger et al., 2019b). For most aquatic species, due to a high content of repetitive DNA sequences in their genomes, the assembly of the genome was extremely complicated (Houston et al., 2020). For this reason, the HiFi sequencing technology may be a better choice to assemble

the high quality genome of striped catfish. Recent studies demonstrated that repetitive sequences play an important role in regulation of gene expression (Mehrotra and Goyal, 2014). Consequently, a high-quality genome with higher proportion of repetitive regions identified will offer more opportunities to understand the comprehensive biology, especially in aquaculture species (Yue and Wang, 2017). Moreover, by using this improved reference genome, a larger number of DNA markers (e.g., SNPs) with high quality can be identified (Georges et al., 2019; Houston et al., 2020; Malhis and Jones, 2010) . This phenomenon will facilitate MAS and GS to accelerate genetic improvement in selective breeding programs (Yue and Wang, 2017; Zenger et al., 2019). Using a high-quality reference genome can bring several benefits for downstream genetic applications such as reduction of time and funds needed to implement MAS and GS, and achievement of higher genetic gains in the breeding programs (Benevenuto et al., 2019; Houston et al., 2020; Yue and Wang, 2017). In a study on the blueberry (*Vaccinium corymbosum*), using a high-quality reference genome resulted in a higher precision about the location, number, and gene action of QTLs, by that improving our understanding of the genetic architecture of the traits through GWAS analyses and higher chances to uncover the molecular mechanisms underpinning the trait variation (Benevenuto et al., 2019).

In the present study, an improved (chromosome-level) assembly was developed for striped catfish by using PacBio HiFi long reads. The assembly was then scaffolded to chromosome level using Hi-C data. This improved reference assembly represents a step towards improving our understanding of fundamental biological and evolutionary questions and towards the genetic improvement of important aquaculture production traits via genomic-assisted breeding of striped catfish.

## 2. Materials and Methods

### 2.1 Sample collection and DNA extraction

This study was carried out using striped catfish (*Pangasianodon hypophthalmus*) collected from a (8°33' - 10°55' N; 104°30' - 106°50' E) hatchery in the Mekong Delta Vietnam. High quality genomic DNA was extracted from blood of a striped catfish recently, a new study (Wen et al., 2022) suggested that *Pangasianodon hypophthalmus* has a XY sex determination system) using NucleoBond HMW DNA kit (Macherey-Nagel, Germany) for PacBio sequencing and the Maxwell® 16 Blood DNA Purification Kit (Promega, Madison, WI, USA) for Illumina sequencing . DNA quality and quantity were evaluated using

electrophoresis on a 1% agarose gel, and using Quant-iT™ PicoGreen ® dsDNA Reagent and Kits (Thermo Fisher Scientific, Waltham, MA, USA). Then, the DNA was sequenced using two libraries, with short and long reads.

### 2.2 Library construction and DNA sequence

The first library was generated to obtain long PacBio HiFi reads. The HiFi sequencing library was prepared using SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA). The library was further size-selected using BluePippin System from SAGE science. The library was sequenced on one 8M SMRT cell on Sequel II instrument using Sequel II Binding kit 2.0 and Sequencing chemistry v2.0. The sequencing produced 3,902,121 polymerase reads (~ 270.8 Gb of raw data) with an average polymerase read length of 69.4 kb and an average insert length of 16.4 kb. Circular Consensus Sequences (CCS) (PacBio, CA, USA) were generated using CCS pipeline (SMRT Link v9.0.0.92188) with following settings: minimum number of passes: 1, minimum predicted accuracy: 0.9. CCS reads with at least 99% accuracy are called HiFi reads. This sequencing was performed by the Norwegian Sequencing Centre (www.sequencing.uio.no) (accessed on 1 October 2020) from the University of Oslo, Norway.

In addition to PacBio reads, we employed Illumina paired-end sequencing by synthesis technology to sequence the striped catfish genome using the whole genome shotgun approach. Paired-end libraries with 350 and 550 bp fragments were constructed using a TruSeq DNA PCR-Free Kit (Illumina) according to manufacturer protocols. The libraries were sequenced using Illumina NextSeq 500 sequencing platform with Illumina protocols for whole-genome shotgun sequencing with paired end reads (2 x 150 bp). This work was performed by the Interdisciplinary Center for Biomedical Research (GIGA) at the University of Liege, Belgium.

### 2.3 Sequence data processing and genome assembly

Raw genomic library data generated by Illumina were quality checked using FastQC v.0.11.9 (Andrews, 2010). We performed data filtering and trimmed adapter sequences, low-quality reads and PCR duplication using Trimmomatic v 0.39 (Bolger et al., 2014) with parameters ILLUMINACLIP: TruSeq3-PE.fa:2:30:20 LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:50.

The initial characterization of the striped catfish genome was estimated through *k-mer* (k=21 in this study) analysis of the HiFi data by Jellyfish v 2.3.0 (Marçais and Kingsford, 2011) . The histograms were uploaded to GenomeScope version 2.0 (Vurture et al., 2017) for estimation of genome size, level of duplication and heterozygosity.

The HiFi reads were assembled into contigs using Hifiasm-0.16.1 (Cheng et al., 2021) using default parameter values. To scaffold genome, the cleaned Hi-C reads (with a total of 90 Gb) from a previous study on striped catfish (Gao et al., 2021) were mapped to the primary assembly by BWA version 0.7.17 (Li and Durbin, 2009) to generate a BAM file which was subsequently converted to BED file. Then, SALSA version 2.3 (Ghurye et al., 2019) used this BED file, which contained the mapping information of Hi-C reads on the assembly, to scaffold primary assembly.

To build chromosome-level assembly scaffolds, the cleaned Hi-C reads were first mapped to the assembled genome using BWA version 0.7.17 (Li and Durbin, 2009). Scaffolds were then clustered, ordered, and oriented using Lachesis (Burton et al., 2013), with the parameter set "CLUSTER_N = 30,CLUSTER_MIN_RE_SITES = 25, CLUSTER_MAX_LINK_DENSITY = 2, CLUSTER_NONINFORMATIVE_RATIO = 3". Figure 1 shows the assembly and annotation workflow.

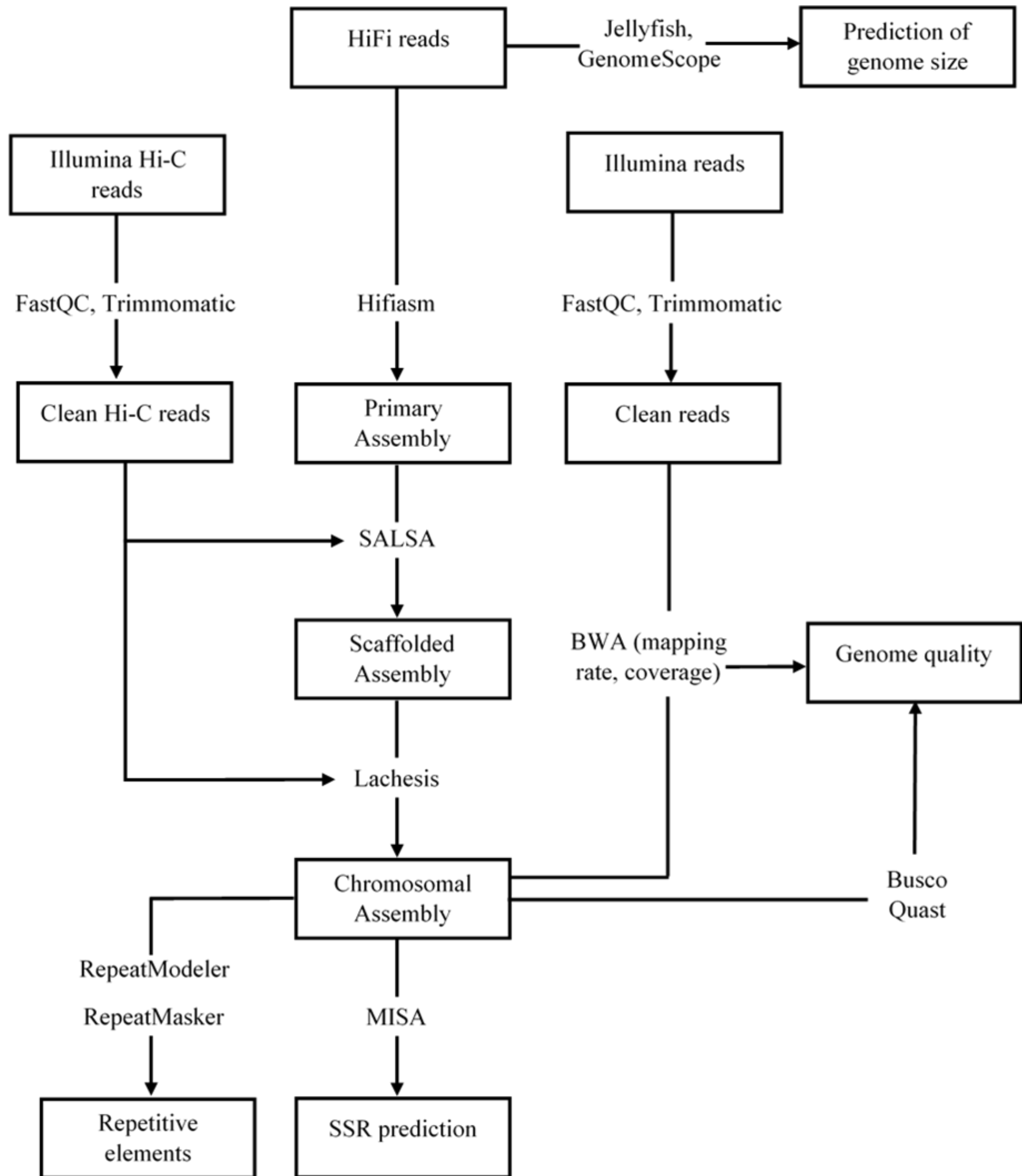


Figure 1: Detailed workflow for de novo whole-genome assembly and annotation

### 2.4 Genome quality assessment

The quality of our final genome assembly was evaluated in terms of three data sets:

1) BUSCO Actinopterygii Genes. Quality assessment was conducted using BUSCO software version 5.2.2 (Simão et al., 2015) with default parameters by searching the genome against 3,640 single-copy orthologues from actinopterygii_odb10 database (https://busco-data.ezlab.org/v5/data/lineages/)(accessed on 10 November 2021);
2) Illumina short DNA reads. To assess the accuracy of our genome assembly, we aligned the Illumina short reads to the assembly using BWA version 0.7.17 (Li and Durbin, 2009) to evaluate the mapping and covering rate;
3) Comparison to other assemblies. Contig N50 values were calculated for comparison with those of previous studies on striped catfish (Gao et al., 2021; Kim et al., 2018) and other Siluriformes species by QUAST version 5.0.2 (Gurevich et al., 2013). In addition, we also mapped our newly assembled genome against the previously published one: (Genbank accession no: GCA_003671635.1) (Kim et al., 2018) and GCA_016801045.1 (Gao et al., 2021) to identify the gaps using MUMMER version 3.23 (Kurtz et al., 2004).

### 2.5 Identification of repetitive elements and simple sequence repeat markers

Prediction of repeat elements was based on de novo and homology methods. In the homology approach, we searched the genome for repetitive DNA elements using RepeatMasker version 4.1.2 (Tarailo-Graovac and Chen, 2009) with parameters "-species siluroidei" based on the known repeats library (Dfam 3.5 database, https://www.dfam.org/releases/Dfam_3.5/families/) (accessed on 15 November 2021). To identify repetitive element de novo, RepeatModeler version 2.0.2 (Flynn et al., 2020) was primarily used to generate a repeats library. Then, this library was aligned to the assembled genome with RepeatMasker version 4.1.2 (Tarailo-Graovac and Chen, 2009). In addition, the final assembled scaffolds were analyzed to identify simple sequence repeats (SSRs) such as di-, tri-, tetra-, pentra-, and hexa-nucleotide repeats using MISA version 2.1 (Thiel et al., 2003).

## 3. Results

### 3.1 DNA sequencing

For HiFi data, a total of 17.28 Gb (Q ≥ 20) were generated, corresponding to 1,68,528 reads with an average read length of 14,791 bp, and a median of read quality of Q = 30. The distribution of read lengths and read qualities is presented in Figure S1. The data obtained from the two paired-end libraries of 350-bp and 550-bp insert sizes reached 51.2 Gb and 49.5 Gb, respectively (Table 1). A *k-mer* analysis predicted that the striped catfish had an estimated genome size of 713,9 Mb, the heterozygosity rate of the genome was 0.56%, and the repeats content of the genome was 17.5% (Figure S2).

Table 1: Summary of sequencing data for striped catfish genome assembly

| Libraries | Insert size (bp) | Total data (Gb) | Read length (bp) | Sequence coverage (x) * |
|---|---|---|---|---|
| Illumina reads | 350 | 51.2 | 150 | 71.8 |
| Illumina reads | 550 | 49.5 | 150 | 69.3 |
| PacBio (HiFi) reads | 16,400 | 17.28 | 14,791 | 24.2 |
| Total | | 118.0 | | 165.3 |

** The coverage was calculated according to estimated genome size of 713,911,345 bp*

### 3.2 Genome assembly

The initial assembly of the striped catfish genome was based on HiFi reads only using Hifiasm *de novo* assembler, which resulted in a genome size of 788.1 Mb with 845 contigs and a contig N50 of 14.7 Mb. Next, the contigs were scaffolded using SALSA with Hi-C data, which generated a genome size of 788.4 Mb with 381 scaffolds and scaffold N50 of 21.8 Mb (Table 2). We also found that the average GC content in the striped catfish (38.96%) was similar to those in other species, that range from 33.62% to 42.01% (Tang et al., 2021). Furthermore, 315 scaffolds (99.3% of genome) were successfully clustered into 30 chromosome groups. Finally, we obtained a high-quality chromosomal-level genome with a total size of 778.9 Mb (Table 3)

Table 2: Summary of the final striped catfish genome assembly

| Category | Contig | | Scaffold | |
|---|---|---|---|---|
| | Length (bp) | Number | Length (bp) | Number |
| Total | 788,121,403 | 845 | 788,355,903 | 381 |
| Largest | 30,145,618 | NA | 35,439,358 | NA |
| N50 | 14,675,983 | 20 | 21,837,136 | 15 |
| N60 | 11,176,154 | 26 | 19,579,817 | 19 |
| N70 | 6,525,798 | 35 | 17,728,427 | 24 |
| N80 | 2,330,994 | 55 | 14,533,765 | 28 |
| N90 | 459,384 | 139 | 1,263,789 | 54 |

*N/A: not applicable.*

Table 3: Statistics of chromosomal level assembly of Striped catfish

| Chr ID | Length (bp) | Chr ID | Length (bp) | Chr ID | Length (bp) |
|---|---|---|---|---|---|
| Chr1 | 57,614,776 | Chr11 | 30,308,490 | Chr21 | 20,241,537 |
| Chr2 | 57,569,486 | Chr12 | 30,160,703 | Chr22 | 20,107,636 |
| Chr3 | 47,895,633 | Chr13 | 27,700,161 | Chr23 | 19,020,310 |
| Chr4 | 40,963,295 | Chr14 | 26,554,853 | Chr24 | 18,006,869 |
| Chr5 | 37,055,990 | Chr15 | 26,362,405 | Chr25 | 16,472,023 |
| Chr6 | 33,881,059 | Chr16 | 25,536,520 | Chr26 | 12,481,112 |
| Chr7 | 33,777,686 | Chr17 | 25,407,408 | Chr27 | 6,262,317 |
| Chr8 | 33,526,538 | Chr18 | 22,657,484 | Chr28 | 2,381,093 |
| Chr9 | 32,744,715 | Chr19 | 22,587,380 | Chr29 | 1,580,235 |
| Chr10 | 32,245,061 | Chr20 | 20,661,286 | Chr30 | 1,373,485 |
| Total chromosome-level length | | | 783,137,546 | | |
| Total length | | | 788,355,903 | | |
| Chromosome length/total length | | | 99.3% | | |

### 3.3 Genome quality evaluation

BUSCO analysis of 3640 actinopterygii single-copy orthologues revealed that the proportions of complete (C), complete and single-copy (S), complete and duplicated (D), fragmented (F), and missing (M) genes were 96.0%, 93.5%, 2.5%, 1.0% and 3.0%, respectively (Table 4). 3,497 (96.0%) core genes were completely identified in our assembly genome, suggesting a high completeness of genome assembly. Using short reads, approximately 99.62% of these reads were mapped to our genome assembly, and these reads covered 94.5% of the total assembly.

Table 4: Genome assessment based on BUSCO annotations.

| Index | Number |
|---|---|
| Complete BUSCOs (C) | 3,497 |
| Complete and single-copy BUSCOs (S) | 3,405 |
| Complete and duplicated BUSCOs (D) | 92 |
| Fragmented BUSCOs (F) | 35 |
| Missing BUSCOs (M) | 108 |
| Total BUSCO groups searched (n) | 3,640 |
| C:96.0%[S:93.5%,D:2.5%],F:1.0%,M:3.0%,n:3,640 | |

For assessment using other assemblies, we compared our genome assembly to the two previous versions of the striped catfish genome as well as to those from other *Siluriformes* species. Results showed that our genome assembly had a similar size to those of other *Siluriformes* species (Table 5), and the contig N50 of the present genome assembly is significantly higher than those from the other previously published ones, including bighead catfish (GCA_011419295.1), walking catfish (GCA_003987875.1), and channel catfish (GCA_004006655.2) (Table 5). Moreover, the contig N50 value of our version of the genome (14.7 Mb) was 245-fold greater than the previous version of (Kim et al., 2018) based only on Illumina reads (0.06 Mb) and 4.2-fold greater than the version of (Gao et al., 2021) based on Illumina reads, Nanopore long reads and Hi-C reads (3.5 Mb) (Table 6). In addition, the results from aligning our genome against the two previous versions indicated that 29,987 and 24,587 gaps were filled with respect to the versions of (Kim et al., 2018) and (Gao et al., 2021), respectively. These gaps corresponded to 72.7 Mb and 45.8 Mb of sequence in our new assembly, respectively.

Table 5: Comparison of the genome assemblies of various *Siluriformes* species.

| Species | Genome size (Mb) | Number of contig | Contig N50 |
|---|---|---|---|
| *Pangasianodon hypophthalmus* (from the present study) | 788.4 | 845 | 14.7 |
| *Clarias macrocephalus* (GCA_011419295.1) | 883.3 | 44,869 | 0.05 |
| *Clarias batrachus* (GCA_003987875.1) | 821.7 | 78,047 | 0.02 |
| *Ictalurus punctatus* (GCA_004006655.2) | 1,002.3 | 5,816 | 2.7 |
| *Ageneiosus marmoratus* (GCA_003347165.1) | 1,030 | 169,048 | 0.007 |
| *Ompok bimaculatus* (GCA_009108245.1) | 718.1 | 27,068 | 0.08 |
| *Bagarius yarrelli* (GCA_005784505.1) | 570.8 | 928 | 1.9 |
| *Tachysurus fulvidraco* (GCA_003724035.1) | 713.8 | 2,402 | 1.0 |

Table 6: Comparison of quality metrics of this study and the previous striped catfish genome assemblies

| Genomic feature | This study | Gao et al. (2021) | Kim et al. (2018) |
|---|---|---|---|
| The size of genome (Mb) | 788.4 | 742.6 | 715.7 |
| Number of contigs | 845 | 821 | 23,340 |
| Contig N50 (Mb) | 14.7 | 3.5 | 0.06 |
| Longest contig (Mb) | 30.1 | 16.1 | 0.5 |
| GC content (%) | 38.9 | 38.9 | 38.7 |
| Repetitive regions (%) | 39.3 | 36.9 | 33.8 |
| Complete BUSCOs (C) (%) | 96.0 | 93.3 | 92.3 |

### 3.4 Repetitive genome elements and SSR markers

The results from an analysis of repeated elements showed that 308.36 Mb (39.11%) and 331.92 (42.10%) of the striped catfish genome consisted of repeat sequences when estimated from the *de novo* and homology approaches, respectively. For *de novo* approach, approximately 57.2 Mb (7.25%) of class I retrotransposons were identified (long interspersed nuclear elements [LINEs], 3.87%; short interspersed nuclear elements [SINEs], 0.39%; total long terminal repeat elements, 2.99%). In addition, 101.79 Mb (12.91%) class II DNA transposons and 112.36 Mb (14.25%) unclassified elements were identified (Table S1).

A total of 960,574 SSRs were obtained from the striped catfish genome assembly. The biggest fraction of SSRs were dinucleotides (53.2%), followed by mononucleotides (32.8%), trinucleotides (7.7%), tetranucleotides (5.6%), pentanucleotides (0.5%) and hexanucleotides (0.2%). Among these SSRs, T (49.2%) and A (48.6%) accounted for 97.8% of the total mononucleotide repeats. AC (17.3%), TG (15.9%), GT (15.7%), CA (14.3%) and TC (8.0%) accounted for 71.1% of the dinucleotide repeats, whereas AAT (11.3%), TTA (10.1%), ATT (9.9%), TAA (9.3%), TAT (7.7%) and ATA (7.4%) accounted for 55.7% of trinucleotide repeats (Table S2).

## 4 Discussion

Genome assembly is the process of reconstructing a genome from randomly sampled sequence fragments in which overlapped sequence fragments are referred to as contigs (in the ideal case, one contig per chromosome) (Berlin et al., 2015; Logsdon et al., 2020). The quality of a genome assembly can be often evaluated using metrics related to contigs (contiguity), such as N50, and the ability to complete the whole structure of the genome (completeness), such as BUSCO score (Ellis et al., 2021; Molina-Mora et al., 2020). In the present study, a high quality genome was assembled for striped catfish. This genome included 30 chromosomes, which is consistent with previously reported chromosome analysis in this species (Sreeputhorn et al., 2017). The contigs N50 value of the present genome assembly is significantly higher than in the previous versions (Table 6). In addition, the BUSCO value of the genome obtained in our study (96.0 %) was slightly superior to that in (Z. Gao et al., 2021) and (Kim et al., 2018) with 93.3% and 92.3%, respectively. Our findings were in agreement with previous works, where BUSCO scores and contig N50 of genome assembly were increased when using PacBio HiFi reads (Chin and Khalak, 2019; Driguez et al., 2021; Sharma et al., 2021). Moreover, we also obtained high mapping and coverage rates when we mapped the Illumina short reads on the obtained genome assembly. Together, these results indicated that our new genome assembly was more contiguous and complete than the previous published ones.

The long-read sequencing technologies can generate very long reads that make them invaluable for resolving complex repeats regions that cannot be assembled using shorter reads (Zimin and Salzberg,022). Among these technologies, the HiFi sequencing with both long and highly accurate reads have made it possible to overcome the assembly of highly repetitive regions (Yang et al., 2022). In this study, the highly accurate long-read HiFi sequencing data used to generate the genome assembly have led to remarkable improvements, especially in finding new repetitive sequences. A higher number of repetitive elements (39,1%) were identified in our assembly compared to the previous genome assembly of (Z. Gao et al., 2021)

with 36.9% and (Kim et al., 2018) with 33.8%. Our results are in accordance with prior work stating that a genome assembly using HiFi data has significantly improved continuity and accuracy in many complex regions of the genome, and that the HiFi sequence data assembled an additional 10% of duplicated regions comparing to one assembled using continuous long-reads (CLRs) which are currently the most common PacBio data type with lower accuracy (85-92%) (Vollger et al., 2020). In addition, our assembly also had a higher proportion of repeat sequences than Murray cod (*Maccullochella peelii*) with 23.38% (Austin et al., 2017), but lower than zebrafish (Howe et al., 2013) with 52.0%. Moreover, we identified 72.7 Mb and 45.8 Mb of new sequences that were unassembled in previous genome of (Kim et al., 2018) and (Z. Gao et al., 2021), respectively. Combining this information with the increase of repetitive regions in our new assembly suggests that the new identified sequences might be located in repetitive regions of the genome. Taken together, these results are indicative of the high quality and completeness of the present genome assembly and its potential as an alternative to the previous versions of the striped catfish genome.

The improvement of the quality of a reference genome may often lead to novel discoveries. For example, improvements to rainbow trout (*Oncorhynchus mykiss*) assembly identified novel structure variants between 2 separate lineages and featured a scaffolded sex-determination gene (sdY) in the Y chromosome sequence (G. Gao et al., 2021). The drastic improvements in the present genome could also lead to such new discoveries. It will also contribute to accelerate marker-assisted selection and functional genomic studies in striped catfish. Moreover, it could also facilitate more accurate marker ordering and fine mapping for quantitative trait loci and genome wide association studies.

## 5 Conclusions

In this study, we have generated a high quality chromosome-level reference genome assembly for striped catfish based on HiFi and Hi-C sequencing data. Both the sequence continuity (contig N50) and genome completeness (BUSCO score) were remarkably higher than those of the previously released striped catfish genomes, implicating the advantage of HiFi reads on de novo genome assembly. Moving forward, the reference genome presented here will serve as a springboard for further studies on the molecular evolution, comparative genomics, genotypic and phenotypic variation, genome structure, and genetic studies related to selective breeding on striped catfish.

**Acknowledgments**

We thank the staff of Genomics Platform, GIGA, University of Liege, the Nor- wegian Sequencing Centre, University of Oslo, Norway for sequencing DNA samples, and the Con- sortium des Équipements de Calcul Intensif (CÉCI) for providing computational resources.

**Data accessibility:** Supplementary material of this study is presented in Appendices

# Chapter 6

# DNA reprogramming and modification under a salinity selection programme during early development in striped catfish (*Pangasianodon hypophthalmus*)

**** *Manuscript in preparation*

Dang Quang Hieu[1], Bui Thi Bich Hang[3], Dao Minh Hai[2,3], De Pooter Simon[1], Frederic Silvestre[1], Jérôme Lambert[1], Do Thi Thanh Huong[3], Duong Thuy Yen[3], Nguyen Thanh Phuong[3], Frédéric Farnir[2], Patrick Kestemont[1].

[1]. *Research Unit in Environmental and Evolutionary Biology, Institute of Life, Earth & Environnment, University of Namur, rue de Bruxelles 61, B-5000 Namur, Belgium*

[2]. *FARAH/Sustainable Animal Production, Faculty of Veterinary Medicine, University of Liege (B43), 4000-Liege, Belgium*

[3] *College of Aquaculture and Fisheries, Can Tho University, Viet Nam.*

# Chapter 6: DNA reprogramming and modification under a salinity selection programme during early development in striped catfish (*Pangasianodon hypophthalmus*)

## Preamble

Understanding the molecular mechanisms underpinning phenotypic response of animals to environmental stress, saline condition in our case, is central to biology. Recently, interest in epigenetics has surged as it has become increasingly clear that epigenetic mechanisms can provide a measurable link between environment and phenotype. We demonstrate in Chapter 4 that the selection program has succeeded in developing a new strain of striped catfish better adapted to saline conditions. In addition, the results also indicate that the fish originating from parents reared in saline condition performed better than those coming from parents kept in freshwater. The results suggest a role of epigenetics for difference observed between fish originating from difference environment. To date, the epigenetic factors affecting salinity tolerance in fish are not properly understood, and would draw attention in future studies, especially in striped catfish. If confirmed, the presence of such epigenetic effects could have impact on seed production strategies. Therefore, a combined salinity exposure experiment was conducted at the embryonic to study changes in DNA methylation, a most significant indicator of epigenetic impact. Samples were collected at different stages during early development, namely fertilization, early blastula, late blastula, early gastrula, late gastrula, hatching, 1, 3, and 6 days post-hatching. DNA reprogramming was studied to determine how the methylation pattern occurs during early development. Global CpG methylation and expression of genes related to DNA methylation, osmoregulation and stress were analysed.

## Abstract

Striped catfish *Pangasianodon hypophthalmus* is a freshwater species intensively cultivated in Vietnam as well as elsewhere in South-East Asia. Due to climate change and salinity intrusion in the lower part of the Mekong Delta, a genetic selection programme is in progress to select a strain of striped catfish better adapted to brackish water conditions. The present study was conducted to evaluate the efficiency of the selection program through methylation and growth related processes. A first objective was to determine DNA reprogramming during the early developmental stages under normal freshwater conditions. A second objective was to evaluate the impact of an increase of salinity on DNA reprogramming in a normal and selected strain of striped catfish. Under freshwater rearing conditions, global CpG methylation increased remarkably from fertilization (40.73%±8.83) to the early blastula stages (54.69%±1.87%), followed by a steady raise until the gastrula stage (64.30%±4.83). The expression of genes related to the methylation process was significantly altered during early development. In particular, *dnmt1* mRNA showed a dramatic decrease in the early stages, from fertilization to early gastrulation, and the level remained stable in the later stages. Similarly, *mecp2* accumulation decreased significantly from fertilization to early gastrulation. Conversely, *dnmt3ba4* was significantly upregulated during the same period and then gradually decreased until 1 day after hatching (dph). A remarkable increase from early gastrula stage to somite development was observed in *dnmt3aa* transcript accumulations while *dnmt3ab3* and *dnmt3ab* expression showed similar patterns with the gradual increase in late gastrulation at the end of the experiment, except for a significant drop in *dnmt3ab* at 6 dph. *Tet3* expression increased significantly from late blastula to late gastrula and then dropped to a moderate level at 1 dph. Exposure to a salinity of 2.5 ppt during early development (6 dph larvae) significantly increased *dnmt* genes expression. In addition, the selection programme was conducted on striped catfish exposed to improve tolerant capacity to salinity. Particularly, the F1 generation was cross-bred with fish from 3 different locations and then subjected to 10 ppt salinity until the maturation stage. They were divided into selection and random groups, with the former consisting of 50% fast-growing fish and the latter being randomly selected. At the same time, another group was raised in freshwater (freshwater group) as favourable conditions for the species. Subsequently, the offspring of selection and random groups showed a slightly altered expression of *dnmt1*, *gh1* and *igf1*. In conclusion, these results indicate that DNA reprogramming is dynamic during early striped catfish development and is likely regulated by changes in *dnmt*, *mecp2*, and *tet3* expression. Furthermore, the significant increase in *dnmt* genes expression of 6dph larvae incubated at 2.5 ppt suggests an important role of the salt environment on the methylation process during early development of striped catfish.

## 1. Introduction

In a recent past, it was generally accepted that genetic diversity is the only source of heritable variation in natural populations and that random mutations can explain this genetic variability and drive evolution through natural selection (Mayer and Provine 1998). However, a number of researches have indicated that epigenetic variation can also contribute to heritable changes within a population, and subsequently lead to rapid evolution (Bossdorf et al. 2008; Skinner 2015). Within epigenetic modifications, DNA methylation is the most studied form related to transcriptional regulation, genomic imprinting, deletion of repetitive elements, and DNA-protein interaction during organismal development (Jähner et al. 1982; Li et al. 1993; Kass et al. 1997; Iida et al. 2006). In eukaryotes, DNA methylation occurs primarily through the transfer of a methyl group at the 5-position of cytosine to form 5-methylcytosine (*5 mC*) through regulating DNA family proteins (cytosine-5)-methyltransferases (Dnmt), primarily in *CpG* (cytosine-phosphate-guanine) dinucleotides (Moore et al. 2013). DNA methylation plays a key role for the cellular memory mechanism to maintain cellular identities. However DNA methylation reprogramming is required to obtain a totipotent state and produce the next generation (Zeng and Chen 2019). During embryonic development, epigenetic reprogramming is mainly related to the erase most methylation marks inherited from gametes at preimplantation stages and the reestablishment of global DNA methylation patterns upon implantation (Zeng and Chen 2019). Reprogramming involves methylation deletion in gametes after fertilization, followed by the re-establishment of new methylation marks during the mid-blastula stages (Wang and Bhandari 2019). The process is modified by the regulation of DNMTs in which DNMT1 is responsible for the maintenance of DNA methylation patterns and DNMT3a/b plays the role of *de novo* establishment of DNA methylation (Dabe et al. 2015). In addition, Ten-Eleven translocation (TET) affects DNA reprogramming through a process of active DNA demethylation. These enzymes oxidize 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Wu and Zhang 2017). In medaka, *tet* gene expression was significantly altered in correlation with DNA methylation levels during embryonic stages (Wang and Bhandari 2019). Furthermore, the recruitment of MeCP2 to methylated CpG lead to repress transcription (Fuks et al. 2003). It is also well known that MeCP2 silences the transcription process by recruiting the repressive histone deacetylase machinery, which removes acetyl groups from histones, resulting in gene silencing (Jones et al. 1998; Nan et al. 1998).

DNA methylation is one of the most crucial forms of epigenetic marks that are linked to transcriptional regulation, genomic imprinting, deletion of repetitive elements, and DNA-protein interaction during organism development (Jähner et al. 1982; Li et al. 1993; Kass et al. 1997; Iida et al. 2006). Reprogramming of the epigenome after fertilization erases most epigenetic marks passed on by gametes and establishes new marks during mid-blastula stages (Wang and Bhandari 2019). Although the process of methylation during embryogenesis has been well documented in plant and mammalian organisms (Dean et al. 2003; Feng et al. 2010), such knowledge is not available for fish, where the process was studied so far on model fishes such as zebrafish (Mhanni and McGowan 2004; Potok et al. 2013), medaka (*Oryzias latipes*) (Walter et al. 2002; Dasmahapatra and Khan 2016) and mangrove rivulus (Fellous et al. 2018). In mammals, methylated DNA is demethylated in two ways: passive and active demethylations. Passive demethylation, also known as replication-mediated demethylation, is due to a dilution of the methylation signal on DNA strands that do not receive methylation marks through the DNMT-mediated recruitment process (Rougier et al. 1998; Inoue et al. 2011; Inoue and Zhang 2011; Guo et al. 2014). Otherwise, active DNA demethylation is induced by TET-mediated oxidation of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), followed by replication-dependent dilution of oxidized 5mC or excision, under the action of thymine DNA glycosylase, of 5-fC and 5-caC, coupled with base excision repair (Wu and Zhang 2017).

The epigenome is globally sensitive to the changes of environmental factors, especially during embryonic development (Morey et al. 2015; Ohbo and Tomizawa 2015; Vougiouklakis et al. 2017). Reprogramming of DNA methylation has been suggested to represent a critical window that is extremely sensitive to environmental stressors (Fellous et al. 2018). A previous study revealed that environmental stressors, including cadmium and temperature, upregulate *dnmt1a* and *dnmt3b* gene expressions in zebrafish during embryonic development (Zhou et al. 2021). Salinity is one of the most important environmental factors directly affecting teleost tolerance and behaviour (Varsamos et al. 2005). Additionally, salinity exposure led to alteration of putative promoters of *gh1* protein expression, resulting in the expression of its mRNA in the smooth-tongued sole (*Cynoglossus semilaevis*) (Si et al. 2021). However, nothing is known so far about the possible impact of changes in salinity occuring during fish development on DNA methylation.

The striped catfish *Pangasianodon hypophthalmus* (Sauvage, 1878) (Pangasiidae; Siluriformes) is a freshwater species very important for the aquaculture industry in Vietnam, with export sales reaching

US$2.3 billion in 2019 (FAO 2020). Rising sea level resulting from global warming negatively impacts the natural environment and consequently the production of striped catfish (Nguyen et al. 2014a). Furthermore, the construction of dams upstream the Mekong River is altering its hydrology by significantly reducing the flow, especially during the dry season. The changes exacerbate the salinity intrusion in the Mekong Delta, the main production area for striped catfish (Binh et al. 2020). A breeding selection programme is one of the available methods to increase tolerance to environmental stressors by identifying and selecting parents with the highest possible additive genetic merit for the traits of interest (Farias et al. 2017). In aquaculture, some species have been selected for genetic gain (Atlantic salmon *Salmo salar* and channel catfish *Ictalurus punctatus*) (Gjedrem and Rye 2018), disease resistance (Pacific white shrimp *Litopenaeus vannamei* and red tilapia *Oreochromis* spp.) (Argue et al. 2002; Suebsong et al. 2019) and vegetal feed adaptation (rainbow trout *Oncorhynchus mykiss* and Amago salmon *Oncorhynchus masou ishikawae*) (Yamamoto et al. 2016; Callet et al. 2017). In striped catfish, a selection programme based on growth rate in freshwater was conducted. The third generation showed a heritability value of 0.34 and a direct response to selection for high growth of 9.3% per generation (Vu et al. 2019). Another selection programme has been carried out on striped catfish with salinity tolerance as main target (in study $N^0$ 2). Here, striped catfish were subjected (randomly selected groups) or not to 10 ppt during its life cycle. The results showed that the offspring of the 10 ppt exposed groups could tolerate high salinity better than the unexposed group at all developmental stages. In addition, growth performance was improved in the offspring of the groups selected for salinity. Although many species have been selected for improved performance, there is no research focusing on transgenerational inheritance to date. Therefore, this study aims to uncover DNA reprogramming in striped catfish during development through modification of the DNA methylation process, as well as the effects of salinity and salinity tolerance-based selection programme on genes involved in DNA methylation in its offspring.

## 2. Material and methods

### 2.1 Experimental animals

Striped catfish eggs were collected from three broodstock groups in the selection programme described in the previous chapter. Briefly, broodstock of F0 were selected from three different hatcheries located in An Giang, Vinh Long, and Can Tho provinces, with 10 males and 10 females from each hatchery. The first generation (F1) was produced from a full factorial cross of the broodstock, 30 females being

crossed with 30 males to create 900 full-sib families. Healthy larvae with no abnormal signs, uniform size, swimming actively and quickly responding to external stimuli were divided into three groups for culture until maturation. The first group was cultured in freshwater (Freshwater group, representing broodstock sources currently used in Mekong Delta). The second and third ones were both chronically exposed and cultured under 10 ppt. In particular, the second group, named Selection group, resulted from upwardly selection of 50% of fast-growing fish at three stages, *i.e* 100 g, 300 g and 1000 g, while the third one was undergone in a random selection process (Random group). These fish matured and were able to be reproduced after around 3 years in the corresponding conditions. A new generation (F2) was obtained from random couples from the three F1 groups (*i.e.* from the selected group (10♂ X 10♀), the random group (10♂ X 10♀), and the freshwater group (8♂ X 8♀) (Figure 1)). F2 embryos were collected to conduct the current experiment.

A pre-experiment estimated that the conditions of egg incubation up to 2.5 ppt did not affect significantly the fertilization rate, embryonic development, hatching rate and early morphological development of larvae in comparison with freshwater conditions, except the longer hatching time of the freshwater group (Chapter 4). Therefore, a water salinity of 2.5 ppt was compared with freshwater in the experiment (Figure 1). Particularly, a dry fertilization process was used, where eggs and milt were gently mixed. The fertilised eggs from three groups were immediately incubated into the freshwater and 2.5 ppt conditions. Embryonic development was monitored under a binocular microscope. After hatching, the larvae were transferred into a 1 $m^3$ composite tank with the corresponding salinity level to nurse until 6-day post-hatching (dph). Samples were collected at the following stages: fertilisation, early blastula, late blastula, early gastrula, late gastrula, somite development, hatching, 1 dph, 3 dph and 6 dph. Pooled samples were transferred to a cryotube containing a solution to preserve RNA that was prepared following the protocol detailed at https://www.protocols.io/view/RNAlater-Recipe-c56y9d). The tubes were stored at −80°C until further analysis.

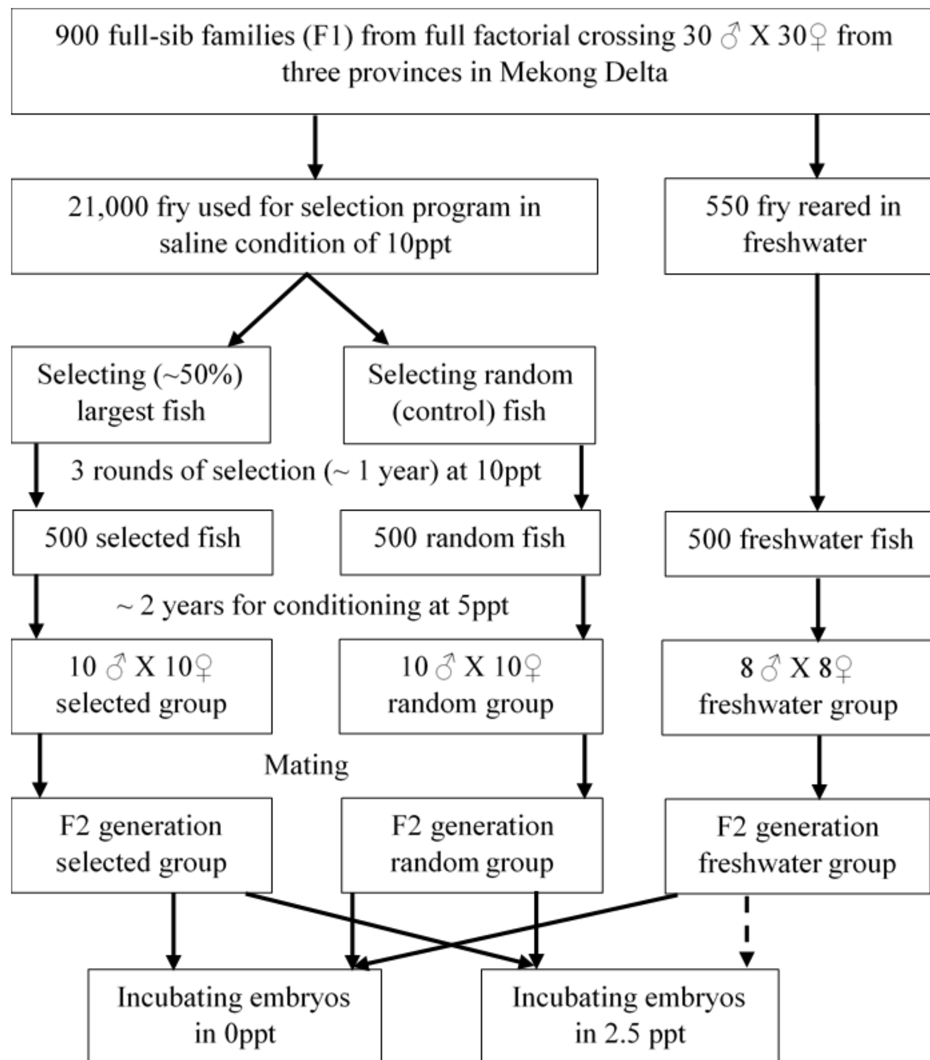


**Figure 1**: Schematic representation of mating design for obtaining the experimental fish. Dash arrow indicate that the all larvae in the treatment were died in 6 dph

**2.2 Quantification to overall DNA methylation**

The global DNA methylation level was analysed using a LUminometric Methylation Assay (LUMA) (Fellous et al. 2018). Genomic DNA from pooled embryos and larvae were extracted using affinity chromatography (Machery-Nagel 740901, Germany) according to the protocol of manufacturer. DNA concentration and quality were determined with a NanoDrop-2000 spectrophotometer (Thermo Scientific,

Waltham, MA, USA) and after migration on agarose gel (1.5%), respectively. Samples with degraded DNA were rejected. LUMA assay was conducted to measure the global DNA methylation based on the original method (Pilsner et al. 2010; Head et al. 2014). The reaction comprised 20 µl, including 2 µl 10X Tango Buffer, 0.5 µl of Mun1, and 0.5 of either *HpaII* or *MspI* (input DNA) (all enzymes and buffer supplied by New England BioLabs) and 600 ng of DNA. A digestion step was done in the Thermocycler (Westburg) at 37°C for 4 h followed by a period of enzyme inactivation at 65°C for 20 min before cooling at 10°C. Next, 15 µl of annealing buffer (Qiagen) was added and 30 µl of the mixture was loaded into a pyrosequencing plate. The plate was loaded into a Pyromark Q24 (Qiagen) and the dispensation order for nucleotides was GTGTCACATGTGTG. The level of global DNA methylation was calculated following the formula: % of Methylation = 100*(1-((*HpaII*(G)/*MunI*(T))/*MspI*(G)/(*MunI*(T)))), where G represents the peak height for *HpaII* (cuts the recognition site when unmethylated) or MspI (insensitive to methylation status of the recognition site) and T represents the peak height for MunI (input DNA). Since the data are normalized to the *MunI* peak, the variable DNA input should in theory not influence the level of DNA methylation. To prevent underestimating methylation values, pyrograms with a quality control value of a signal-to-noise $<6$ (ratio between *HpaII* (G) peak height and background) and/or nonspecific peaks (evidence of DNA degradation) were removed (Head et al. 2014).

### 2.3 Gene expression

RNA of pooled embryos and larvae was extracted using a Tri-Reagent solution (Ambion, Thermofisher Scientific) according to the manufacturer's instructions. The RNA concentration and quality were determined by NanoDrop-2000 spectrophotometer (Thermo Scientific, Waltham, MA, USA) and after migration on agarose gel (1%), respectively. Genomic DNA contamination was removed by digestion using rDNAse I (Thermo Fischer Scientific, Waltham, MA, USA). Then, 2.5 µg of total RNA was reverse-transcribed using the RevertAid RT kit (Thermo Fischer Scientific) according to the manufacturer's instructions. The cDNA samples were diluted 1:30 and used to determine the expression of 11 genes such as DNA methyltransferases 1 (*dnmt1*), DNA methyltransferases 3aa (*dnmt3aa*), DNA methyltransferases 3ab (*dnmt3ab*), DNA methyltransferases 3ab3 (*dnmt3ab3*), DNA methyltransferases 3ba4 (*dnmt3ba4*), *mecp2*, Ten-Eleven Translocation (*tet3)*, $Na^+/K^+$-ATPase α1a (*nak1a*), solute carrier protein family protein 12 group A, member 1 (*slc12a1*), growth hormone (*gh1*) and insulin-like growth factor 1 (*igf1*). The reactions were performed using SsoAdvanced™ Universal SYBR® Green Supermix (Bio-Rad, Hercules, CA, USA) on QuantStudio 5 (Applied Biosystems, Waltham, MA, USA). The thermal conditions were set

at 95 ºC for 3 min in the initial denaturation step, followed by 40 cycles at 95 ºC for 30 s and 60 ºC for 30 s. The primers used were designed using Primer-BLAST (Ye et al. 2012) from the striped catfish sequences published on GenBank. The efficiency of each gene was tested before analysing. Details of the primers are given in Table 1. A standard curve of dilution series of the pooled cDNA of all samples was included to calculate the PCR efficiency and normalise the transcript levels. The relative levels of RNA were quantified for each gene based on the standard curve. Ratios of target genes/housekeeping gene were subsequently calculated for relative expression of target gene. Pre-tests were run with potential housekeeping genes such as beta-actin (*actb*), elongation factor factor 1 alpha (*ef1α*) and ribosome protein L27 (*rpl27*). The results showed the lowest variability in *actb* expression, which is consistent with the reports on zebrafish embryos and larvae (Casadei et al. 2011).

**Table 1. qPCR primers of 12 genes screened for epigenetic impact.**

| Gene | Sequence (5'->3') | Direction | Product length | GenBank Accession Number |
|---|---|---|---|---|
| *actb* | CCCAAACCTTAAGTTCAGCCA<br>CGACCCACAATGGATGGGAA | Forward<br>Reverse | 130 | XM_026929614.2 |
| *dnmt1* | TCTCCCAGCACATCAAAGGG<br>TTGCGCTTGGGTACTCTTGA | Forward<br>Reverse | 146 | XM_026929737.2 |
| *dnmt3aa* | GACGCACAGCCAAGTTCGAT<br>GCCAGACGACTCATGTTGGA | Forward<br>Reverse | 187 | XM_034301944.1 |
| *dnmt3ab* | TAATCCAGCACGGAAAGAGGG<br>ACCCATGGCCACTACGTTTT | Forward<br>Reverse | 130 | XM_034307930.1 |
| *dnmt3ab3* | GGTGATGCCATCCAACACGG<br>CTCACACTGTCAGTCATGCGAT | Forward<br>Reverse | 90 | XM_034307931.1 |
| *dnmt3ba4* | TGGAATTGCCACAGGCTACC<br>GAAGGTCAAACGGTCCCCAT | Forward<br>Reverse | 188 | XM_034306779.1 |
| *mecp2* | AGCGGCGATCCATCATTCG<br>GAAGGCTTTCCCTTCTGGGTTGA | Forward<br>Reverse | 149 | XM_026927959.2 |
| *tet3* | GCCCCTCAGGCATACCAAAA<br>AAGAGTGCAAACCACAGTGC | Forward<br>Reverse | 174 | XM_026931142.2 |
| *nka1a* | CCCAAGGGCAAGAAGGACAAG<br>ATTTGACCCATTCTGGGGTTGT | Forward<br>Reverse | 211 | XM_026931934.1 |
| *slc12a1* | ACGTCGTTTTCTGGTAACGGT<br>TCCTGCAGTGAGGGTCTGAT | Forward<br>Reverse | 167 | XM_026912120.1 |
| *igf1* | CCGGGGCTTTTATTTCAGCAA<br>TCGTGGCGCTTTACCAGATT | Forward<br>Reverse | 154 | NM_001200295.1 |
| *gh1* | CCAGCCTGGATGAGAACG<br>GGGATCTCCTGCACTTGG | Forward<br>Reverse | 162 | M63713 |

### 2.4 Statistical analyses

Statistical analyses were performed in the SPSS software version 20.0. Regarding the LUMA and gene expression data during development in freshwater, an ANOVA1, followed by a Tukey's *post hoc* test was applied. ANOVA2 and Tukey *post hoc* test were performed for data at 6 dph. The normality and the homogeneity of variance were tested and the data was transformed by log10 when these assumptions were not met. All tests were set up at a confidence level of 95% ($p<0.05$).

## 3. Results

### 3.1 Methylation of CpG genomic DNA

We observed a significant increase in the level of global DNA methylation during early development in striped catfish (Figure 2) ($p< 0.001$). The level of methylation was significantly lower in the fertilised embryos (40.73% ±8.83) followed by a rapid rise in the early blastula stage (54.69% ± 1.87%). The level of methylation continuously increased during development up to the late gastrula stage (64.30%±4.83), which is significantly different from that in early blastula, before slightly dropping at somite developmental stage (60.44%±2.35). DNA methylation was maintained stable in the remaining experimental stages, from hatching to 6 dph, ranging from 67.98 ± 3.63 % to 73.96 ± 1.04 %.

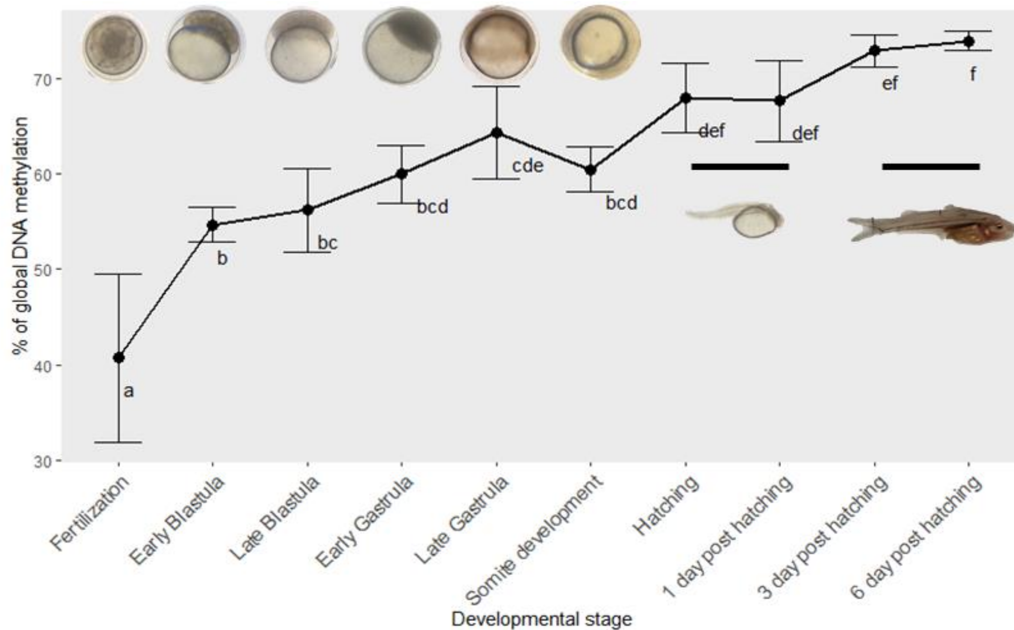


**Figure 2:** Global DNA methylation level during the development of striped catfish. Results are given for a quantity of 600 ng of genomic DNA using the LUminometric Methylation Assay (LUMA). Results are given as mean ± SD (n=6). Different lowercase letters denote statistically significant differences between developmental stages ($p < 0.05$).

### 3.2 Expressions of *dnmt* genes

Among *dnmt* genes, a significant decrease was observed in the relative expression of *dnmt1* from the fertilisation (13.943 ± 1.228) to early gastrula (0.597 ± 0.312), (Fig 3a) ($p<0.001$). Later on the level of mRNA transcript remained constantly low until the end of the experiment. Conversely, the accumulation of *dnmt3ba4* mRNA showed a rapid increase during the same period, from 0.003 ± 0.0008 to 5.174 ± 0.728, followed by a gradual decrease until 1 dph stage (0.135 ± 0.038) ($p<0.001$) (Fig 3b), and the expression remained then low at the subsequent stages. Regarding *dnmt3aa*, its expression further increased from the stage of early gastrula (0.005 ± 0.003) to somite development (1.876 ± 0.082), then declined at 1 dph stage (0.654 ± 0.048) ($p<0.001$) (Figure 3c). Furthermore, its expression continued to increase at the stages of 3 and 6 dph. In terms of *dnmt3ab3* and *dnmt3ab*, their levels of mRNA transcript displayed similar increased patterns during early development ($p<0.001$). Particularly, significant increases appeared initially at the stage of early gastrula after which the levels remained constant until 1 dph. Subsequently, a second

significant rise was seen in *dnmt3ab3* at 6 dph while *dnmt3ab* was maximum at 3 dph, followed by a decrease at 6 dph (Fig 3 d,e).

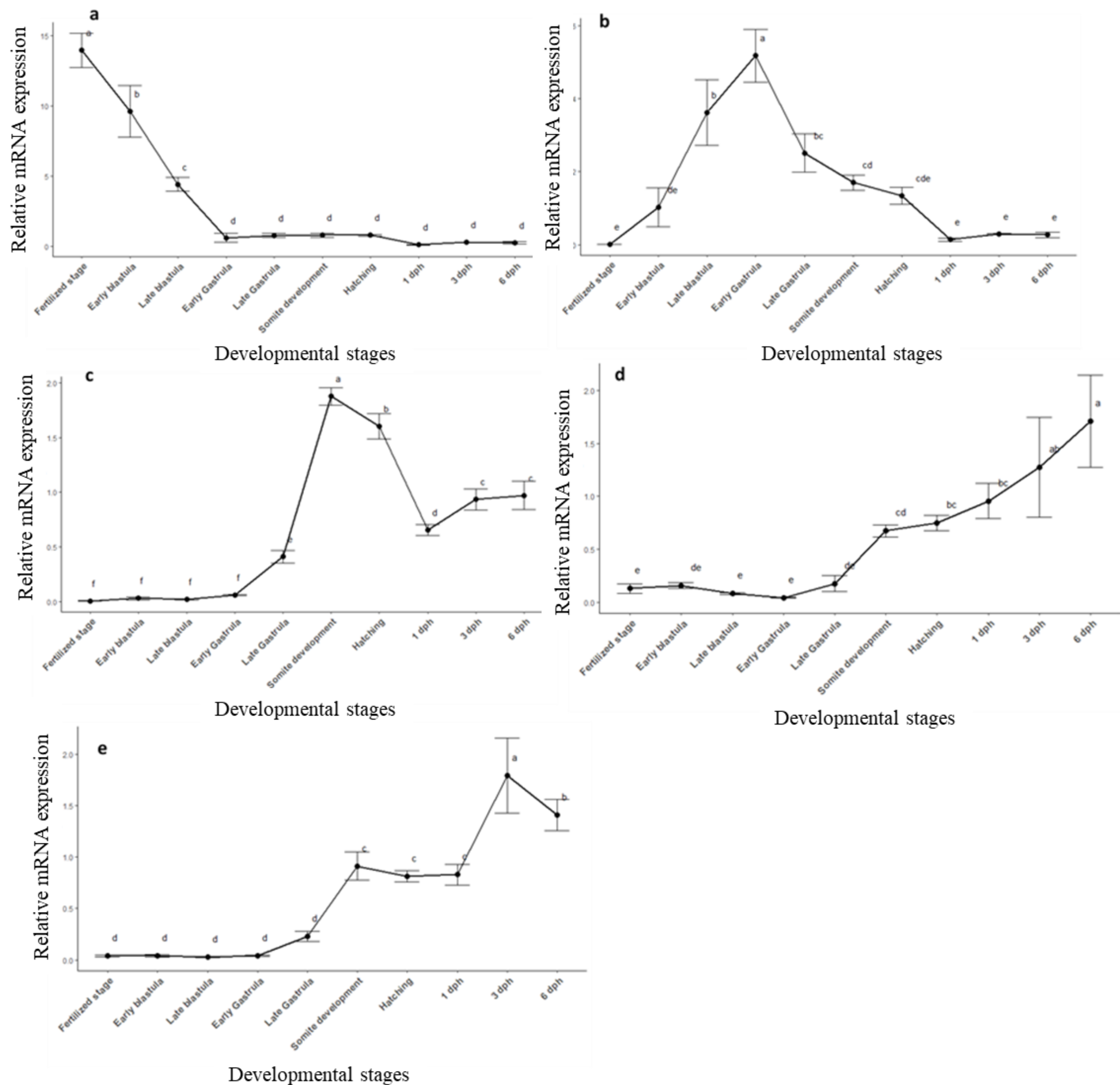


**Figure 3:** Relative expression of *dnmt* genes in striped catfish during early stages. (a) *dnmt1*, (b) *dnmt3ba4*, (c) *dnmt3aa*, (d) *dnmt3ab3,* (e) *dnmt3ab*. Results are given as mean ± SD (n=6). Different lowercase letters denote statistically significant differences between development stages ($p < 0.05$).

### 3.3 Expressions of *mecp2* and *tet3*

The transcript of *mecp2* decreased significantly, dropping about sixfold from fertilised to late blastula stages (p<0.001) (Fig. 4a), then its level was consistent up to the end of the experiment. The expression of the *tet3* gene increased rapidly from the late blastula (0.047 ± 0.006) to late gastrula stages (1.687 ± 0.330), followed by a gradual decline to the stage of 1 dph (0.726 ± 0.082) (p<0.001), which was maintained subsequently (Fig. 4b)

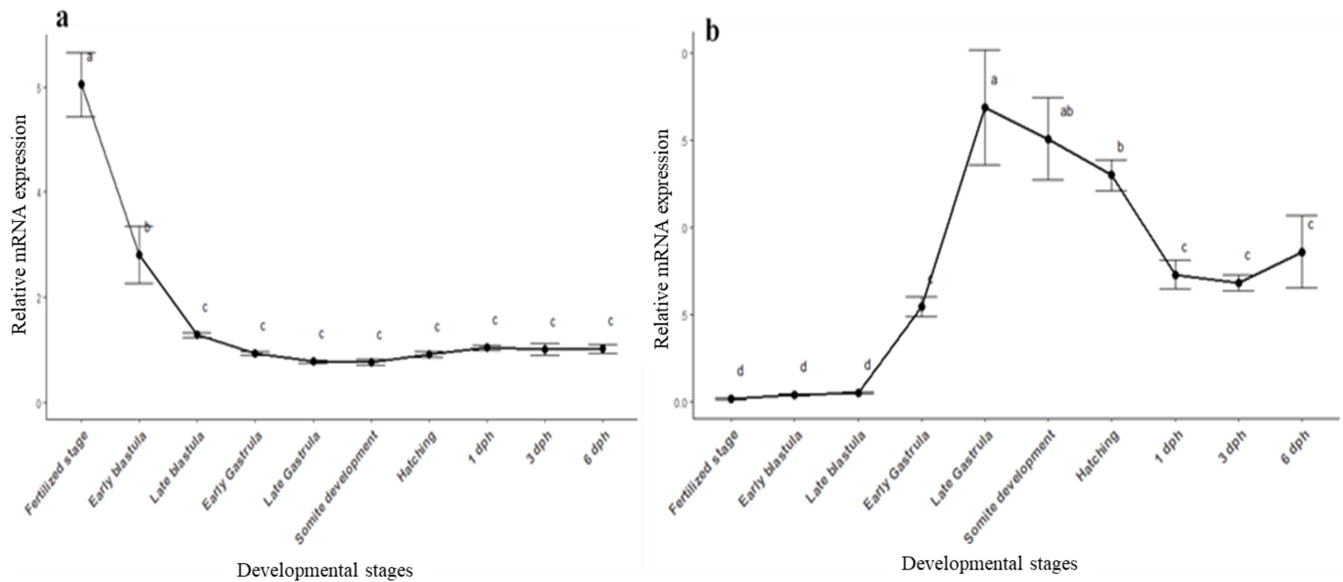


**Figure 4:** Relative expression of *mecp2* and *tet3* genes in striped catfish during early stages. (a) *mecp2,* (b) *tet3*. Results are given as mean ± SD (n=6). Different lowercase letters denote statistically significant differences between developmental stages ($p < 0.05$)

### 3.4 Expression of the *dnmt1* genes in 6 dph larvae

All larvae from the freshwater treatment and incubated at 2.5 ppt died at 6 dph, therefore only 5 treatments accounted for the analyses, *i.e* freshwater, random, selection incubated at 0 ppt and random and selection incubated at 2.5 ppt. Generally, salinity significantly affected the expression of *dnmt* genes ($p < 0.001$, except *dnmt3aa* with p =0.007). In particular, *dnmt1* expression was significantly enhanced in larvae incubated at 2.5 ppt during embryonic development compared to that in 0 ppt (Fig 5a). Additionally, this gene expression of freshwater group was significantly upregulated in comparison with the random groups incubated at 0 ppt condition. The level of *dnmt3aa* mRNA of selected larvae incubated at 2.5 ppt

was significantly higher than that in random one exposed to 0 ppt (Fig 5b). Remaining genes, *i.e, dnmt3ab*, *dnmt3ab3* and *dnmt3ba4* showed the same pattern of upregulation in fish submitted to 2.5 ppt in comparison with 0 ppt group (Fig 5c,d,e).

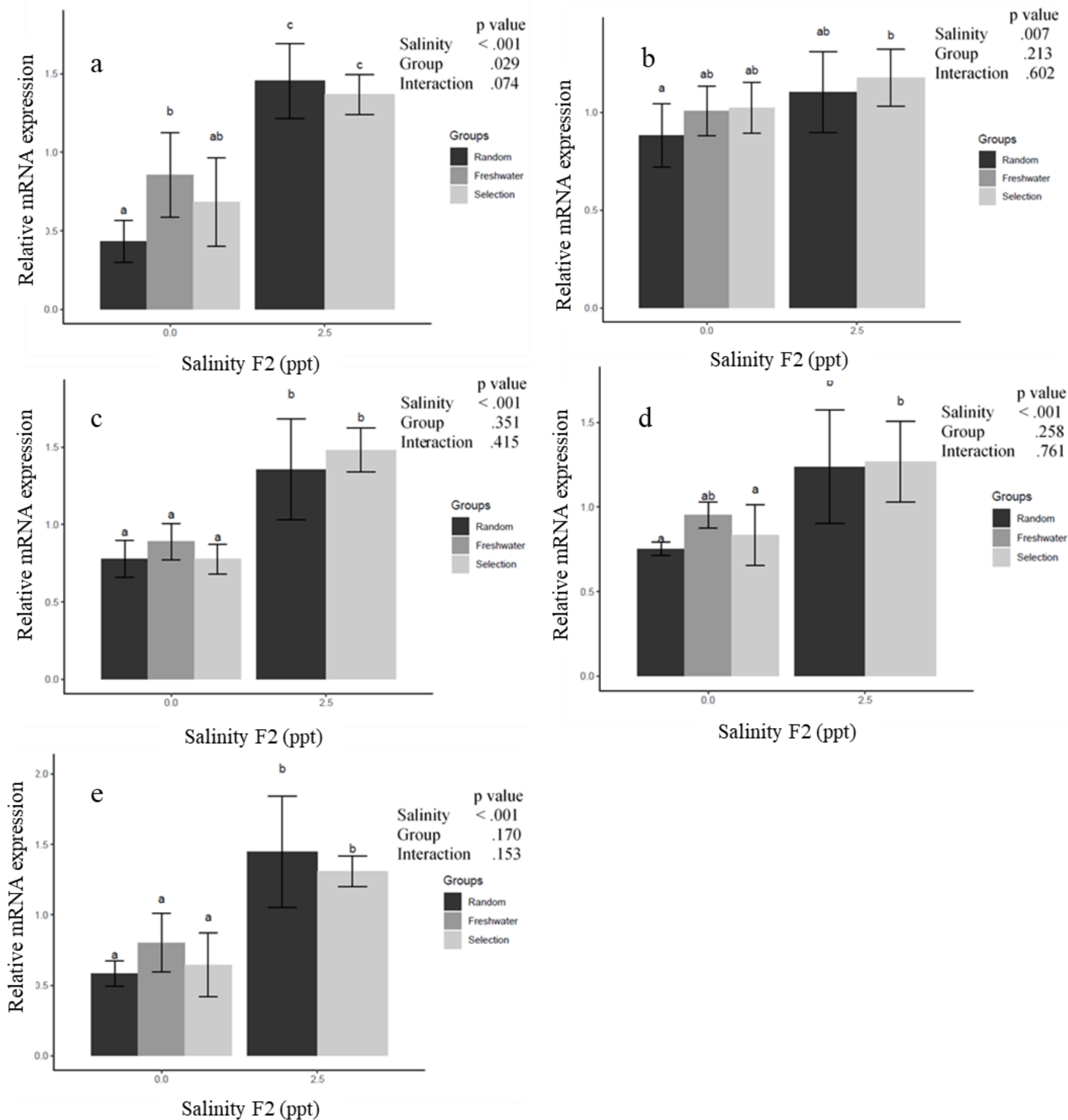


**Figure 5**: Relative expression of *dnmt* genes in striped catfish groups at 6 dph under different salinities. (a) *dnmt1*, (b) *dnmt 3aa*, (c) *dnmt3ab*, (d) *dnmt3ab3*, (e) *dnmt3ba4*. Results are given as mean ± SD (n=6). Different lowercase letters denote statistically significant differences between treatments analysed by Tukey post hoc test (p < 0.05).

### 3.5 Expression of the *mecp2* and *tet3* genes in 6 dph larvae

Salinity did not affect the levels of *mecp2* and *tet3* transcripts (Fig 6a, b). The only significant difference was observed between the expressions of *mecp2* in freshwater and random larvae incubated in 0 ppt (Fig 8a) ($p<0.05$).

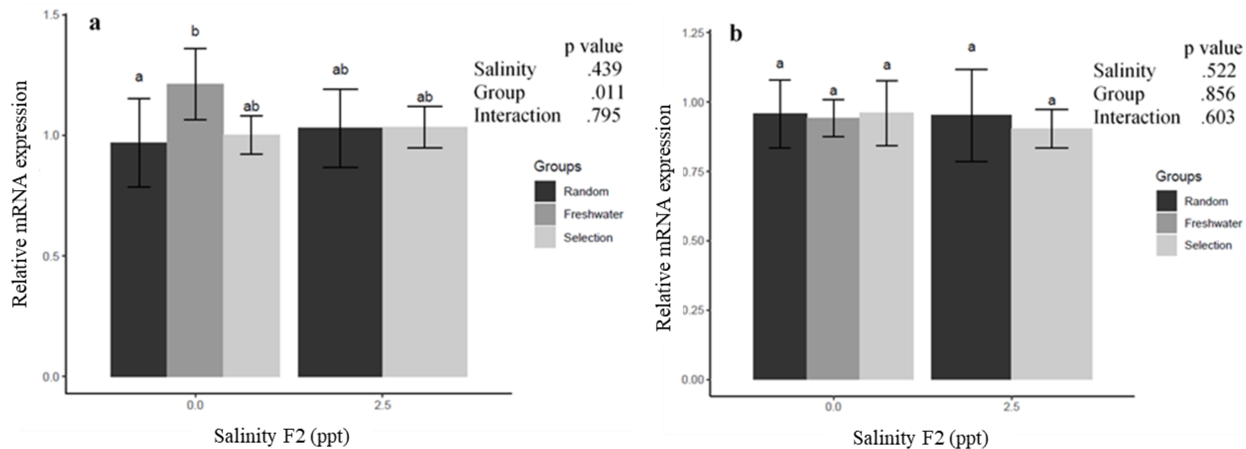


**Figure 6:** Relative gene expression in striped catfish groups at 6 dph under different salinities. (a) *mecp2*, (b) *tet3*. Results are given as mean ± SD (n=6). Different lowercase letters denote statistically significant differences between treatments analysed by Tukey post hoc test ($p < 0.05$).

### 3.6 Expression of genes related to osmoregulation (*nka1a* and *slc12a1*) and growth performance (*gh1* and *igf1*) in 6 dph larvae

Two-way Anova showed a significant effect of salinity on the transcript levels of *nka1a* ($p<0.05$) (Fig 7a) despite of no significant between all treatments. Conversely, *slc12a1* expressions were remarkably different between groups ($p<0.05$) with the higher level of mRNA in the random group compared to the other groups (Fig 7b). Both individual group and interaction of salinity and group significantly affected on *gh1* expression ($p<0.01$) (Fig 7c). Furthermore, the salinity of 2.5 ppt significantly increased the mRNA levels in random fish compared to selection ones. Besides, the individual salinity significantly upregulated the expressions of *igf1* (Fig 7d) ($p<0.001$).

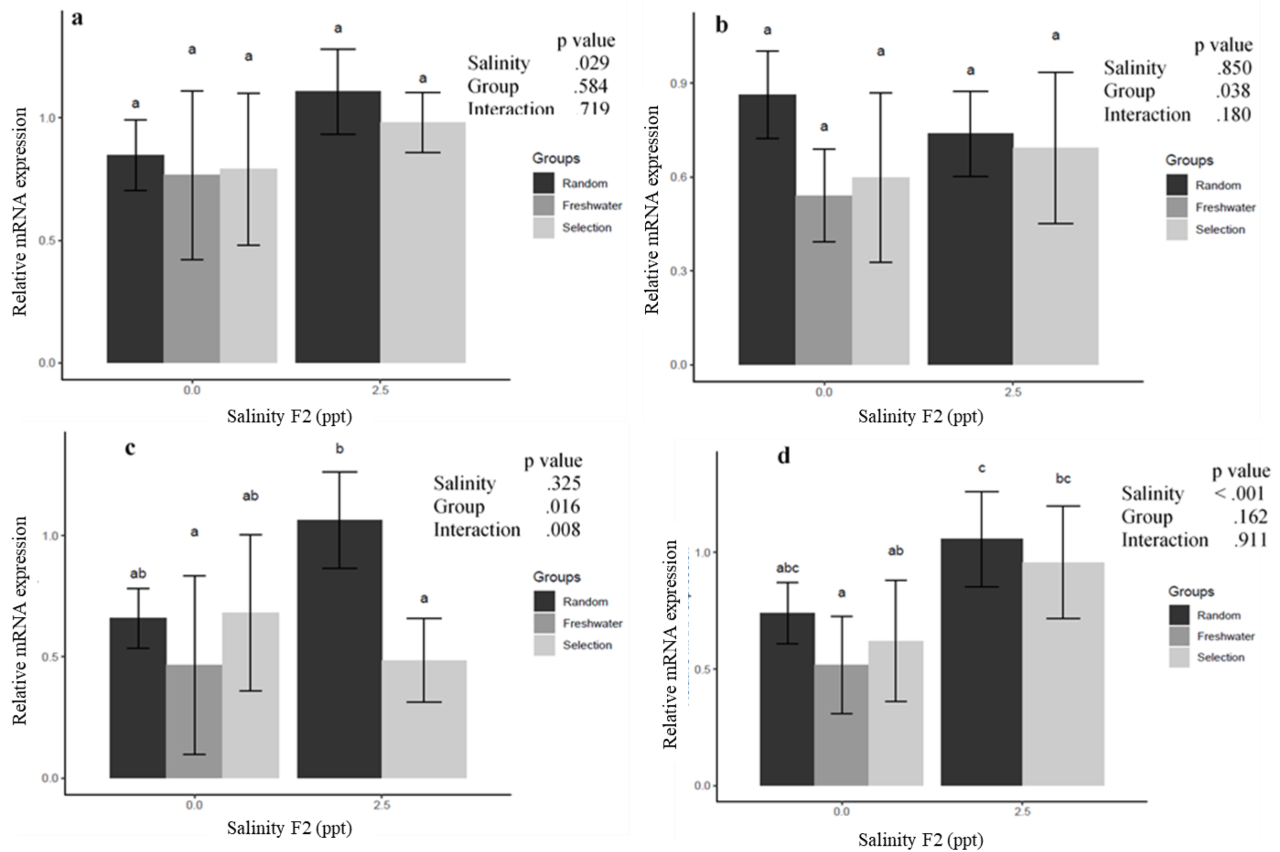


Figure 7: Relative expression of genes related to osmoregulation and growth performance in striped catfish groups at 6 dph under different salinities. (a) *nka1a*, (b) *slc12a1*, (c) *gh1* and (d) *igf1*. Results are given as mean ± SD (n=6). Different lowercase letters denote statistically significant differences between treatments analysed by Tukey post hoc test ($p < 0.05$).

## 4. Discussion

Three groups of the second generation of striped catfish embryos and larvae were obtained from breeders and submitted to a selection programme based on salinity stress. We explored DNA reprogramming as well as the effects induced by salinity on the methylation process by measuring global CpG methylation levels and marker genes expression.

### 4.1 DNA reprogramming during early development

We are the first team investigating the DNA reprogramming during early development of striped catfish. During early development, the overall CpG methylation of genomic DNA is dynamic, due to the significant increase in methylation after fertilization until the early blastula stages, followed by another increase in methylation until the gastrula stage. Interestingly, the timing of DNA remethylation in striped catfish was highly consistent with that of zebrafish and medaka, where genomic DNA undergoes the greatest remethylation from the 1-2 cell stage to the blastula stage, followed by less increase at the gastrula stage (Mhanni and McGowan 2004; Wang and Bhandari 2019). It is also similar to the process in mice, where the *de novo* methylation process occurs slightly earlier, at the morula stage (Santos et al. 2002). These results suggest that the timing of DNA reprogramming is rather conserved between many species, with exceptions such as in self-fertilised fish, *Kryptolebias marmoratus,* where the process takes place in later stages (Fellous et al. 2018). Global DNA methylation in zebrafish is 70-95 %, highest in sperm (91-95 %) and lowest in oocytes (75-80 %) (Jiang et al. 2013; Potok et al. 2013). Similarly, the paternal medaka genome is highly hypermethylated while the maternal genome is hypomethylated, at 83.03 % and 26.59 %, respectively (Wang and Bhandari 2019). Particularly, methylation levels in the sperm genome are erased during the first cell division cycles and the embryonic genome remains hypomethylated from the zygote to the 16-cell stage before *de novo* remethylation started in early morula stage (Wang and Bhandari 2019). The findings require a further investigation to estimate whether or not DNA methylation in paternal and maternal gametes is transmitted in striped catfish. Some mammals and fish have the level of methylation in late blastula stages approximately equal to sperm methylation (Jiang et al. 2013; Wang et al. 2014; Wang and Bhandari 2019), to some extent, it can be indicated that paternal inheritance undergoes methylation suppression during early development in striped catfish. However, the exact mechanism need to be clarify.

In vertebrates, DNA methyltransferase catalyzes methylation processes, which occur primarily at CpG sites (Wang and Bhandari 2019). There are two types of DNA methyltransferases in mammals that are involved in the regulation of DNA methylation, namely the DNMT1 and DNMT3 families (Bestor 2000). In the current study, the relative expression of *dnmt1* exhibited a significant decrease in expression in the early gastrula, which was strongly contrasted to the DNA remethylation process observed during embryogenesis. DNMT1 is the most important enzyme responsible for maintaining the DNA methylation pattern during replication (Branco et al. 2008; Geiman and Muegge 2010). Expression of *dnmt1* was relatively high during the demethylation process in zebrafish, presumably to maintain DNA methylation

status, consistent with rapid DNA replication and short cell cycles (Wang and Bhandari 2019). Subsequently, a dramatic decrease in expression of the gene occurred in the early gastrula when the remethylation process was completed and the genome reached the hypermethylation state (Wang and Bhandari 2019). This is consistent with the results in striped catfish that showed a significantly lower level of *dnmt1* mRNA in gastrulation compared to that in the earlier stage. These results indicates that the need to maintain rapid de novo methylation is terminated in this stage with the stability of methylation pattern (Wang and Bhandari 2019). In addition, expression of *dnmt3ba*, *dnmt3aa* and *dnmt3ab/ab3* genes was detected and altered during early development. The first interesting observation was that *dnmt3ba4* accumulations increased dramatically, coinciding with the pattern of global DNA methylation levels up to the early gastrula stage, indicating an important role for DNMT3BA in *de novo* methylation in striped catfish. DNMT3BB.1 and DNMT3BA shared the function of *de novo* methylation that significantly increased with the remethylation process (Okano et al. 1999). Additionally, *dnmt3aa* and *dnmt3ab/ab3* expression in striped catfish were initiated when *dnmt3ba* expression was terminated, showing that the contributions of *dnmt3aa* and *dnmt3ab/ab3* to *de novo* methylation. However, because of the significant increases in methylation marks after late gastrula, this suggests that *dnmt3a* potentially regulate other functions in striped catfish. Homozygous knockout of *dnmt3a* in mice resulted in imprinting defects, stunted growth and premature, postnatal death (Okano et al. 1999). Furthermore, acute deletion of *dnmt3a* and *dnmt3b* in mouse embryonic somatic cells did not affect global methylation marks but prevented *de novo* methylation of newly integrated proviruses (Okano et al. 1999). Intriguingly, *dnmt3a/b* knockout caused progressive depletion of methylation marks with continued culture of these cells (Chen et al. 2003). These results suggest that while *dnmt1* is primarily responsible for maintaining methylation, *dnmt3a* and *dnmt3b* are also required for the stability and faithfulness of this process, besides *de novo* methylation function (Zeng and Chen 2019).

Although the level of 5hmC was not estimated in striped catfish, we revealed the presence of *tet3* during early development, possibly demonstrating that active demethylation plays an important role during this period. The mRNA transcripts of *tet3* had low levels from fertilization to the late blastula stages, when the remethylation process occurred primarily, followed by a peak at the end of gastrulation. TET3 is an important component of active demethylation and its absence results in a large loss of 5hmc and a simultaneous increase in 5mC (Gu et al. 2011; Jinsuk et al. 2015). It will be interesting to study the regulation of *tet3* during cell division cycles. The significant increase in the gene from the end of blastulation to the end of gastrulation could contribute to the end of the demethylation process by converting

some of the 5mC to 5hmC. In medaka, a similar pattern of *tet3* increase was reported at the end of the demethylation process, followed by stable methylation marks during gastrulation (Wang and Bhandari 2019). This finding may support the hypothesis of *tet3* function on the regression of the remethylation process, but further studies are needed to clarify it. While the functions of DNA methylation and DNMT proteins have been well studied during mammalian development (Dean et al. 2003; Swales and Spears 2005), the mechanisms of methyl-CpG-binding domain proteins, generally thought to govern normal embryogenesis, remain largely unknown (Bogdanović and Veenstra 2009). DNA methylation can suppress transcription through the recruitment of MeCP2 to methylated CpG dinucleotides (Fuks et al. 2003). It is also well known that MeCP2 silences the transcription process by recruiting the repressive histone deacetylase machinery, which removes acetyl groups from histones, resulting in gene silencing (Jones et al. 1998; Nan et al. 1998). Expression of *mecp2* was the highest at the fertilization stage, followed by a dramatic decrease until the blastula stage in striped catfish. A possibly complicated mechanism has been suggested that MeCP2 deficiency enhances cell reprogramming by stimulating IGF1/AKT/mTOR signaling and activating ribosomal protein-mediated cell cycle gene translation at the onset of reprogramming (Zhang et al. 2018). Interestingly, the expression of *dnmt1* and *mecp2* showed a strong similar pattern during early development in striped catfish. This could be explained by the fact that MeCP2 is a target of DNMT1 and the two work together for the maintenance of methylation *in vivo* (Kimura and Shiota 2003).

**4.2 Alterations of expressions of genes related to methylation, osmoregulatory and growth processes in saline conditions.**

The epigenome is critically important and globally sensitive to environmental factors (Marsit 2015; Legault et al. 2018; Norouzitallab et al. 2019), especially during early embryonic development due to its epigenetic regulation associated with correct cell fate determination (Morey et al. 2015; Ohbo and Tomizawa 2015; Vougiouklakis et al. 2017). In our study, salinity exposure induced increase expression of all *dnmt* genes, i.e. *dnmt1*, *dnmt3aa*, *dnmt3ab*, *dnmt3ab3*, and *dnmt3ba4* in striped catfish larvae at 6dph stage. This is consistent with results in the mandarin fish *Siniperca chuatsi* (3 months) , which showed significantly higher levels of *dnmt1* and *dnmt3a/b* mRNA in the gills and gut of fish subjected to 10 ppt compared to those in freshwater after 7 days of exposure (Zhou et al. 2021). Other abiotic environmental stressors, cadmium and temperature, also induced increased expression of *dnmt1a* and *dnmt3b* genes in zebrafish during embryonic development (Zhou et al. 2021). Overall, this indicates that exposure of striped catfish to 2.5 ppt salinity during embryo can trigger transcriptional changes in DNA methylation-related

genes. Diadromous fish must regulate the gene expression pathways to adapt to changes in the osmotic environment when they enter seawater or vice versa. Therefore, salinity changes can impact the epigenome of aquatic vertebrates that may be a significant marker during the migrations (Hammond et al. 2015). Furthermore, an increasing body of research indicates that DNA methylation patterns are much more dynamic and context-dependent than originally thought (Baubec and Schübeler 2014; Ambrosi et al. 2017). Classically, the high level of methylation in CpG-rich promoters has been involved with gene silencing through the blockage of the transcription initiation machinery (Bird 2002; Edwards et al. 2017). However, hypermethylation in promoter regions are also indicated for gene activation (Smith et al. 2020). Additionally, the modifications of DNA methylation occur in different regions such as within promoter, especially transcriptional factor binding site and enhancers (Bogdanović et al. 2016; Li et al. 2017), exons (Moghadam et al. 2017; Uren Webster et al. 2018) and introns regions (Anastasiadi et al. 2018). These findings indicate that methylation changes due to salinity impact may result various consequences that need to be clarify using modern technology in the further studies.

Regarding osmoregulatory genes, two-way ANOVA analysis showed a significant effect in *nka1a* expressions under salinity incubation. *Nka1a* and *slc12a1* encoding $Na^+/K^+$ α-subunit of Na+/K+-ATPase and $Na^+$-$K^+$-$2Cl^-$ 2 are major components in osmoregulatory process related to ions exchanges (Grosell 2006). . Although it is suggested that true osmoregulation in teleost embryos begins in gastrulation and is functional through the closure of the yolk plug and this regulation appears to be correlated with the development of chloride cells in the yolk sac integument (Alderdice 1988), the tolerated capacity is depended on organogenesis development during the early stages (Varsamos et al. 2005). Furthermore, previous research has only estimated salinity tolerance in striped catfish up to the hatching stage (Hossain et al. 2021), but no reports have been made on the subsequent development of larvae exposed to salinity at the early embryonic stage. The present study suggests that striped catfish eggs can hatch at 2.5 ppt, but that its osmoregulatory capacity is not functional enough for subsequent development. However, additional research is needed to determine if osmoregulatory capacity exists in the early stages of the species.

In addition to osmoregulatory capacity, the current study examined the growth performance of the species under salinity exposure through the genetic expressions of *gh1* and *igf1*. It is evident that osmotic conditions affect the growth ability of fish (Bœuf and Payan 2001). The growth hormone/insulin-like growth factor or hypothalamus-pituitary growth axis plays a major role in the regulation of growth in aquatic animals (Fox et al. 2010; Fu et al. 2019). Exposure to 2.5 ppt salinity during early stages increased

*igf1* mRNA transcripts after 6 dph, while this change did not affect *gh1* expression. An increasing *igf1* expression under hyperosmotic environment was also found in different species. In particular, Nile tilapia and Mozambique tilapia (*Oreochromis mossambicus*) have been reported to have *igf1* expressions upregulated during saltwater acclimation (Magdeldin et al. 2007; El-Leithy et al. 2019). Despite the non-significant difference observed in *gh1* expression by exposure to 2.5 ppt, the gene expression was significantly affected by the combined effect of selection programme and incubated salinity and the slight increase was shown in fish subjected to 2.5 ppt compared to those in 0 ppt in random fish. This finding suggests that salinity increases *gh1* mRNA expression, but that the difference between parent interplays on its expressions. Besides, GH hormone and cortisol are primarily responsible for adaptation to seawater (Eckert et al. 2001; Yada et al. 2004). Intriguingly, the half smooth tongued sole (*Cynoglossus semilaevis*) showed lower levels of methylation in the putative *gh* promoter, concomitant with higher expression of *gh* mRNA in fish exposed to 30 ppt compared to those exposed to 15 ppt (Si et al. 2021). This may indicate that the increase in *gh1* mRNA under tolerated salinity through alterations in DNA methylation leads to a subsequent increase in *igf1* expression during early striped catfish development.

### 4.3 Effects of the selection programs on methylation, osmoregulatory and growth process in the offspring.

In addition to the effects of salinity, the present research examined the transgenerational inheritance through gene expressions during a selection programme based on salinity tolerance which was the first study conducted on teleosts to date. The study showed that only the first generation fish acclimated to salt water, *i.e.* the random and selection groups, were able to produce larvae capable of developing under salt stress conditions, while all larvae in the freshwater group died at 6 dph. The result was supported by study $N^0$ 2 which showed that the salinity-based selection could improve the hyperosmotic tolerance in the offspring compared to that of unselected fish. In general, in a 0 psu environment, most methylation-associated genes had the highest level in freshwater groups, followed by that of selection and the lowest in random fish, in which significant differences were observed in *dnmt1* and *mecp2*. Interestingly, inverse patterns were found in the expression of genes related to salinity tolerance and growth performance, with the lowest expression observed in the freshwater group. This may be explained by the fact that DNMT1 and MECP2 are major components in maintaining DNA methylation and CpG dinucleotide methylation through MeCP2 recruitment, respectively, and therefore both can suppress gene expression (Fuks et al. 2003; Geiman and Muegge 2010). In addition, it was found that the weight gain of fry was significantly

higher in the random and selection groups than in the freshwater group (in study $N^0$ 2). There may be a relationship between the breeding programme and the genetic expressions of the offspring, and then the growth performance. However, these findings in the current study are not sufficiently certain to conclude on the effectiveness of the selection programme. This hypothesis needs to be clarified in later generations. The main differences between the two groups are that the first generation in the selection group had a higher average weight gain than the random fish. There was an interaction effect of salinity and selection programme on *gh1* expression and osmotic conditions of 2.5 ppt significantly increased *gh1* expression in the random group while this increase did not occur in the selection group. A similar pattern was shown in *igf1* expressions. Although GH is the key factor and is strongly correlated with growth rates in fish (Moriyama et al. 1990, 1993; McLean and Donaldson 1993), GH and IGF-1 have been widely considered important for seawater adaptation, or hypoosmoregulatory function (McCormick 1995; Sakamoto and McCormick 2006). Indeed, environmental salinity induced osmoregulatory stress in fishes and GH increase was shown to be important for maintaining hypoosmoregulatory function (Deane and Woo 2009). In addition, considering that weight gains of random and selected fry did not differ at salinities below 10 ppt (in study $N^0$ 2), it may be suggested that *gh1* and *igf1* genes in random fish were increased to response to osmotic stress that did not occur in selected fish.

## 5. Conclusion

In conclusion, this study revealed that DNA reprogramming in striped catfish was dynamic with alterations in global CpG methylation levels through the regulations of *dnmt*, *tet3*, and *mecp2* genes. Methylation reprogramming occurred primarily in two stages: a dramatic increase from fertilization to early blastula, followed by a gradual increase to gastrulation stage. Only parents acclimated to saltwater, *i.e.* the random and selection groups, were able to produce offspring capable of developing under salt stress conditions up to 6 dph. Furthermore, salinity exposure during early development significantly increased the expression of methylated genes in 6 dph larvae. In addition, the offspring of the freshwater fish had the highest expressions of *dnmt1* and *mecp2*, which differed significantly from those of the random fish. In a 2.5 ppt environment, the increase in *gh1* and *igf1* expression was probably more associated with hypo-osmoregulatory function than with growth function. Future studies should focus on the process of DNA methylation in sperm and oocytes as well as in the stages of cell divisions to determine the exact reprogramming of DNA in the early stage of striped catfish. Besides, modern methods as reduced representation and/or whole genome bisulfite sequencings need to be approached to clarify the methylation

in specific genes and then functional changes under salinity effect. It is also recommended to continue to study transgenerational inheritance in future generations of selection programme.

# Chapter 7

# General discussion

# Chapter 7: General discussion

## 1. Harnessing genomics in breeding program to support sustainable development of striped catfish aquaculture in Mekong Delta, Vietnam

Selective breeding is a long-term genetic improvement procedure involving selecting and breeding individuals with traits of interest in a population. These traits include commercially important production traits (such as growth, feed efficiency, meat color), but also many other traits, provided that these traits have an at least partial genetic basis. Among these additional traits, immunity and disease resistance, , sex determination, stress tolerance, reproductive traits, color and pigmentation, (Dao et al., 2022 ; Wong et al., 2022) are of importance to improve the performances and reduce the environmental footprint of aquaculture (Houston et al., 2020). Recent advances in high throughput generation of 'omics' data coupled with the development of better computational approaches have allowed a better understanding of the biology and of the genetic architecture of complex traits. These progresses can also be used to accelerate the genetic improvements in aquaculture species (Nguyen et al., 2022 ; Hai et al., 2022). For example, incorporating genomic information on favourable alleles at major QTL (MAS) or integrating all markers in the prediction of breeding values (GS) in selection breeding schemes allows expediting the exploitation of untapped genetic potential (Houston et al., 2020; Wong et al., 2022). These techniques could provide significant contributions to the sustainable development of aquaculture in the future.

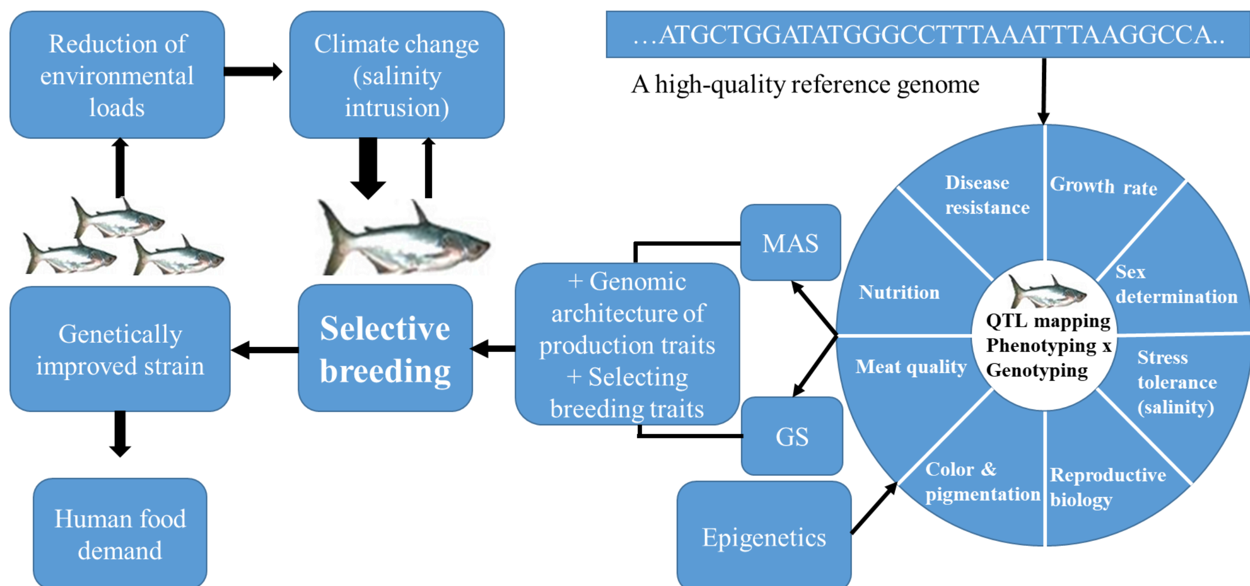


Figure 1: Harnessing genomics in breeding program to to sustainable development of striped catfish

### 1.1 Improving the quality of current seed production of striped catfish

The lack of a high-quality reference genome has so far hindered research concerning the genome, genetics, and breeding improvement of striped catfish, the latter being considered as the best approach to improve the quality of striped catfish seed (Dao et al., 2022; Nguyen et al., 2013). In this thesis, we have assembled a high-quality genome from HiFi sequencing data with remarkable improvement compared to two previous versions of the genome. The updated version may facilitate genome discovery research and enable more understanding of this complex biological system (Wong et al., 2022; Yue and Wang, 2017). The higher quality genome version may increase the production and improve product quality. Identification of targeted markers associated to commercially important QTL, and use of these QTL in MAS and/or in GS can increase the accuracy and efficiency of these breedin programs (Yue, 2014). For striped catfish, such approaches could target an improvement of salinity tolerance. Using high-quality reference genome can bring several benefits for downstream genetic applications such as reduction of time and funds needed to implement MAS and GS, achievement of higher genetic gains in the breeding programs (Benevenuto et al., 2019; Houston et al., 2020; Yue and Wang, 2017), and higher chances to uncover the molecular mechanisms underpinning the trait variation (Benevenuto et al., 2019). Indeed, using high-quality genome leads to a higher precision about the location, number, and gene action of QTLs, thereby improving our understanding of the genetic architecture of the traits through GWAS analyses (Appels et al., 2018; de los Campos et al., 2013). Altogether, it can be concluded that investing time and resources to obtain a high-quality reference genome is worthwhile given the benefits it confers to downstream genetic analyses and in the decision-making process for breeding programs (Appels et al., 2018; Benevenuto et al., 2019; de los Campos et al., 2013).

Geneticists struggle to get improved maps of the studied genomes because the improved accuracies of the maps lead to improved accuracies in the localization of the identified mutations. Recent improvements in the sequencing technologies include the HiFi long-read sequencing used in our study (Wenger et al., 2019b). This technology provides an improved access to more complex regions of repetitive DNA (Huddleston et al., 2017; Logsdon et al., 2020), which were notoriously difficult to sequence using preceding technologies. These regions sometimes contain regulatory elements that are important for traits of interest to the researchers (Mehrotra and Goyal, 2014). It is therefore of importance to have an as accurate as possible knowledge of these regions. In addition, using a low quality genome as reference genome for identifying DNA markers may generate aberrant DNA markers and lead to inaccurate linkage markers maps (Yue, 2014; Zenger et al., 2019). Using these maps leads to biased positions of the identified genes (or QTLs) in positional mapping. These inaccuracies might then lead to lack of power (due to the incorrect

relative positioning of the markers, a region of interest has not been detected) or to false positive results (a region of interest has been detected in an incorrect marker interval) in genome mapping. For these reasons, continuously improving the genomic map of the sequence remains important. In this thesis, the genome assembled from HiFi reads has a significantly increased quality, with a higher proportion of identified repetitive regions (demonstrated by the increase in the size of the assembled genome, of the contig N50 value, the proportion of repetitive regions and the BUSCO value). This improvement will lead to an increased number of DNA markers as well as a better quality of the marker maps. As a result, applications such as MAS and GS should have increased power and accuracy. Consequently, this new version will be a better tool for people working on the genetics of striped catfish.

Repetitive sequences, including transposable elements (TEs) and simple repeats, constitute up to 45% of the genome in mouse, 51% in golden snub-nosed monkey and 50 -70% in human (Biémont, 2010; de Koning et al., 2011; Wang et al., 2019). The repetitive sequences may influence host gene expression at both transcriptional and post-transcriptional levels through cis and trans mechanisms and participate in the regulation of diverse biological and pathological processes (Lu et al., 2020). For this reason, the repeat regions in a genome assembly may have important functional implications. In the present study, a higher number of repetitive elements (39.1%) were identified in our assembly compared to the previous genome assembly of Kim et al. (2018) and Gao et al. (2021) with 33.8% and 36.9%, respectively. The improved sequence will offer more opportunities to elucidate complex physiological processes regulated by genes located in repetitive sequences (Bailey et al., 2002; Dennis and Eichler, 2016; Lu et al., 2020). Moreover, it has been reported that many organism have the potential to generate new variation in response to stressful conditions through the modulation of epigenetic marks associated with TE regions (Dowen et al., 2012; Rey et al., 2016; Yu et al., 2013). Epigenetic components and TEs can mechanistically interact with each other to form a complex molecular network potentially permitting organisms to rapidly and lastingly cope with these drastic environmental changes (Rey et al., 2016). A high-quality genome with larger proportion of repeat regions identified can provide a better foundation that can improve the quality of basic research such as on epigenetic mechanism, DNA markers discovery, GWAS and GS, and access to novel genetic variation. These results are important primary steps to genetic improvement of striped catfish.

According to Taniguchi et al. (2003), genetic improvement of aquaculture species is classified into active and passive strategies, in which selective breeding is seen as an active strategy, whereas maintenance of genetic diversity and reduction of inbreeding depression remain the focus of passive strategies. Inbreeding depression resulting from breeding of closely related sibs and/or low number of breeders, may further impede the continuous reproduction of good quality aquaculture species by reducing their

performance and reproductive fitness traits (Shikano et al., 2001; Su et al., 1996; Khedkar et al., 2013). In the Mekong Delta, although the supply of striped catfish seed for farming is currently totally dependent on the operations of hatcheries, the genetic management of the broodstock is not routinely applied and has not been studied extensively (Ha et al., 2009; Tam et al., 2010). Some hatcheries do not have a medium to long-term broodstock management plan, and potential broodstock are kept without consideration for genetic management. In addition to the inefficient use of the genetic material, the lack of genetic management leads to risks associated with inbreeding, with negative effects on growth performance, an increase in the incidence of abnormalities, and a decrease in the survival rates (Tam et al., 2010; Pham et al., 2021). In fact, negative effects of inbreeding due to non-proper management of the broodstock resources has been reported in Pham et al., (2021). It is demonstrated to be one of the main reasons for the reduction of the survival rate of striped catfish in the grow-out stage, with survival rates ranging between 40 and 50%. Therefore, the pedigree information of broodstock in hatcheries, which is essential for a proper genetic management of the fish in hatcheries, is urgently required in order to sustain the high quality of seed needed for catfish aquaculture in the Mekong Delta.

However, recording pedigree information in striped catfish hatcheries is neither cheap nor easy. When using genomic markers for parentage assignment, the costs may involve fish marking (physical PIT tags), data management, and, most importantly, the production of genomic information (Raul et al., 2006; Vandeputte and Haffray, 2014). In my PhD thesis, a high quality reference genome for striped catfish of Mekong Delta has been assembed and is available for public use (Hai et al., 2022). We also provide a new method (*Shallowped*) and detailed guidelines for assigning parentage from sequencing data directly. The result from empirical data show that assignment rate from our method was 94% with mean cover depth being ~ 2x. This rate was higher than in the study of Trọng et al (2013) with the Cervus program used for parentage assignment on Nile tilapia (*Oreochromis niloticus*). With 12 microsatellites and 122 SNPs the mean assignment rates were 18% and 39%, respectively. In addition, our method has been demonstrated to be effective at an even lower depth coverage (lower ~1X) if more markers are used, which is not a problem when using SWGS. Moreover, the cost for sequencing is decreasing rapidly (Preston et al., 2021), especially in the case of low cover depth. Lower cover depth leads to lower expenses for sequencing, allowing more individuals to be sequenced for the same cost. Alternatively, sequencing could focus on most informative regions, with a higher cover in these regions but an equivalent sequencing effort and cost. Currently, the total cost for operation of one hatchery is 39,376 euro/ year (Pham et al., 2021) for an average of 2,700 broodstock. If a hatchery uses our program for parentage analysis, the total cost for material (sequencing, PIT tags, reader, expected cover depth being ~2X) is 3,479 euro/year when assuming that broodstock is

used during seven years. This additional cost of 8.8% seems affordable for hatcheries and should be largely counterbalanced by the potential benefits due to an improved genetic pool of fish.

Moreover, the sequence data for parentage assignment can also be used for other useful genomic analyses for selective breeding programs. For example, it has been reported that low depth sequencing can be a good strategy for GWAS (Li et al., 2011). Identifying regions of interest for a trait (such as quantitative trait loci (QTL)), and integrating genomic information on these regions in the selection programs (marker-assisted selection (MAS)) is a well-known strategy to improve the efficiency of the selection schemes (Yue and Wang, 2017). Similar conclusions can be drawn for GS studies; the optimal prediction accuracy can be obtained with low depth (around 1-2) sequencing of many individuals (Gorjanc et al., 2015). With the additional benefits from combining shallow sequencing data for multiple purposes, applying our new program *Shallowped* becomes a feasible approach for striped catfish hatcheries to obtain pedigree data that will be used for producing high quality seed resource for striped catfish industry in the Mekong delta, Vietnam.

### 1.2. The potential for catfish farming in saline intrusion areas

The expansion of catfish aquaculture in the Mekong Delta is dependent on tidal regime and salt water intrusion that would negatively affect the striped catfish culture industry (Nguyen et al., 2014). It has been considered that a water salinity higher than 4 ppt is not suitable for striped catfish culture (De Silva and Nguyen, 2011; Nguyen et al., 2014) (Figure 1). In the Mekong Delta, the risk from saltwater intrusion reduces the suitable area for catfish farming in the coastal provinces (Tien Giang, Ben Tre, Tra Vinh, and Soc Trang), from January to April. Consequently, most farmers do not stock fingerlings in these months. A new strain of striped catfish with better salinity tolerance generated from our breeding program can be expected to improve the growth performance of striped catfish cultured in these coastal provinces. In the current study, the targeted salinity of 10 ppt was set in a selective breeding program. Results from challenging tests with different salinity levels from G2 generations show that the salinity tolerance of striped catfish from selected groups reared in 15 ppt was similar to fish from 10 ppt (Figure 2 to Figure 4 in study N°2). These results suggest that the target salinity level for next-generation selection could be adjusted to 15 ppt to increase a genetic gain of salinity tolerance in breeding programs. In addition, the results (unpublished data, PANGAGEN project) from analyzing meat quality parameters including proximate composition, mineral content, fatty acids, fillet rate, and amino axit show that the salinity used in the selection program (after one generation of selection) did not impact on meat quality of striped catfish. Together, these results imply that with this new striped catfish strain (from the present generation or from

the next generations of selection), the farmers could stock the fingerlings in months with a high level of salinity without effects on meat quality. This would help fixing one of the main concerns of the processing sector, thus increasing the productivity of cultured striped catfish.

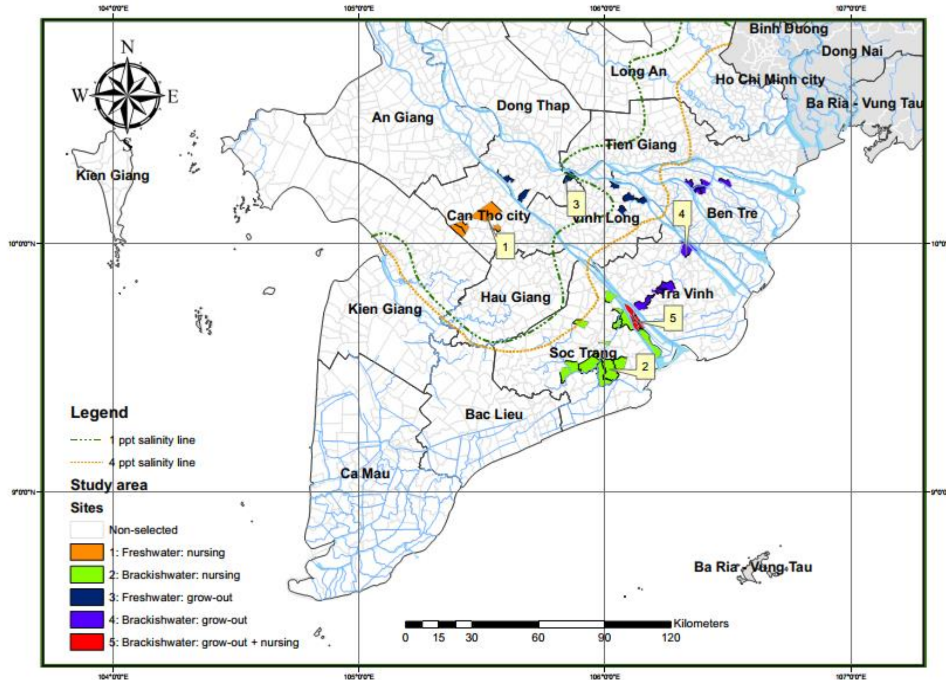


Figure 2: Salinity intrusion in Mekong Delta in 2020 (Son et al., 2022, PANGAGEN)

According to Gjedrem et al. (2012), a substantial benefit of breeding programs stems from reduced production cost per unit. Genetically improved fish grow faster, and are more efficient feed converters. Therefore, overall aquaculture production will increase faster as use of improved stock increases, and finally reduce the cost for water and feed requirement. As the growth rate of fish increases, the time to market size becomes shorter. In the present study, the new salinity-tolerant fish proved to grow and survive better in saline conditions than non-selected animals. The survival and growth rate of the selected strain were significantly higher than those from normal fish when they were cultured in 10 ppt during 8 months (Table 1 from Chapter 4). The results from survey on farmers of China, Vietnam and Malaysia showed that the willingness to pay a price premium for improved seed with respect to growth and survival was in the order of 20% (Olesen et al., 2015). In a study of Nguyen et al. (2016), the extra cost for using improved salinity-tolerant strain of striped catfish is slightly less than 0.4 % of the present production cost. These results suggest that a new salinity-tolerant fish produced from our program is likely to be accepted by farmers. Taken together, these results suggest that our new striped catfish strain could be an excellent choice

for culturing in salinity-instrusion areas. They will help the catfish farmers to reduce the risks due to increasing-salinity period, reach market size in shorter time, improve growth performance and survival rate, and then reduce the production costs.

### 1.3 Reduction of environmental loads in striped catfish farming

One of the key aspects of sustainable development for aquaculture is to minimize environmental impact (Lundh, 2020; Olesen et al., 2011). According to Sae-Lim et al. (2017), although aquaculture has no direct impact on climate change because aquatic species do not emit greenhouse gasses (GHG) as do ruminants, aquaculture activities, such as logistics, input power, and feed production indirectly affect climate change. Several studies confirmed that reduction in feed production and reduced amount of feed needed for producing each kg of fish can reduce global warming potential (d'Orbcastel et al., 2009 ; Henriksson et al., 2015a, 2015b). Moreover, the major ingredients for protein and energy resources in fish feed for farmed fish and shrimp species are fish meal and fish oil that are mainly from the wild fish stocks (Hasan and Shipton, 2021; Lundh, 2020). The increasing demand for aquafeeds and the use of natural resources may contribute to overfishing and influence natural ecosystems. Selective breeding for increased production, such as growth, have also been shown to have improved feed conversion, through correlated changes in feed efficiency; therefore, selection for increased growth rate can be used to improve feed efficiency (Gjedrem et al., 2012). In the study of Thodesen et al. (1999b), the feed conversion ratio (FCR) of Atlantic salmon was improved by 20% during five generations of selection. In intensive striped catfish aquaculture production, the highest production cost was for feed, which accounted for about 81% of the total production cost (Hasan and Shipton, 2021). In 2014, commercial feed production was estimated at 1,830,075 tons, with a protein level ranging from 35 to 26%, for whole demand of striped catfish farming in the Mekong Delta (Hasan and Shipton, 2021). In our study, the FCR of selected group was slightly improved compared to the normal group for both freshwater and saline conditions (Figure 4 in Chapter 4). These results imply that if a new striped catfish strain produced from the present selection program were cultured in the Mekong Delta (in both freshwater or/and saline regions), a large amount of feed needed for farming will be saved. This phenomenon will contribute to minimizing the impact of striped catfish farming on climate change, for example, to reduce greenhouse gas emissions through a reduction in environmental loads (Figure 1)

At the present time, the majority of striped catfish broodstock resources for hatcheries in the Mekong Delta are still based on wild stocks, or farmed animals from breeding programs without proper selection and/or control of inbreeding (Pham et al., 2021), that are usually poorly adapted to life in captivity

comparing to genetically improved or well-managed stocks. This situation implies poor growth rates and animal welfare, high mortality, inefficient use of resources such as feeds, water, and energy, and higher cost per kg of fish produced. Consequently, a higher carbon footprint with a negative impact on climate change per kg fish produced is expected. In the present study, the selection program not only generated faster growth fish with higher feed utilization but also enhanced the survival rate, especially in the commercial stage (Tabble 1 from Chapter 4). The higher survival rate means that the new strain has more robustness and adaptation to variation in environmental factors. These attributes of the new striped catfish strain may minimize the use of chemicals/drugs such as antibiotics, disinfection, vitamin, and vaccines that can affect the quality of the fish and the natural environment (Lundh, 2020). Further, as the growth rate increase due to the selection program, the time to culturing striped catfish for market size becomes shorter. The results from culturing after 8 months show that to get the market size (around 1 kg/per fish (Phan et al., 2009)), the selected group only need around 7 months while normal fish required over 8 months (reduction of around 14.2% in time to harvest) (Figure 3A). This reduction will save a lot of water and energy for striped catfish farming. In the Mekong Delta, the average cost for energy is 1,409 USD, most of them used for pumping to exchange water in grow-out ponds (Nguyen et al., 2018). Together these results suggest that selective breeding is a long-term, cost-effective strategy that can best minimize the detrimental effects of climate change (in our case, salinity increase) on striped catfish aquaculture and also have a reverse effect on reducing the impact of striped catfish aquaculture on climate change.

## 2. Potential genetic adaptation of striped catfish to salinity

In this thesis, individuals from the selected and random groups, with their parents living in saline conditions, showed an increased adaptation in all salinity treatments, especially with high levels of salinity (in Chapter 4). And, the offspring originating from parents selected on size “selected group” tend to perform better in salinity tolerance than those of parents from randomly chosen “random group”. During the embryonic stage, eggs from the selected and random groups had significantly greater gastrula and hatching rates than those collected from the freshwater group, with their parents without historical salinity exposure, when treated with a salinity stress of 10 ppt (Figure 2 of Chapter 4). In larvae up to the fry stage, only fish from the selected and random groups survived when the salinity reached 15 ppt. In the comparison of fish at fingerling stage exposed to salinity of 20 ppt, the survival rate in the freshwater group was significantly lower than for fish from the selected and random groups (Figure 3 & 4 in Chapter 4). The same trend was found at adult stage when fish were cultured 8 months under salinity of 10 ppt (Table 1 in Chapter 4). The descendants of fish from the freshwater group displayed significantly increased cumulative mortality relative to the fish from the selected and random groups during the culture time. These results suggest the

role of genetic basis for a difference of salinity tolerance between fish groups. The selection in the previous generation has probably increased the frequency of favorable alleles for salinity tolerance in the selected cohort. The transmission of these variants to the next generation leads to better performances for the offspring of the selected fish than for those of the randomly selected fish, a first step towards genetic adaptation to the new environmental conditions (Donelson et al., 2019). A genetic component of the salinity tolerance has been observed in several studies and species, such as Arctic charr (*Salvelinus alpinus*) (Norman et al., 2011) and Nile tilapia (*Oreochromis niloticus*) (Gu et al., 2018; Rengmark et al., 2007). In striped catfish, a number of putative genes related to salinity tolerance have been identified using RNA sequencing (Nguyen et al., 2014; Viet et al., 2016). The increased salinity tolerance in selection programs has been reported for tilapia in several previous studies (Jaspe and Caipang, 2011; Tran et al., 2008). Similarly, Afonso et al. (1998) have provided an extensive review of successful selection programs for stress tolerance on rainbow trout and common carp. Progressive genetic adaptation to salinity has been documented in laboratory strains of guppy (*Poecilia reticulate*) (Shikano et al., 2001; Shikano and Fujio, 1998a, 1998b). Purcell et al. (2008) also showed that mosquitofish *(Gambusia affinis)* living in saline environment required genetic adaptation through natural selection for higher individual salinity tolerance than fish from freshwater conditions where no such selection for salinity tolerance exists.

Genetic adaptation in a population occurs through the selection of favorable phenotypes that are heritable from one generation to another, leading to a shift in the frequency of alleles that change the mean of trait values (changing distribution of offspring phenotypes) under the altered conditions (Donelson et al., 2019). This process can sometimes occur remarkably rapid in a few generations when existing genetic variants, termed "standing genetic variation" and recombination rates are high (Barrick and Lenski, 2013). The reason being that it does not require the appearance of new beneficial mutations which potentially requires many generations. Furthermore, gametes, embryos and larvae might undergo selection for alleles that provide advantage in the parental environment, particularly in highly fecund species as fish (Torda et al., 2017). The coefficient of variation is a measure of the phenotypic diversity of the traits of interest in the tested population. A high value of this coefficient coupled with a reasonable heritability indicate that the high genetic diversity underlying the trait could be efficiently tapped in a selection program (Tahapari et al., 2018). In our study, growth was found to have a relatively high genetic variance, given its relatively large value of the coefficient of variation for weight (28.8 ± 4.0%) in the F1 generation at 10 ppt (Table 1 in Chapter 4) and a reasonable heritability (0.29). This indicates a good opportunity for a successful selection program targeting the process of genetic adaptation to the increase in salinity.

Salinity serves as a key barrier for the distribution of almost all aquatic organisms. The maintenance of plasma ion concentration with distinct salinity levels require an expensive adjustment of the osmoregulation, which consumes up to 50% of total energy (Gilles and Patrick, 2001; Hasan et al., 2017). Most freshwater fish need to regulate their blood osmolality in the 280 – 360 mOsm/kg to maintain homeostasis and iso-osmotic salinity of freshwater fish normally range from 10 – 12 ppt. In contrast, fish in seawater have to maintain a body fluid composition close to, but slightly higher than that of freshwater species, at around 300-400 mOsm/kg (Nguyen, 2015; Varsamos et al., 2005). In this study, average plasma osmotic pressures of striped catfish in 0 and 5ppt treatments were hyperosmotic to their treatment environment. When fish were exposed to the two highest tested salinities (15 and 20 ppt), fish experienced hyperosmotic conditions relative to their own plasma. Like most teleost, catfish can limit ion and water exchange using barriers at the skin surface via scales and a mucus layer when they live in their natural environment (Nguyen et al., 2014). However, there is an associated cost for these activities because they must use energy for osmoregulation to maintain their osmotic balance, in this case ranging from 268 to 286 mOsm/kg (Figure 5 in Chapter 4). In freshwater, plasma osmolalites were higher in fish (mean of three groups; 276.4 ± 6.4 mOsm/kg) than in the water environment (1.2 ± 0.6 mOsm/kg), while the reverse was true at 20 ppt (405.5 ± 28.3 mOsm/kg for the fish and 525.7 ± 6.6 mOsm/kg for the water environment). These values indicate that striped catfish are able to regulate internal osmolality to adapt to the salinity fluctuations. Interestingly, the osmolalities at 20 ppt in selected fish (417.3 ± 35.9 mOsm/kg) and random fish (414.9 ± 16.4 mOsm/kg) were significantly higher than in the freshwater group (384.5 ± 22.1 mOsm/kg). These differences indicate that the fish from the selected and random groups may have a better tolerance to stressful osmotic conditions in comparison to those from the freshwater group. As a result, selected and random fish use less energy for their osmoregulation in a hyperosmotic environment. These results may be used to explain better growth performance (especially with survival rate) of the selected and random groups compared to the freshwater group (Figure 4 in Chapter 4), especially at 20 ppt. As a consequence, strong selection is thus expected to drive the adaptation toward novel salinity levels (Goehlich et al., 2021)

### 3. Epigenetic considerations for the improvement of salinity tolerance of striped catfish

In aquaculture, maintaining and improving the production requires an understanding of genetic and physiological mechanisms that control the desired traits. Elucidation of these mechanisms has led to the development of pioneering approaches for applications in aquaculture (Gavery and Roberts, 2017). Given clear recent evidence about environmental influences on epigenetic mechanisms in aquatic species and the relationship between these mechanisms and phenotype, potential applications of epigenetics to aquaculture

could involve environmental manipulation. In aquaculture, two stages in the life-cycle (during larval development and during broodstock conditioning) are particularly sensitive to environmental influence on epigenetic mechanisms. The environmental conditions experienced in early life could influence the phenotype later in life (a memory of the environment), which could be beneficial in controlled aquaculture settings. Mechanisms behind such alteration of the phenotype may possibly be explained by a change of the methylation pattern of the genome induced by a change in the environment, which could alter the transcriptional capabilities (Bossdorf et al., 2007). In Olympia oyster, early exposure of larvae to ocean acidification influences juvenile traits (Hettinger et al., 2012). In addition, broodstock conditioning is also important to consider for the potential transmission of environmentally-induced epigenetic information between parents and their offspring. It has been reported that epigenetic transmission has the potential to be transmitted not only from the maternal side, but also from the paternal side (Rodgers et al., 2015; Soubry et al., 2014). According to Zhao et al. (2017), adult Manila clams exposed to low pH during gonadal maturation have faster- growing offspring compared to controls. The larvae of the Sydney rock oyster produced by parents incubated under low-pH conditions are larger and develop faster in low-pH conditions and also have higher fitness as adults (Parker et al., 2015, 2012). In the present thesis, analysis of the results showed an alteration of the DNA methylation pattern of striped catfish exposed to saline water at an early stage. These changes seem to induce changes in the expression of osmoregulation and growth performance genes. Moreover, our findings (in study N$^0$2, figure 2 to 4) indicate that offspring of striped catfish populations with historical salinity exposure seem more resistant to salinity stress than the offspring of populations without historical exposure. The difference in salinity tolerance between freshwater and random groups of individuals, two groups assumed to be reared in similar conditions but with different histories of exposition to salinity, might also suggest an epigenetic basis for the differences observed between fish originating from different environments. Taken together, these results suggest that exposure to salinity can change the epigenetic marks related to salinity tolerance on striped catfish and these changes may partly be transmitted to next generations.

According to Goddard and Whitelaw (2014), the epigenetic status of sites in the genome can be considered as a phenotype, that depends on the conventional DNA sequence, environmental effects (such as saline factor in our study), stochastic events in larvae and broodstock state and perhaps inherited epimutations. Therefore, it is possible that epigenetic status could be treated as a selection criterion and epigenetic markers could be integrated in selection programs in aquaculture (Kumar and Janmejay, 2021). In our case, if epigenetic markers associated with salinity tolerance on striped catfish were confirmed, they can be used in combination with genetic selection to improve the salinity tolerance of striped catfish.

However, the study of epigenetics and its application is still very scarce in aquaculture and unanswered questions remain (Kumar and Janmejay, 2021). In striped catfish, our study is the first one investigating the basic epigenetic mechanisms related to saline stimuli. One of the main constraints to the development of epigenetic solutions in aquaculture is the lack of a high-quality genome of several commercially important species, and data on the epigenome are even more limited. Moreover, the studies related to epigenetics often require a long time investment due to the long and complex life cycle of high-value aquatic species (Granada et al., 2018). This issue can become more complicated in the case of striped catfish with a very long generation time, with approximately three years to mature. A high-quality genome of striped catfish will facilitate many further investigations on epigenetics not only for salinity tolerance but also for other economic traits.

## 4. The future development of striped catfish aquaculture in Mekong Delta.

The striped catfish aquaculture plays a vital role in the agricultural sector in Vietnam, and is in a fast growth phase in terms of culture area and production. This fish will continue to be a key aquatic export commodity of Vietnam and contribute increasingly to Vietnamese aquaculture production and to the economy, the great bulk of which will trickle down to the poor rural communities of the Mekong Delta (De Silva and Nguyen, 2011; Hasan and Shipton, 2021; Phan et al., 2009). The export markets for Vietnamese catfish products have been diversified into more than 140 countries and territories (VASEP,2019). The future of striped catfish products requires a proper policy in widening export, expanding local and regional markets, and the development of brand image for striped catfish products from Vietnam. To meet the demand from markets, the culture area and production of striped catfish will increase in the coming years. However, the striped catfish industry is now challenged by several issues, that may impact its sustainability, including potential pollution from wastewater and effluents, disease outbreaks, reduction of seed quality, and especially saline intrusion as a result of climate change (Pham et al., 2021; Hoa et al., 2021; Hai et al., 2020). These emerging issues require new strategies for the sustainable development of striped catfish aquaculture in the future.

1) Applying selection programs to develop robust fish, that can better adapt to variations of the external environment including climate change, will be the best strategy for sustainable development of striped catfish, and dissemination of genetically improved stocks will in-turn efficiently increase aquaculture production and reduce environmental load, including GHG-emissions. In the case of our study, a selective breeding program has succeeded to develop a robust striped catfish with better growth performance in both freshwater and saline conditions.

The other selection programs focused on other economically important traits such as disease resistance and meat quality should be implemented in the future. Farming of genetically improved striped catfish will improve growth rates and animal welfare, survival rate, efficient use of resources such as feeds, water and energy, and lower cost per kg of produced fish.

2) The sustainability of striped catfish farming in the Mekong Delta is associated to the external aquatic environment. With more intense farming, high amount of effluents from striped catfish ponds that cannot be treated properly are considered as a source of pollution causing degradation of the water quality in the Mekong river in future. For this reason, control of effluent water quality from striped catfish pond culture systems to minimize environmental pollution is a crucial issue to enable sustainable development of striped catfish farming in the Mekong Delta. To minimize environmental pollution from striped catfish culture, the farmers should be encouraged to apply VietGAP, GlobalGAP, BAP, and ASC standards in their farming activities. These practices combined with the use of genetically improved striped catfish should significantly contribute to the sustainable development of the striped catfish industry in Mekong Delta in the future.

The Vietnamese government has recognized the role and potential of striped catfish species in the context of aquaculture development. Therefore, different supportive policies from the government have been established to enhance the sustainable development of the striped catfish industry in the Mekong Delta, especially under the impacts of climate change such as Vietnam Government Resolution No. 120 NQ-CP (dated 17/January/2017), Decision No. 3550/QD-BNN-TCTS and Resolution 120/NQ-CP (dated 12/August /2021). Several technical trainings and consultation programs for catfish farmers have been intensively implemented via universities, research institutions, and local organizations to improve their skills in striped catfish farming practices. These policies have brought many benefits to the development of striped catfish in the Mekong Delta. However, to deal with new emerging problems, the government needs to issue other policies for sustainable development for striped catfish farming and other aquatic species. For striped catfish aquaculture, public policies should focus on continuing selection programs for growth and salinity tolerance traits and for economically important traits such as disease resistance and/or meat quality based on genomic resources generated in the current study. However, the cost of commercial implementation of such a selective breeding program is unusually high and can be prohibitive for long-generation interval species such as striped catfish. For this reason, the policymakers and governance should provide incentives (such as public funding) to boost selective breeding programs on the most important and highest volume species, which may not be a priority for investment by private breeders due to long time for

in return to investment. With other aquatic species in the Mekong Delta, the public policies should support basic research for these species such as biology, immunology, reproduction, or even implementing selective breeding for traits of interest. This knowledge will enhance the technology for the culture of these species, which are fundamental steps to generating diversity in farmed species as well as contributing to the sustainable development of aquaculture in the Mekong Delta in the future.

# Chapter 8

# Conclusions and perspectives

---

# Chapter 8: Conclusions and perspectives

## 1. Conclusions

The striped catfish industry is a critical component of economic development in the Mekong Delta, with projections for remarkable expansion in the coming decades to meet the increasing demands from the international market. Selective breeding has offered substantial opportunities to enhance production efficiency and help tackle challenges in aquaculture via cumulative and permanent genetic improvement of farmed strains. In the present study, we have succeeded in implementing a selective breeding program to develop a productive strain of striped catfish with better tolerance to salinity after one generation of selection under saline environment. After one generation of selection, average direct response to selection for growth was 18% and a moderate realized heritability (0.29) was estimated for body weight. The improved salinity tolerance of striped catfish developed in this program is of practical significance in the context of Mekong Delta aquaculture, where the striped catfish breeding activities are projected to be greatly affected by salinity intrusion due to climate change. Using this new strain of striped catfish should allow farmers in the coastal provinces in Mekong Delta to increase the number of crops per year as well as to improve their productivity. Besides, results from the analysis of basic epigenetic mechanisms demonstrated that the methylation process during early development of this species was affected by the saline environment, and that these changes seem to be passed on to the next generation. The results suggest that in addition to the genetic effects, epigenetic components may play an important role in the improvement of salinity tolerance of striped catfish.

Moreover, to assist selective breeding programs, we have developed a new parentage assignment algorithm allowing to reconstruct the pedigree using genomic information from SWGS data. Using directly low coverage sequence data allows reducing the sequencing costs to be as low as possible. Furthermore, the cost and labor of developing marker panels can be avoided. The package program *Shallowped* for this algorithm is available for public use and we believe that it will be a useful tool in parentage assignment not only for aquaculture but also for a wide range of research communities. In addition, a high-quality genome of striped catfish was assembled using highly accurate long-read HiFi sequencing data, an important step to deliver reliable markers, which will allow to increase the power of this algorithm. With the improvement of the quality of the striped catfish genome, a large amount of “omics” data can be generated to understand more factors underlying the biology, including epigenetic mechanism and genetic architecture of complex traits. These results should accelerate genetic gain in aquaculture species throug selective breeding programs exploiting this new knowledge.

## 2. Perspectives

Nowadays, one of the biggest challenges faced by human mankind currently is to provide healthy food to the growing population projected to rise to over 9 billion by 2050, while addressing the adverse impacts of climate change and environmental degradation on the resource base (Lundh, 2020). In 2015, report from FAO (Food and Agriculture Organization) of the United Nations indicated that about 11% (or 814 million) of the world population were undernourished (FAO, 2015). Aquaculture has a crucial and rapidly increasing role in food security and economic stability worldwide, providing over half of the world's seafood supply, with projections for significant expansion in the coming decades to meet the increasing demands for animal protein production related with human population growth (Houston et al., 2020; Nguyen et al., 2022). According to Anderson et al. (2017), more than 90% of global aquaculture occurs in low and middle income countries, where it provides major contributions to the Sustainable Development Goals of the United Nations, either directly through human consumption or indirectly through economic growth. To improve the sustainable development of future aquaculture, we need to increase the number and efficiency of selective breeding programs (Houston et al., 2020; Gjedrem et al., 2012). Genetic improvement originating from a well-managed breeding programme makes possible cumulative increases in production traits, and facilitates adaptation to emerging challenges including climate change or infectious disease outbreaks (Sae-Lim et al., 2017; Gjedrem and Rye, 2016; Houston et al., 2020). Striped catfish is one of the three top consumed freshwater fish types (after carps and tilapia), and VietNam is the source of more than 90 percent of the world's striped catfish exports, which has increased 50-fold in the last decade (FAO,2019). Recently, striped catfish aquculture has been developed in several Asian countries such as Bangladesh, India, China, Indonesia, Myanmar, Nepal, and Pakistan. With the important role of striped catfish production in international markets, the success of selective breeding of saline-tolerant striped catfish in the present study could play a crucial role in contributing to food security and economic stability not only in Vietnam but also in the world under the impact of climate change.

In the future, the combined impact of sea level rise, changes in the upstream freshwater flow, and tidal regime on salinity intrusion due to climate change will cause more impacts on striped catfish farming owing to shrinkage of the freshwater zone. For this reason, selective breeding of saline-tolerant striped catfish needs to be continued for more generations to increase the genetic gain on salinity tolerance. Importantly, MAS or GS could be used to increase the efficiency of selection. With a high-quality genome of striped catfish, there is ample room for a reliable identification of numerous SNPs and other DNA markers. These DNA markers could be used to identify QTLs not only for salinity tolerance traits, but also for other economically important traits of striped catfish. The DNA markers associated to a trait of interest

can be used in selecting progenies carrying favourable alleles via MAS or GS. For GS, the availability of a large set of markers, shallow sequencing and the use of imputation should allow to target accurate breeding values estimations using the latest evaluation techniques. However, exclusive emphasis on production traits of selective breeding program may lead to detrimental correlated responses in other traits, particularly those associated with fitness. For example, on livestock, although selection for milk production in dairy cattle was extremely successful, there was a substantial undesired decrease in fertility over the same period (Georges et al., 2019). Thus, next selection schemes on striped catfish (or other aquatic species) should focus on the balance between animal health, fertility, production and environmental impact.

Although the application of epigenetic knowledge could significantly affect the productivity and sustainability of aquaculture practice, epigenetics and its application are at an early stage of development, especially for striped catfish for which no other study exists to date. Many questions remain unanswered regarding the epigenetic mechanisms not only for salinity tolerance traits, but also for other economically important traits for aquatic species in general, and for striped catfish in particular. Further studies should investigate on how the genome sequence impacts the epigenome, how epigenetic events interact with genotype to influence the induction of phenotypic traits under saline conditions and use this fundamental information to improve salinity tolerance of striped catfish. Moreover, in next selection generation, if transgenerational inheritance of epimutations associated to salinity tolerance are confirmed, the influence of the epigenetic processes on the estimation of breeding value and accuracy of selection programs for genetic improvement of salinity tolerance should be considered and evaluated.

Recently, with the rapid development of biotechnological innovations, promising approaches to tackle production barriers in aquaculture are emerging. For example, these innovations include the use of genome editing technologies to make targeted changes to the genomes of aquaculture species, resulting in improved health and performance, or the use of reproductive biotechnologies such as surrogate broodstock to expedite genetic gain, or combinations of both approaches (Houston et al., 2020). Genome editing tools such as engineered CRISPR-Cas9 systems are invaluable for understanding genetic regulation of economically important traits and have the potential to accelerate genetic gain in aquaculture breeding programs (Cong et al., 2013; Mali et al., 2013). In a breeding program, a key factor in the rate of genetic gain is the length of the generation interval. Decreasing generation time potentially increases the genetic gain rate via the reduction of the generation interval (Houston et al., 2020). This is particularly true in aquaculture, considering that many of the major aquaculture species have relatively long generation intervals (for example, up to more than 3 years to get maturation of striped catfish). When combined with GS to predict breeding value of embryos or juveniles, surrogate broodstock technology could potentially

reduce the generation interval without substaintial loss of selection accuracy (Zenger et al., 2019; Yoshizaki and Yazawa, 2019). Generally, the breeding program should ideally utilize genomic information to fast-track the genetic improvement of farmed aquatic species, which will be necessary to secure the sustainable growth of aquaculture as one of the most promising solutions to the current global food security challenge.

Despite the fact that aquaculture does not directly emit GHGs, aquaculture activities, such as input power, transport, and feed production contribute to GHG emissions. LCA (life cycle analysis) is a method to quantify the use of resources and emission of pollutants in the entire production chain for a product. In aquaculture, environmental impacts related to aquatic commodities have been quantified in various LCA studies (Henriksson et al., 2015a, 2015b). Selective breeding programs for increased production traits are expected to enhance the efficiency of resource utilization (feed, energy, water, drug/chemicals, and land) of a production system, through correlated changes in feed efficiency or shorter production period. Applications of LCA to define breeding goals that maximize production, while minimizing environmental impacts have been implemented in the study of Besson et al. (2016) on African catfish (*Clarias gariepinus*). With that method, it is possible to evaluate different breeding strategies that maximize total genetic gain and minimize environmental loads. For these benefits, this method should be considered for incorporation in selective breeding programs in the future to optimize the selection target.

# References

# References

**-A-**

Abdelrahman, H., ElHady, M., Alcivar-Warren, A., Allen, S., Al-Tobasei, R., Bao, L., Beck, B., Blackburn, H., Bosworth, B., et al., 2017. Aquaculture genomics, genetics and breeding in the United States: Current status, challenges, and priorities for future research. BMC Genomics 18. https://doi.org/10.1186/s12864-017-3557-1

Álvarez-Venegas, R., De-la-Peña, C., 2016. Editorial: Recent advances of epigenetics in crop biotechnology. Front. Plant Sci. 7, 1–3. https://doi.org/10.3389/fpls.2016.00413

Anderson, E. C., & Garza, J. C. (2006). The Power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics Society of America*, *172*(April), 2567–2582. doi:10.1534/genetics.105.048074.

Afonso, J., Gines, R., Montero, D., 1998. Selection programmes for stress tolerance in fish. Genet. Breed. Mediterr. Aquac. species. 245, 235–245.

Ahmed, N., Thompson, S., Glaser, M., 2019. Global Aquaculture Productivity, Environmental Sustainability, and Climate Change Adaptability. Environ. Manage. 63, 159–172.

Allred, S., Shao, W., Schilling, M.W., Petrie-Hanson, L., Allen, P.J., 2019. An assessment of red fillet prevalence in the catfish industry. Aquaculture 507, 203–210.

Anderson, J.L., Asche, F., Garlock, T., Chu, J., 2017. Aquaculture: Its role in the future of food. Front. Econ. Glob. 17, 159–173. https://doi.org/10.1108/S1574-871520170000017011

Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J., Pozniak, C.J., et al., 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science (80-. ). 361. https://doi.org/10.1126/science.aar7191

Aranguren Caro, L.F., Alghamdi, F., De Belder, K., Lin, J., Mai, H.N., Millabas, J., Alrehaili, Y., Alazwari, A., Algetham, S., Dhar, A.K., 2021. The effect of salinity on Enterocytozoon hepatopenaei infection in Penaeus vannamei under experimental conditions. BMC Vet. Res. 17, 65.

Amornsakun, T., Vo, V.H., Petchsupa, N., Pau, T.M., Hassan, A. Bin, 2017. Effects of water salinity on hatching of egg, growth and survival of larvae and fingerlings of snake head fish, *channa striatus*. Songklanakarin J. Sci. Technol. 39, 137–142. https://doi.org/10.14456/sjst-psu.2017.16.

Ágústsson T, Sundell K, Sakamoto T, et al (2003) Pituitary gene expression of somatolactin, prolactin, and growth hormone during Atlantic salmon parr–smolt transformation. Aquaculture 222:229–238. https://doi.org/10.1016/S0044-8486(03)00124-8

Alderdice DF (1988) 3 Osmotic and Ionic Regulation in Teleost Eggs and Larvae. Fish Physiol 11:163–251. https://doi.org/10.1016/S1546-5098(08)60200-9

Ambrosi C, Manzo M, Baubec T (2017) Dynamics and context-dependent roles of DNA methylation. J Mol Biol 429:1459–1475

Anastasiadi D, Esteve-Codina A, Piferrer F (2018) Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. Epigenetics Chromatin 11:1–17

Arnesen AM, Toften H, Agustsson T, et al (2003) Osmoregulation, feed intake, growth and growth hormone levels in 0+ Atlantic salmon (*Salmo salar* L.) transferred to seawater at different stages of smolt development. Aquaculture 222:167–187. https://doi.org/10.1016/S0044-8486(03)00109-1

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., Gouil, Q., 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21, 1–16. https://doi.org/10.1186/s13059-020-1935-5

Andrews, S., 2010. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [WWW Document]. URL http://www.bioinformatics.babraham.ac.uk/ projects/fastqc/ (accessed 8.17.19).

Austin, C.M., Tan, M.H., Harrisson, K.A., Lee, Y.P., Croft, L.J., Sunnucks, P., Pavlova, A., Gan, H.M., 2017. De novo genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (Maccullochella peelii), from Illumina and Nanopore sequencing read. Gigascience 6, 1–19. https://doi.org/10.1093/gigascience/gix063

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81–92. doi: 10.1038/nrg.2015.28

Argue, B.J., Arce, S.M., Lotz, J.M., Moss, S.M., 2002. Selective breeding of Pacific white shrimp (*Litopenaeus vannamei*) for growth and resistance to Taura Syndrome Virus. Aquaculture 204, 447–460. https://doi.org/10.1016/S0044-8486(01)00830-4

Arora, K., Shah, M., Johnson, M., Sanghvi, R., Shelton, J., Nagulapalli, K., Oschwald, D.M., Zody, M.C., Germer, S., Jobanputra, V., Carter, J., Robine, N., 2019. Deep whole-genome sequencing of 3 cancer cell lines on 2 sequencing platforms. Sci. Rep. 9, 1–13. https://doi.org/10.1038/s41598-019-55636-3

Artemov, A. V., Mugue, N.S., Rastorguev, S.M., Zhenilo, S., Mazur, A.M., Tsygankova, S. V., Boulygina, E.S., Kaplun, D., Nedoluzhko, A. V., Medvedeva, Y.A., Prokhortchouk, E.B., 2017. Genome-wide DNA methylation profiling reveals epigenetic adaptation of stickleback to marine and freshwater conditions. Mol. Biol. Evol. 34, 2203–2213. https://doi.org/10.1093/molbev/msx156

Azzi, A., Dallmann, R., Casserly, A., Rehrauer, H., Patrignani, A., Maier, B., Kramer, A., Brown, S.A., 2014. Circadian behavior is light-reprogrammed by plastic DNA methylation. Nat. Neurosci. 17, 377–382. https://doi.org/10.1038/nn.3651

**-B-**

Baerwald, M.R., Meek, M.H., Stephens, M.R., Nagarajan, R.P., Goodbla, A.M., Tomalty, K.M.H., Thorgaard, G.H., May, B., Nichols, K.M., 2016. Migration-related phenotypic divergence is associated with epigenetic modifications in rainbow trout. Mol. Ecol. 25, 1785–1800.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., Eichler, E.E., 2002. Recent segmental duplications in the human genome. Science (80-. ). 297, 1003–1007. https://doi.org/10.1126/science.1072047

Bansal, V., Boucher, C., 2019. Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going? iScience 18, 37–41. https://doi.org/10.1016/j.isci.2019.06.035

Barzon, L., Lavezzo, E., Militello, V., Toppo, S., Palù, G., 2011. Applications of next-generation sequencing technologies to diagnostic virology. Int. J. Mol. Sci. 12, 7861–7884.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*(10), 1–7. doi: 10.1371/journal.pone.0003376

Baubec T, Schübeler D (2014) Genomic patterns and context specific interpretation of DNA methylation. Curr Opin Genet Dev 25:85–92

Bandyopadhyay, P., Das Mohapatra, P.K., 2009. Effect of a probiotic bacterium Bacillus circulans PB7 in the formulated diets: On growth, nutritional quality and immunity of *Catla catla* (Ham.). Fish Physiol. Biochem. 35, 467–478. https://doi.org/10.1007/s10695-008-9272-8

Barrick, J.E., Lenski, R.E., 2013. Genome dynamics during experimental evolution. Nat. Rev. Genet. 14, 827–839. https://doi.org/10.1038/nrg3564

Barton, B., Morgan, J., Vijayan, M., Adams, S., 2002. Physiological and condition-related indicators of environmental stress in fish. Biological indicators of aquatic ecosystem stress, 111-148

Besson, M., Aubin, J., Komen, H., Poelman, M., Quillet, E., Vandeputte, M., Van Arendonk, J.A.M., De Boer, I.J.M., 2016. Environmental impacts of genetic improvement of growth rate and feed conversion ratio in fish farming under rearing density and nitrogen output limitations. J. Clean. Prod. 116, 100–109. https://doi.org/10.1016/j.jclepro.2015.12.084

Bland, J.M., Bett-Garber, K.L., Li, C.H., Brashear, S.S., Lea, J.M., Bechtel, P.J., 2018. Comparison of sensory and instrumental methods for the analysis of texture of cooked individually quick frozen and fresh-frozen catfish fillets. Food Sci. Nutr. 6, 1692–1705. https://doi.org/10.1002/fsn3.737

Bentsen, H.B., Gjerde, B., Eknath, A.E., de Vera, M.S.P., Velasco, R.R., Danting, J.C., Dionisio, E.E., Longalong, F.M., Reyes, R.A., Abella, T.A., Tayamen, M.M., Ponzoni, R.W., 2017. Genetic improvement of farmed tilapias: Response to five generations of selection for increased body weight at harvest in *Oreochromis niloticus* and the further impact of the project. Aquaculture 468, 206–217. https://doi.org/10.1016/j.aquaculture.2016.10.018

Borode, A.O., Balogun, A.M., Omoyeni, B.A., 2002. Effect of salinity on embryonic development, hatchability, and growth of african catfish, *clarias gariepinus*, eggs and larvae. J. Appl. Aquac. 12, 89–93. https://doi.org/10.1300/J028v12n04_08

Bestor TH (2000) The DNA methyltransferases of mammals. Hum Mol Genet 9:2395–2402

Binh D Van, Kantoush S, Sumi T (2020) Changes to long-term discharge and sediment loads in the Vietnamese Mekong Delta caused by upstream dams. Geomorphology 353:107011. https://doi.org/https://doi.org/10.1016/j.geomorph.2019.107011

Bird A (2002) DNA methylation patterns and epigenetic memory. Genes Dev 16:6–21

Bœuf G, Payan P (2001) How should salinity influence fish growth? Comp Biochem Physiol Part C Toxicol Pharmacol 130:411–423. https://doi.org/10.1016/S1532-0456(01)00268-X

Bogdanović O, Smits AH, de la Calle Mustienes E, et al (2016) Active DNA demethylation at enhancers during the vertebrate phylotypic period. Nat Genet 48:417–426

Bogdanović O, Veenstra GJC (2009) DNA methylation and methyl-CpG binding proteins: developmental requirements and function. Chromosoma 118:549–565. https://doi.org/10.1007/s00412-009-0221-9

Bossdorf O, Richards CL, Pigliucci M (2008) Epigenetics for ecologists. Ecol Lett 11:106–115

Branco MR, Oda M, Reik W (2008) Safeguarding parental identity: Dnmt1 maintains imprints during epigenetic reprogramming in early embryogenesis. Genes Dev 22:1567–1571. https://doi.org/10.1101/gad.1690508

Breves JP, Fujimoto CK, Phipps-Costin SK, et al (2017) Variation in branchial expression among insulin-like growth-factor binding proteins (igfbps) during Atlantic salmon smoltification and seawater exposure. BMC Physiol 17:1–11

Boichard, D., Barbotte, L., Genestout, L. (2014) AccurAssign, software for accurate maximum-likelihood parentage assignment. 10. World Congress on Genetics Applied to Livestock Production (WCGALP), Aug 2014, Vancouver, Canada. ffhal-01193905

Bonin, A., Bellemain, E., Eidesen, P. B., Pompanon, F., Brochmann, C., & Taberlet, P. (2004). How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, *13*(11), 3261–3273. doi: 10.1111/j.1365-294X.2004.02346.x

Bresadola, L., Link, V., Buerkle, C.A., Lexer, C., Wegmann, D. (2020). Estimating and accounting for genotyping errors in RAD-seq experiments. *Mol Ecol Resour*.20: 856– 870.https://doi.org/ 10.1111/1755-0998.13153

Beagan, J.J., Drees, E.E.E., Stathi, P., Eijk, P.P., Meulenbroeks, L., Kessler, F., Middeldorp, J.M., Pegtel, D.M., Zijlstra, J.M., Sie, D., Heideman, D.A.M., Thunnissen, E., Smit, L., de Jong, D., Mouliere, F., Ylstra, B., Roemer, M.G.M., van Dijk, E., 2021. PCR-Free Shallow Whole Genome Sequencing for Chromosomal Copy Number Detection from Plasma of Cancer Patients Is an Efficient Alternative to the Conventional PCR-Based Approach. J. Mol. Diagnostics 23, 1553–1563.

Becker, W.A., 1968. Manual of procedures in quantitative genetics. Washingt. Univ., Press. WA, U.S.A.

Benevenuto, J., Ferrão, L.F. V., Amadeu, R.R., Munoz, P., 2019. How can a high-quality genome assembly help plant breeders? Gigascience 8, 1–4. https://doi.org/10.1093/gigascience/giz068

Bentsen, H.B., Olesen, I., 2002. Designing aquaculture mass selection programs to avoid high inbreeding rates. Aquaculture 204, 349–359. https://doi.org/10.1016/S0044-8486(01)00846-8

Bernard, C., Edwige, Q., Francine, K., Marie-gwénola, H., Muriel, M., André, F., Laurent, L., Jean-pierre, H., Marc, V., 2004. Enhanced individual selection for selecting fast growing fish : the " PROSPER " method , with application on brown trout ( *Salmo trutta fario* ) 36, 643–661.

Biémont, C., 2010. A brief history of the status of transposable elements: From junk DNA to major players in evolution. Genetics 186, 1085–1093. https://doi.org/10.1534/genetics.110.124180

Bjornsson, H.T., Fallin, M.D., Feinberg, A.P., 2004. An integrated epigenetic and genetic approach to common human disease. Trends Genet. 20, 350–358. https://doi.org/10.1016/j.tig.2004. 06.009

Borrell, Y.J., Gallego, V., García-Fernández, C., Mazzeo, I., Pérez, L., Asturiano, J.F., Carleos, C.E., Vázquez, E., Sánchez, J.A., Blanco, G., 2011. Assessment of parental contributions to fast- and slow-growing progenies in the sea bream Sparus aurata L. using a new multiplex PCR. Aquaculture 314, 58–65. https://doi.org/10.1016/j.aquaculture.2011.01.028

Berlin, K., Koren, S., Chin, C.S., Drake, J.P., Landolin, J.M., Phillippy, A.M., 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol. 33, 623–630. https://doi.org/10.1038/nbt.3238

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A., Aury, J.M., Louis, A., Dehais, P., Bardou, P., Montfort, J., Klopp, C., Cabau, C., Gaspin, C., Thorgaard, G.H., Boussaha, M., Quillet, E., Guyomard, R., Galiana, D., Bobe, J., Volff, J.N., Genêt, C., Wincker, P., Jaillon, O., Crollius, H.R., Guiguen, Y., 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat. Commun. 5. https://doi.org/10.1038/ncomms4657

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brawand, D., Wagner, C.E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A.Y., Lim, Z.W., Bezault, E., Turner-Maier, J.,et al., 2015. The genomic substrate for adaptive radiation in African cichlid fish. Nature 513, 375–381. https://doi.org/10.1038/nature13726

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., Shendure, J., 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. 31, 1119–1125. https://doi.org/10.1038/nbt.2727

Bossdorf, O., Richards, C.L., Pigliucci, M., 2007. Epigenetics for ecologists. Ecol. Lett. 071117033013002-??? https://doi.org/10.1111/j.1461-0248.2007.01130.x

Bui, T.M., Phan, L.T., Ingram, B. a., Nguyen, T.T.T., Gooley, G.J., Nguyen, H. V., Nguyen, P.T., De Silva, S.S., 2010a. Seed production practices of striped catfish, Pangasianodon hypophthalmus in the Mekong Delta region, Vietnam. Aquaculture 306, 92–100. https://doi.org/10.1016/j.aquaculture.-2010.06.016

Bui, T.M., Phan, L.T., Ingram, B.A., Nguyen, T.T.T., Gooley, G.J., Nguyen, H. V., Nguyen, P.T., De Silva, S.S., 2010b. Seed production practices of striped catfish, Pangasianodon hypophthalmus in the Mekong Delta region, Vietnam. Aquaculture. https://doi.org/10.1016/j.aquaculture.2010.06.016

Burgerhout, E., Mommens, M., Johnsen, H., Aunsmo, A., Santi, N., Andersen, O., 2017. Genetic background and embryonic temperature affect DNA methylation and expression of myogenin and muscle development in Atlantic salmon (Salmo salar). PLoS One 12. https://doi.org/10.1371/journal.pone.0179918

**-C-**

Caco, P., 1998. Description of the sexual cycle related to the environment and set up of the artificial propagation in pangasius bocourtl(sauvage, 1880) and pangasius hypophthalmus (sauvage, 1878), reared in floating cages and in ponds in the Mekong Delta. Aquac. Res. Unit, 1, 71.

Cacot, P., Legendre, M., Dan, T.Q., Tung, L.T., Liem, P.T., Mariojouls, C., Lazard, J., 2002. Induced ovulation of Pangasius bocourti (Sauvage, 1880) with a progressive hCG treatment. Aquaculture 213, 199–206. https://doi.org/10.1016/S0044-8486(02)00033-9

Callet, T., Médale, F., Larroquet, L., Surget, A., Aguirre, P., Kerneis, T., Labbé, L., Quillet, E., Geurden, I., Skiba-Cassy, S., Dupont-Nivet, M., 2017. Successful selection of rainbow trout (Oncorhynchus mykiss) on their ability to grow with a diet completely devoid of fishmeal and fish oil, and correlated changes in nutritional traits. PLoS One 12, 1–21. https://doi.org/10.1371/journal.pone. 0186705

Campos, C., Valente, L.M.P., Conceição, L.E.C., Engrola, S., Fernandes, J.M.O., 2013. Temperature affects methylation of the myogenin putative promoter, its expression and muscle cellularity in Senegalese sole larvae. Epigenetics 8, 389–397. https://doi.org/10.4161/epi.24178.

Casadei R, Pelleri MC, Vitale L, et al (2011) Identification of housekeeping genes suitable for gene expression analysis in the zebrafish. Gene Expr Patterns 11:271–276. https://doi.org/https://doi.org/ 10.1016/ j.gep.2011.01.003

Campbell, I.C., 2012. Biodiversity of the Mekong Delta. pp. 293–313. https://doi.org/10.1007/978-94-007-3962-8_11

Coetzer, W.G., Downs, C.T., Perrin, M.R., Willows-Munro, S., 2017. Testing of microsatellite multiplexes for individual identification of Cape Parrots ( Poicephalus robustus ): paternity testing and monitoring trade. PeerJ 5, e2900. https://doi.org/10.7717/peerj.2900

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L. a, 2013. Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., … Zhang, F. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. Science (New York, N.Y.). Science 339, 819–823. https://doi.org/10.1126/science.1231143.Multiplex

Cnaani, A., Hulata, G., 2011. Improving salinity tolerance in tilapias: Past experience and future prospects. Isr. J. Aquac. - Bamidgeh 63. https://doi.org/10.46989/001c.20590

Chen T, Ueda Y, Dodge JE, et al (2003) Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. Mol Cell Biol 23:5594–5605

Chaisson, M.J.P., Wilson, R.K., Eichler, E.E., 2015. Genetic variation and the de novo assembly of human genomes. Nat. Rev. Genet. 16, 627–640. https://doi.org/10.1038/nrg3933

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H., 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods 18, 170–175. https://doi.org/10.1038/s41592-020-01056-5

Chin, C.-S., Khalak, A., 2019. Human Genome Assembly in 100 Minutes. bioRxiv 705616.

Cercueil, A., Bellemain, E., Manel, S., 2002. PARENTE: Computer program for parentage analysis. J. Hered. 93, 458–459. https://doi.org/10.1093/jhered/93.6.458

CGIAR, 2016. The drought and salinity intrusion in the Mekong River Delta of Vietnam. CGIAR Res. Centers Southeast Asia 54 pp.

Chin, S.F., Santonja, A., Grzelak, M., Ahn, S., Sammut, S.J., Clifford, H., Rueda, O.M., Pugh, M., Goldgraben, M.A., Bardwell, H.A., Cho, E.Y., Provenzano, E., Rojo, F., Alba, E., Caldas, C., 2018. Shallow whole genome sequencing for robust copy number profiling of formalin-fixed paraffin-embedded breast cancers. Exp. Mol. Pathol. 104, 161–169. https://doi.org/10.1016/j.yexmp.018.03.006

Covelo-Soto, L., Leunda, P.M., Pérez-Figueroa, A., Morán, P., 2015. Genome-wide methylation study of diploid and triploid brown trout ( Salmo trutta L.). Anim. Genet. 46, 280–288. https://doi.org/10.1111/age.12287

**-D-**

Danzmann, R.G., 1997. PROBMAX: A Computer Program for Assigning Unknown Parentage in Pedigree Analysis From Known Genotypic Pools of Parents and Progeny. J. Hered. 88, 333–333. https://doi.org/10.1093/oxfordjournals.jhered.a023112

Dashrath, K., Chakrapani, V., Kumar, A., Rambhau, A., Sundaray, J.K., Ninawe, A., Jayasankar, P., 2017. Status and future perspectives of single nucleotide polymorphisms ( SNPs ) markers in farmed fi shes : Way ahead using next generation sequencing. Gene Reports 6, 81–86.

Daxinger, L., Whitelaw, E., 2012. Understanding transgenerational epigenetic inheritance via the gametes in mammals. Nat. Rev. Genet. 13, 153–162. https://doi.org/10.1038/nrg3188.

Dabe EC, Sanford RS, Kohn AB, et al (2015) DNA Methylation in Basal Metazoans: Insights from Ctenophores. Integr Comp Biol 55:1096–1110. https://doi.org/10.1093/icb/icv086

Dasmahapatra AK, Khan IA (2016) Modulation of DNA methylation machineries in Japanese rice fish (*Oryzias latipes*) embryogenesis by ethanol and 5-azacytidine. Comp Biochem Physiol Part C Toxicol Pharmacol 179:174–183

Dao, H., Duong, Y., Pham, L., Bui, T., Vo, S., 2022. Selective breeding of saline-tolerant striped catfish (*Pangasianodon hypophthalmus*) for sustainable catfish farming in climate vulnerable Mekong Delta , Vietnam 25. https://doi.org/10.1016/j.aqrep.2022.101263

d'Orbcastel, E.R., Blancheton, J.-P., Aubin, J., 2009. Towards environmentally sustainable aquaculture: Comparison between two trout farming systems using Life Cycle Assessment. Aquac. Eng. 40, 113–119. https://doi.org/10.1016/j.aquaeng.2008.12.002

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., Calus, M.P.L., 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193, 327–345. https://doi.org/10.1534/genetics.112.143313

Dean W, Santos F, Reik W (2003) Epigenetic reprogramming in early mammalian development and following somatic nuclear transfer. In: Seminars in cell & developmental biology. Elsevier, pp 93–100

Deane EE, Woo NYS (2005) Upregulation of the Somatotropic Axis Is Correlated with Increased G6PDH Expression in Black Sea Bream Adapted to Iso-osmotic Salinity. Ann N Y Acad Sci 1040:293–296

Deane EE, Woo NYS (2009) Modulation of fish growth hormone levels by salinity, temperature, pollutants and aquaculture related stress: a review. Rev Fish Biol Fish 19:97–120. https://doi.org/10.1007/s11160-008-9091-0

Drennon K, Moriyama S, Kawauchi H, et al (2003) Development of an enzyme-linked immunosorbent assay for the measurement of plasma growth hormone (GH) levels in channel catfish (*Ictalurus punctatus*): assessment of environmental salinity and GH secretogogues on plasma GH levels. Gen Comp Endocrinol 133:314–322. https://doi.org/10.1016/S0016-6480(03)00194-1

De Silva, S.S., Nguyen, P.T., 2011. Striped catfish farming in the Mekong Delta, Vietnam: a tumultuous path to a global success. Rev. Aquac. 3, 45–73.

De Silva, S.S., Soto, D., 2009. Climate change and aquaculture: potential impacts, adaptation and mitigation. In: Climate change implications for fisheries and aquaculture: Overview of current scientific knowledge. Fisheries and Aquaculture Technical Paper No. 583. FAO, Rome. 10–13.

Donelson, J.M., Sunday, J.M., Figueira, W.F., Gaitán-Espitia, J.D., Hobday, A.J., Johnson, C.R., Leis, J.M., Ling, S.D., Marshall, D., Pandolfi, J.M., Pecl, G., Rodgers, G.G., Booth, D.J., Munday, P.L., 2019. Understanding interactions between plasticity, adaptation and range shifts in response to marine environmental change. Philos. Trans. R. Soc. B Biol. Sci. 374, 20180186.

Dong, Z., Nguyen, N.H., Zhu, W., 2015. Genetic evaluation of a selective breeding program for common carp *Cyprinus carpio* conducted from 2004 to 2014. BMC Genet. 16, 1–9.

de Koning, A.P.J., Gu, W., Castoe, T.A., Batzer, M.A., Pollock, D.D., 2011. Repetitive elements may comprise over Two-Thirds of the human genome. PLoS Genet. 7.

Dennis, M.Y., Eichler, E.E., 2016. Human adaptation and evolution by segmental duplication. Curr. Opin. Genet. Dev. 41, 44–52. https://doi.org/10.1016/j.gde.2016.08.001

Dodds, K.G., McEwan, J.C., Brauning, R., Anderson, R.M., Stijn, T.C., Kristjánsson, T., Clarke, S.M., 2015. Construction of relatedness matrices using genotyping-by-sequencing data. BMC Genomics 16, 1–15. https://doi.org/10.1186/s12864-015-2252-3

Dodds, K.G., McEwan, J.C., Brauning, R., Van, T.C., Suzanne, S.J., Mary McEwan, R.K., Clarke, S.M., 2019. Exclusion and genomic relatedness methods for assignment of parentage using genotyping-by-sequencing data. G3 Genes, Genomes, Genet. 9, 3239–3247. https://doi.org/10.1534/g3.119.400501

DOF, 2020. Directorate of Fisheries of Provinces, 2020. The striped catfish production in 2020.

DOF, 2019. Report on striped catfish aquaculture 2019. Directorate of Fisheries (DOF) of Vietnam.

Dolinoy, D.C., Weidman, J.R., Waterland, R.A., Jirtle, R.L., 2006. Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. Environ. Health Perspect. 114, 567–572. https://doi.org/10.1289/ehp.8700

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

Derakhshani, H., Bernier, S.P., Marko, V.A., Surette, M.G., 2020. Completion of draft bacterial genomes by long-read sequencing of synthetic genomic pools. BMC Genomics 21, 1–11. https://doi.org/10.1186/s12864-020-06910-6

Di Genova, A., Buena-Atienza, E., Ossowski, S., Sagot, M.F., 2021. Efficient hybrid de novo assembly of human genomes with WENGAN. Nat. Biotechnol. 39, 422–430. https://doi.org/10.1038/s41587-020-00747-w

Driguez, P., Bougouffa, S., Carty, K., Putra, A., Jabbari, K., Reddy, M., Soppe, R., Cheung, M.S., Fukasawa, Y., Ermini, L., 2021. LeafGo: Leaf to Genome, a quick workflow to produce high-quality de novo plant genomes using long-read sequencing technology. Genome Biol. 22, 1–18. https://doi.org/10.1186/s13059-021-02475-z

Douglas, J. A., Skol, A. D., & Boehnke, M. (2002). Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *American Journal of Human Genetics*, *70*(2), 487–495. doi: 10.1086/338919

Domingos, J.A., Smith-Keune, C., Robinson, N., Loughnan, S., Harrison, P., Jerry, D.R., 2013. Heritability of harvest growth traits and genotype–environment interactions in barramundi, Lates calcarifer (Bloch). Aquaculture 402–403, 66–75. https://doi.org/10.1016/j.aquaculture.2013.03.029

Domiszewski, Z.Z., Grzegorz, B., Dominika, P., Bienkiewicz, G., Plust, D., 2011. Effects of different heat treatments on lipid quality of striped catfish (. Acta Sci. Pol. Technol. Aliment. 10, 359–373.

Dong, S., Kong, J., Zhang, T., Meng, X., Wang, R., 2006. Parentage determination of Chinese shrimp (Fenneropenaeus chinensis) based on microsatellite DNA markers. Aquaculture 258, 283–288. https://doi.org/10.1016/j.aquaculture.2006.04.044

Dowen, R.H., Pelizzola, M., Schmitz, R.J., Lister, R., Dowen, J.M., Nery, J.R., Dixon, J.E., Ecker, J.R., 2012. Widespread dynamic DNA methylation in response to biotic stress. Proc. Natl. Acad. Sci. U. S. A. 109. https://doi.org/10.1073/pnas.1209329109

Doyle, R., Herbinger, C., 1994. The use of DNA fingerprinting for high-intensity , within-family selection in fish breeding.

Duchesne, P., Godbout, M.H., Bernatchez, L., 2002. PAPA (package for the analysis of parental allocation): A computer program for simulated and real parental allocation. Mol. Ecol. Notes 2, 191–193. https://doi.org/10.1046/j.1471-8286.2002.00164.x

Dussault, F.M., Boulding, E.G., 2018. Effect of minor allele frequency on the number of single nucleotide polymorphisms needed for accurate parentage assignment: A methodology illustrated using Atlantic salmon. Aquac. Res. 49, 1368–1372. https://doi.org/10.1111/are.13566

Dussex, N., van der Valk, T., Morales, H.E., Wheat, C.W., Díez-del-Molino, D., von Seth, J., Foster, Y., Kutschera, V.E., Guschanski, K., Rhie, A., Phillippy, A.M., Korlach, J., Howe, K., Chow, W., Pelan, Dalén, L., 2021. Population genomics of the critically endangered kākāpō. Cell Genomics 1, 100002. https://doi.org/10.1016/j.xgen.2021.100002

**-E-**

Estoup, A., Gharbi, K., SanCristobal, M., Chevalet, C., Haffray, P., Guyomard, R., 1998. Parentage assignment using microsatellites in turbot ( *Scophthalmus maximus* ) and rainbow trout ( Oncorhynchus mykiss ) hatchery populations. Can. J. Fish. Aquat. Sci. 55, 715–723. https://doi.org/10.1139/f97-268

Evans, D., 2011. Osmoregulation in fish: an introduction encyclopedia of fish physiolgy: from genome to enviroment, vol Gas exchange, internal homeostatis, and food uptake. Elsevier, Amsterdam.

Evans, S., Langdon, C., 2006. Direct and indirect responses to selection on individual body weight in the Pacific oyster (*Crassostrea gigas*). Aquaculture 261, 546–555.

Eckert SM, Yada T, Shepherd BS, et al (2001) Hormonal Control of Osmoregulation in the Channel Catfish *Ictalurus punctatus*. Gen Comp Endocrinol 122:270–286.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. https://doi.org/10.1186/1471-2105-5-113

Edwards JR, Yarychkivska O, Boulard M, Bestor TH (2017) DNA methylation and DNA methyltransferases. Epigenetics Chromatin 10:1–10

El-Leithy AAA, Hemeda SA, El Naby WSHA, et al (2019) Optimum salinity for Nile tilapia (*Oreochromis niloticus*) growth and mRNA transcripts of ion-regulation, inflammatory, stress- and immune-related genes. Fish Physiol Biochem 45:1217–1232. https://doi.org/10.1007/s10695-019-00640-7

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*(5), 1–10. doi: 10.1371/journal.pone.0019379.

Ellis, E.A., Storer, C.G., Kawahara, A.Y., 2021. De novo genome assemblies of butterflies. Gigascience 10, 1–8. https://doi.org/10.1093/gigascience/giab041

Evenhuis, J.P., Leeds, T.D., Marancik, D.P., Lapatra, S.E., Wiens, G.D., 2015. Rainbow trout (Oncorhynchus mykiss) resistance to columnaris disease is heritable and favorably correlated with bacterial cold water disease resistance. J. Anim. Sci. 93, 1546–1554. https://doi.org/10.2527/jas.2014-8566.

**-F-**

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to Quantitative Genetics, 4th Edition. Longman.

Farias, T., César, J., Silva, L., 2017. Methods of Selection Using the Quantitative Genetics in Aquaculture. Insights in Aquaculture and Biotechnology. iMedPub 1, 1–8.

FAO, 2019. FAO Yearbook of Fishery and Aquaculture Statistics 2017. FAO, Rome.

Feng, S., Cokus, S.J., Zhang, X., Chen, P.Y., Bostick, M., Goll, M.G., Hetzel, J., Jain, J., Strauss, S.H., Halpern, M.E., Ukomadu, C., Sadler, K.C., Pradhan, S., Pellegrini, M., Jacobsen, S.E., 2010. Conservation and divergence of methylation patterning in plants and animals. Proc. Natl. Acad. Sci. U. S. A. 107, 8689–8694. https://doi.org/10.1073/pnas.1002720107

Falconer, D.S., Mackay, T.F.C. (1995). Introduction to Quantitative Genetics. *Essex, England: Longman*.

Fagbenro, O.A., Arowosoge, I.A., 1991. Growth response and nutrient digestability by *Clarias isheriensis* (Sydenham, 1980) fed varying levels of dietary coffee pulp as replacement for maize in low-cost diets. Bioresour. Technol. 37, 253–258. https://doi.org/10.1016/0960-8524(91)90192-M

Falconer, D.S., 1990. Selection in different environments: Effects on environmental sensitivity (reaction norm) and on mean performance. Genet. Res. 56, 57–70. https://doi.org/10.1017/S001667 2300028883

Fellous A, Labed-Veydert T, Locrel M, et al (2018) DNA methylation in adults and during development of the self-fertilizing mangrove rivulus, *Kryptolebias marmoratus*. Ecol Evol 8:6016–6033. https://doi.org/10.1002/ece3.4141

Feng S, Cokus SJ, Zhang X, et al (2010) Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci 107:8689–8694

Fox BK, Breves JP, Davis LK, et al (2010) Tissue-specific regulation of the growth hormone/insulin-like growth factor axis during fasting and re-feeding: importance of muscle expression of IGF-I and IGF-II mRNA in the tilapia. Gen Comp Endocrinol 166:573–580

Fu B, Yu X, Tong J, et al (2019) Comparative transcriptomic analysis of hypothalamus-pituitary-liver axis in bighead carp (*Hypophthalmichthys nobilis*) with differential growth rate. BMC Genomics 20:1–8

Fuks F, Hurd PJ, Wolf D, et al (2003) The Methyl-CpG-binding Protein MeCP2 Links DNA Methylation to Histone Methylation*. J Biol Chem 278:4035–4040.

Fisher, P. J., Malthus, B., Walker, M. C., Corbett, G., & Spelman, R. J. (2009). The number of single nucleotide polymorphisms and on-farm data required for whole-herd parentage testing in dairy cattle herds. *Journal of Dairy Science*, *92*(1), 369–374. doi: 10.3168/jds.2008-1086

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., & Loeb, L. A. (2014). Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl*, 61–70. doi: doi:10.4172/jngsa.1000106

Flanagan, S.P., Forester, B.R., Latch, E.K., Aitken, S.N., Hoban, S., 2018. Guidelines for planning genomic assessment and monitoring of locally adaptive variation to inform species conservation. Evol. Appl. 11, 1035–1052. https://doi.org/10.1111/eva.12569

Fuzzen, M., Bernier, N., Kraak, G., 2011. Stress and reproduction. In: Norris DO, Lopez KH (eds) Hormones and Reproduction of Vertebrates. Acad. Press vol 1, 103–118.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F., 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. 117, 9451–9457. https://doi.org/10.1073/pnas.1921046117

**-G-**

Gao, Z., You, X., Zhang, X., Chen, J., Xu, T., Huang, Y., Lin, X., Xu, J., Bian, C., Shi, Q., 2021. A chromosome-level genome assembly of the striped catfish (*Pangasianodon hypophthalmus*). Genomics 113, 3349–3356. https://doi.org/10.1016/j.ygeno.2021.07.026

Gavery, M.R., Roberts, S.B., 2017. Epigenetic considerations in aquaculture. PeerJ 2017. https://doi.org/10.7717/peerj.4147

Gavery, M.R., Roberts, S.B., 2013. Predominant intragenic methylation is associated with gene expression characteristics in a bivalve mollusc. PeerJ 2013, 1–15. https://doi.org/10.7717/peerj.215

Gao, G., Magadan, S., Waldbieser, G.C., Youngblood, R.C., Wheeler, P.A., Scheffler, B.E., Thorgaard, G.H., Palti, Y., 2021. A long reads-based de-novo assembly of the genome of the Arlee homozygous line reveals chromosomal rearrangements in rainbow trout. G3 Genes, Genomes, Genet. 11. https://doi.org/10.1093/g3journal/jkab052 .

Gao, Z., You, X., Zhang, X., Chen, J., Xu, T., Huang, Y., Lin, X., Xu, J., Bian, C., Shi, Q., 2021. A chromosome-level genome assembly of the striped catfish (*Pangasianodon hypophthalmus*). Genomics 113, 3349–3356. https://doi.org/10.1016/j.ygeno.2021.07.026

Gavrielatos, M., Kyriakidis, K., Spandidos, D.A., Michalopoulos, I., 2021. Benchmarking of next and third generation sequencing technologies and their associated algorithms for de novo genome assembly. Mol. Med. Rep. 23, 1–12. https://doi.org/10.3892/mmr.2021.11890

Georges, M., Charlier, C., Hayes, B., 2019. Harnessing genomic information for livestock improvement. Nat. Rev. Genet. 20, 135–156. https://doi.org/10.1038/s41576-018-0082-2

GEV, 2020. Instructions for Visible Implant Elastomer ( VIE ) Tags. Northwest Marine Technology, Inc. GEV7. 1-5. www.nmt.us.

Gilles, B., Patrick, P., 2001. How should salinity influence fish growth? Comp. Biochem. Physiol. - C Toxicol. Pharmacol. 130, 411–423. https://doi.org/10.1016/S1532-0456(01)00268-X

Gjedrem, T., Rye, M., 2018. Selection response in fish and shellfish: a review. Rev. Aquac. 10, 168–179. https://doi.org/10.1111/raq.12154

Gjedrem, T., Robinson, N., Rye, M., 2012. The importance of selective breeding in aquaculture to meet future demands for animal protein: A review. Aquaculture 350–353, 117–129.

Gjedrem, T., Rye, M., 2016. Selection response in fish and shellfish: a review. Rev. Aquac. 10, 168–179. https://doi.org/10.1111/raq.12154

Guimarães, C.F.M., Mársico, E.T., Monteiro, M.L.G., Lemos, M., Mano, S.B., Conte Junior, C.A., 2016. The chemical quality of frozen Vietnamese Pangasius hypophthalmus fillets. Food Sci. Nutr. 4, 398–408. https://doi.org/10.1002/fsn3.302

Gu, X.H., Jiang, D.L., Huang, Y., Li, B.J., Chen, C.H., Lin, H.R., Xia, J.H., 2018. Identifying a Major QTL Associated with Salinity Tolerance in Nile Tilapia Using QTL-Seq. Mar. Biotechnol. 20, 98–107. https://doi.org/10.1007/s10126-017-9790-4

Ghurye, J., Rhie, A., Walenz, B.P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A.M., Koren, S., 2019. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput. Biol. 15, 1–19. https://doi.org/10.1371/journal.pcbi.1007273

Griot, R., Allal, F., Brard-Fudulea, S., Morvezen, R., Haffray, P., Phocas, F., Vandeputte, M., 2020. APIS: An auto-adaptive parentage inference software that tolerates missing parents. Mol. Ecol. Resour. 20, 579–590. https://doi.org/10.1111/1755-0998.13103

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

Gutierrez, A.P., Yáñ Ez, J.M., Fukui, S., Swift, B., Davidson, W.S., 2015. Genome-Wide association study (GWAS) for growth rate and age at sexual maturation in atlantic salmon (Salmo salar). PLoS One 10, 1–15. https://doi.org/10.1371/journal.pone.0119730

Geiman TM, Muegge K (2010) DNA methylation in early development. Mol Reprod Dev Inc Gamete Res 77:105–113

Goodman JI, Watson RE (2002) Altered DNA methylation: a secondary mechanism involved in carcinogenesis. Annu Rev Pharmacol Toxicol 42:501–525

Grosell M (2006) Intestinal anion exchange in marine fish osmoregulation. J Exp Biol 209:2813–2827. https://doi.org/10.1242/jeb.02345

Gu T-P, Guo F, Yang H, et al (2011) The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature 477:606–610. https://doi.org/10.1038/nature10443

Guo F, Li X, Liang D, et al (2014) Active and Passive Demethylation of Male and Female Pronuclear DNA in the Mammalian Zygote. Cell Stem Cell 15:447–459. https://doi.org/https://doi.org/10.1016/j.stem. 2014.08.003

Geng, X., Sha, J., Liu, S., Bao, L., Zhang, J., Wang, R., Yao, J., Li, C., Feng, J., Sun, F., Sun, L., Jiang, C., Zhang, Y., Chen, A., Dunham, R., Zhi, D., Liu, Z., 2015. A genome-wide association study in catfish reveals the presence of functional hubs of related genes within QTLs for columnaris disease resistance. BMC Genomics 16, 1–12. https://doi.org/10.1186/s12864-015-1409-4

Gjedrem, T., 2012. Genetic improvement for the development of efficient global aquaculture: A personal opinion review. Aquaculture 344–349, 12–22. https://doi.org/10.1016/j.aquaculture.-2012.03.003

Gjedrem, T., 2005. Selection and Breeding Programs in Aquaculture.

Gjedrem, T., 1997. Selective breeding to improve aquaculture production. World Aquac. 28, 33–45.

Gjedrem, T., Baranski, M., 2011. Selective breeding in aquaculture: An introduction, Methods and Technologies in Fish Biology and Fisheries 10.

Gjerde, B., Korsvoll, A., 1999. Realised selection differentials for growth rate and early sexual maturity in Atlantic salmon. Aquac. Eur. 99, 73–74.

Gitterle, T., Rye, M., Salte, R., Cock, J., Johansen, H., Lozano, C., Arturo Suárez, J., Gjerde, B., 2005. Genetic (co)variation in harvest body weight and survival in Penaeus (Litopenaeus) vannamei under standard commercial conditions. Aquaculture 243, 83–92. https://doi.org/10.1016/j.aquaculture.2004.10.015

Glaubitz, J.C., Rhodes, O.E., Dewoody, J.A., 2003. Prospects for inferring pairwise relationships with single nucleotide polymorphisms. Mol. Ecol. 12, 1039–1047. https://doi.org/10.1046/j.1365-294X.2003.01790.x

Goddard, M.E., Whitelaw, E., 2014. The use of epigenetic phenomena for the improvement of sheep and cattle. Front. Genet. 5, 1–6. https://doi.org/10.3389/fgene.2014.00247

Goldstein, D.B., Tterer, C.S., 1999. Microsatellites Evolution and Applications. Heredity (Edinb). 83, 633–635. https://doi.org/10.1177/0261018311403863

Gonzalez-Pena, D., Leeds, T.D., Gao, G., Cleveland, B., Kenney, B., Vallejo, R.L., Palti, Y., 2016. Genome-Wide Association Study for Identifying Genome Loci That Affect Fillet Yield in Rainbow Trout, Plant and Animal Genome XXIV Conference. Plant and Animal Genome. https://doi.org/https://pag.confex.com/pag/xxiv/webprogram/Paper18657.html

González-Recio, O., Toro, M.A., Bach, A., 2015. Past, present, and future of epigenetics applied to livestock breeding. Front. Genet. 6, 1–5. https://doi.org/10.3389/fgene.2015.00305

Gorjanc, G., Cleveland, M.A., Houston, R.D., Hickey, J.M., 2015. Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genet. Sel. Evol. 47. https://doi.org/10.1186/s12711-015-0102-z

Granada, L., Lemos, M.F.L., Cabral, H.N., Bossier, P., Novais, S.C., 2018. Epigenetics in aquaculture – the last frontier. Rev. Aquac. 10, 994–1013. https://doi.org/10.1111/raq.12219

**-H-**

Ha, H.P., Nguyen, T.T.T., Poompuang, S., Na-Nakorn, U., 2009. Microsatellites revealed no genetic differentiation between hatchery and contemporary wild populations of striped catfish, Pangasianodon hypophthalmus (Sauvage 1878) in Vietnam. Aquaculture 291, 154–160.

Hai, T.N., Nguyen, P.T., Van, H.N., Viet, L.Q., 2020. Promoting coastal aquaculture for adaptation to climate change and saltwater intrusion in the Mekong Delta, Vietnam. World Aquac. 19–26.

Hamzah, A., Mekkawy, W., Khaw, H.L., Nguyen, N.H., Yee, H.Y., Abu Bakar, K.R., Mohd Nor, S.A., Ponzoni, R.W., 2017. Genetic parameters for survival during the grow-out period in the GIFT strain of Nile tilapia (*Oreochromis niloticus*) and correlated response to selection for harvest weight. Aquac. Res. 48, 47–55. https://doi.org/10.1111/are.12859

Hasan, M.M., DeFaveri, J., Kuure, S., Dash, S.N., Lehtonen, S., Merilä, J., McCairns, R.J.S., 2017. Sticklebacks adapted to divergent osmotic environments show differences in plasticity for kidney morphology and candidate gene expression. J. Exp. Biol. 220, 2175–2186. https://doi.org/10.1242 /jeb.146027

Ha, T.T.P., van Dijk, H., Bosma, R., Sinh, L.X., 2013. Livelihood capabilities and pathways of shrimp farmers in the Mekong Delta, vietnam. Aquac. Econ. Manag. 17, 1–30. https://doi.org/10.1080/ 13657305.2013.747224

Hai, D.M., Yen, D.T., Liem, P.T., Tam, B.M., Huong, D.T.T., Hang, B.T.B., Hieu, D.Q., Garigliany, M.-M., Coppieters, W., Kestemont, P., Phuong, N.T., Farnir, F., 2022. A High-Quality Genome Assembly of Striped Catfish (Pangasianodon hypophthalmus) Based on Highly Accurate Long-Read HiFi Sequencing Data. Genes (Basel). 13, 923. https://doi.org/10.3390/genes13050923

Hasan, M.R., Shipton, T.A., 2021. Aquafeed value chain analysis of striped catfish in Vietnam. Aquaculture 541, 736798. https://doi.org/10.1016/j.aquaculture.2021.736798

Henriksson, P.J.G., Heijungs, R., Dao, H.M., Phan, L.T., De, G.R., 2015a. Product Carbon Footprints and Their Uncertainties in Comparative Decision Contexts 1–11.

Henriksson, P.J.G., Rico, A., Zhang, W., Ahmad-Al-Nahid, S., Newton, R., Phan, L.T., Zhang, Z., Jaithiang, J., Dao, H.M., Phu, T.M., Little, D.C., Murray, F.J., Satapornvanit, K., Liu, L., Liu, Q., Haque, M.M., Kruijssen, F., De Snoo, G.R., Heijungs, R., Van Bodegom, P.M., Guin??e, J.B., 2015b. Comparison of Asian Aquaculture Products by Use of Statistically Supported Life Cycle Assessment. Environ. Sci. Technol. 49, 14176–14183. https://doi.org/10.1021/acs.est.5b04634

Hieu, D.Q., Hang, B.T.B., Lokesh, J., Garigliany, M.-M., Huong, D.T.T., Yen, D.T., Liem, P.T., Tam, B.M., Hai, D.M., Son, V.N., Phuong, N.T., Farnir, F., Kestemont, P., 2022. Salinity significantly affects intestinal microbiota and gene expression in striped catfish juveniles. Appl. Microbiol. Biotechnol. 106, 3245–3264. https://doi.org/10.1007/s00253-022-11895-1

Hollenbeck, C.M., Johnston, I.A., 2018. Genomic tools and selective breeding in molluscs. Front. Genet. 9, 1–15. https://doi.org/10.3389/fgene.2018.00253

Houston, R.D., Bean, T.P., Macqueen, D.J., Gundappa, M.K., Jin, Y.H., Jenkins, T.L., Selly, S.L.C., Martin, S.A.M., Stevens, J.R., Santos, E.M., Davie, A., Robledo, D., 2020. Harnessing genomics to fast-track genetic improvement in aquaculture. Nat. Rev. Genet. 21, 389–409. https://doi.org/10.1038/s41576-020-0227-y

Hohn, C., Petrie-Hanson, L., 2013. Evaluation of visible implant elastomer tags in zebrafish (*Danio rerio*). Biol. Open 2, 1397–1401. https://doi.org/10.1242/bio.20136460

Hossain, F., Islam, S.M.M., Ashaf-Ud-Doulah, M., Ali, M.S., Islam, M.S., Brown, C., Shahjahan, M., 2021. Influences of Salinity on Embryonic and Larval Development of Striped Catfish *Pangasianodon hypophthalmus*. Front. Mar. Sci. 8, 1–10. https://doi.org/10.3389/fmars.2021.781951

Huong, D.T.T., Quyen, N.T., 2012. Effect of salinity on embryo development and osmoregulation of striped catfish (*Pangasianodon hypophthalmus*) larvae and fingerlings. Can Tho Univ. J. Sci. 21b, 29–37.

Harrison, H.B., Saenz-Agudelo, P., Planes, S., Jones, G.P., Berumen, M.L., 2013. Relative accuracy of three common methods of parentage analysis in natural populations. Mol. Ecol. 22, 1158–1170. https://doi.org/10.1111/mec.12138

Hauser, L., Baird, M., Hilborn, R.A.Y., Seeb, L.W., 2011. An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon ( Oncorhynchus nerka ) population 11, 150–161. https://doi.org/10.1111/j.1755-0998.2010.02961.x

Hauser, M.-T., Aufsatz, W., Jonak, C., Luschnig, C., 2011. Transgenerational epigenetic inheritance in plants. Biochim Biophys Acta 1809, 459–468. https://doi.org/10.1016/j.bbagrm.2011.03.007

Hoa, T.T.T., Boerlage, A.S., Duyen, T.T.M., Thy, D.T.M., Hang, N.T.T., Humphry, R.W., Phuong, N.T., 2021. Nursing stages of striped catfish (*Pangasianodon hypophthalmus*) in Vietnam: Pathogens, diseases and husbandry practices. Aquaculture 533, 736114. https://doi.org/10.1016/j

Huddleston, J., Chaisson, M.J.P., Steinberg, K.M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T.A., Munson, K.M., Kronenberg, Z.N., Vives, L., Peluso, P., Boitano, M., Chin, C.S., Korlach, J., Wilson, R.K., Eichler, E.E., 2017. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 27, 677–685. https://doi.org/10.1101/gr.214007.116

Hammond SA, Nelson CJ, Helbing CC (2015) Environmental influences on the epigenomes of herpetofauna and fish. Biochem Cell Biol 94:95–100. https://doi.org/10.1139/bcb-2015-0111

Head JA, Mittal K, Basu N (2014) Application of the LUminometric Methylation Assay to ecological species: tissue quality requirements and a survey of DNA methylation levels in animals. Mol Ecol Resour 14:943–952. https://doi.org/10.1111/1755-0998.12244

Hieu DQ, Hang BTB, Huong DTT, et al (2021) Salinity affects growth performance, physiology, immune responses and temperature resistance in striped catfish (*Pangasianodon hypophthalmus*) during its early life stages. Fish Physiol Biochem 47:1995–2013. https://doi.org/10.1007/s10695-021-01021-9

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., et al., The zebrafish reference genome sequence and its relationship to the human genome. Nature 496, 498–503. https://doi.org/10.1038/nature12111

Hoffman, J. I., & Amos, W. (2005). Microsatellite genotyping errors: Detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, *14*(2), 599–612. doi: 10.1111/j.1365-294X.2004.02419.x

Holman, L. E., Garcia, D., Onoufriou, A., Hillestad, B., & Johnston, I. A. (2017). A work fl ow used to design low density SNP panels for parentage assignment and traceability in aquaculture species and its validation in Atlantic salmon. *Aquaculture*, *476*(December 2016), 59–64.

Hu, J., Fan, J., Sun, Z., Liu, S., 2020. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics 36, 2253–2255. https://doi.org/10.1093/bioinformatics/btz891

Herbinger, C.M., Doyle, R.W., Pitman, E.R., Paquet, D., Mesa, K.A., Morris, D.B., Wright, J.M., Cook, D., 1995. DNA fingerprint based analysis of paternal and maternal effects on offspring growth and survival in communally reared rainbow trout. Aquaculture 137, 245–256. https://doi.org/10.1016/0044-8486(95)01109-9

Herbinger, C.M., O'Reilly, P.T., Doyle, R.W., Wright, J.M., O'Flynn, F., 1999. Early growth performance of Atlantic salmon full-sib families reared in single family tanks versus in mixed family tanks. Aquaculture 173, 105–116. https://doi.org/10.1016/S0044-8486(98)00479-7.

Herlin, M., Taggart, J.B., McAndrew, B.J., Penman, D.J., 2007. Parentage allocation in a complex situation: A large commercial Atlantic cod (*Gadus morhua*) mass spawning tank. Aquaculture 272, S195–S203. https://doi.org/10.1016/j.aquaculture.2007.08.018

Hettinger, A., Sanford, E., Hill, T.M., Russell, A.D., Sato, K.N.S., Hoey, J., Forsch, M., Page, H.N., Gaylord, B., 2012. Persistent carry-over effects of planktonic exposure to ocean acidification in the Olympia oyster. Ecology 93, 2758–2768. https://doi.org/10.1890/12-0567.1

Holman, L.E., Garcia, D., Onoufriou, A., Hillestad, B., Johnston, I.A., 2017. A work fl ow used to design low density SNP panels for parentage assignment and traceability in aquaculture species and its validation in Atlantic salmon. Aquaculture 476, 59–64. https://doi.org/10.1016/j.aquaculture.2017.04.001

Hong Xia, J., Bai, Z., Meng, Z., Zhang, Y., Wang, L., Liu, F., Jing, W., Yi Wan, Z., Li, J., Lin, H., Hua Yue, G., 2015. Signatures of selection in tilapia revealed by whole genome resequencing. Sci. Rep. 5, 1–10. https://doi.org/10.1038/srep14168

Huisman, J., 2017. Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond. Mol. Ecol. Resour. 17, 1009–1024. https://doi.org/10.1111/1755-0998.12665

Hulata, G., Wohlfarth, G.W., Halevy, A., 1986. Mass selection for growth rate in the Nile tilapia (*Oreochromis niloticus*). Aquaculture 57, 177–184. https://doi.org/10.1016/0044-8486(86)90195-X

Huong, D.T.T., Phuong, N.T., Bayley, M., Wang, T., Thuy, N.H., 2009. Osmotic and ionic regulation, food intake and growth of striped catfish (Pangasianodon hypophthalmus) and sand goby (*Oxyeleotris marmoratus*) exposed to different salinities. Comp. Biochem. Physiol. Part A Mol. Integr. Physiol. 153, S144. https://doi.org/10.1016/j.cbpa.2009.04.272

Huong, D.T.T., Quyen, N.T., 2012. Effect of salinity on embryo development and osmoregulation of striped catfish (Pangasianodon hypophthalmus) larvae and fingerlings. Can Tho Univ. J. Sci. 21b, 29–37.

**-I-**

IPCC, 2007. Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK. 976pp.

Iida A, Shimada A, Shima A, et al (2006) Targeted reduction of the DNA methylation level with 5-azacytidine promotes excision of the medaka fish Tol2 transposable element. Genet Res (Camb) 87:187–193

Inoue A, Shen L, Dai Q, et al (2011) Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. Cell Res 21:1670–1676. https://doi.org/10.1038/cr.2011.189

Inoue A, Zhang Y (2011) Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. Science 334:194. https://doi.org/10.1126/science.1212483

**-J-**

Jabbari, K., Cacciò, S., Païs de Barros, J.P., Desgrès, J., Bernardi, G., 1997. Evolutionary changes in CpG and methylation levels in the genome of vertebrates. Gene 205, 109–118. https://doi.org/10.1016/S0378-1119(97)00475-7

Jablonka, E., Lamb, M.J., 2006. The Changing Concept of Epigenetics. Ann. N. Y. Acad. Sci. 981, 82–96. https://doi.org/10.1111/j.1749-6632.2002.tb04913.x

Jaspe, C.J., Caipang, C.M.A., 2011. Increasing salinity tolerance in tilapias: Selective breeding using locally available strains. AACL Bioflux 4, 437–441.

Jähner D, Stuhlmann H, Stewart CL, et al (1982) De novo methylation and expression of retroviral genomes during mouse embryogenesis. Nature 298:623–628

Jiang L, Zhang J, Wang J-J, et al (2013) Sperm, but Not Oocyte, DNA Methylome Is Inherited by Zebrafish Early Embryos. Cell 153:773–784. https://doi.org/https://doi.org/10.1016/j.cell.2013.04.041

Jinsuk K, Matthias L, A. PW, et al (2015) Simultaneous deletion of the methylcytosine oxidases Tet1 and Tet3 increases transcriptome variability in early embryogenesis. Proc Natl Acad Sci 112:E4236–E4245. https://doi.org/10.1073/pnas.1510510112

Jones PL, Veenstra GJC, Wade PA, et al (1998) Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat Genet 19:187–191. https://doi.org/10.1038/561

Jerry, D.R., Purvis, I.W., Piper, L.R., 1997. Opportunities for genetic improvement in crustacean species. Proc. Assoc. Advmt. Anim. Breed. Genet. 14, 55–58.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H.E., Pedersen, B.S., Rhie, A., Richardson, H., Quinlan, A.R., Snutch, T.P., Tee, L., Paten, B., Phillippy, A.M., Simpson, J.T., Loman, N.J., Loose, M., 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. 36, 338–345. https://doi.org/10.1038/nbt.4060

Jiang, Q., Li, Q., Yu, H., Kong, L., 2016. Inheritance and Variation of Genomic DNA Methylation in Diploid and Triploid Pacific Oyster (*Crassostrea gigas*). Mar. Biotechnol. 18, 124–132. https://doi.org/10.1007/s10126-015-9674-4

Jones, P.A., 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. 13, 484–492. https://doi.org/10.1038/nrg3230

Jones, A. G., Small, C. M., Paczolt, K. A., & Ratterman, N. L. (2010). A practical guide to methods of parentage analysis. *Molecular Ecology Resources*, *10*(1), 6–30. doi: 10.1111/j.1755-0998.2009 .02778.x

Jonsson, B., Jonsson, N., 2014. Early environment influences later performance in fishes. J. Fish Biol. 85, 151–188. https://doi.org/10.1111/jfb.12432

**-K-**

Kalinowski, S. T., Taper, M. L., & Marshall, T. C. (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, *16*(5), 1099–1106. doi: 10.1111/j.1365-294X.2007.03089.x

Kader, T., Goode, D.L., Wong, S.Q., Connaughton, J., Rowley, S.M., Devereux, L., Byrne, D., Fox, S.B., Mir Arnau, G., Tothill, R.W., Campbell, I.G., Gorringe, K.L., 2016. Copy number analysis by low coverage whole genome sequencing using ultra low-input DNA from formalin-fixed paraffin embedded tumor tissue. Genome Med. 8, 121. https://doi.org/10.1186/s13073-016-0375-z.

Kass SU, Pruss D, Wolffe AP (1997) How does DNA methylation repress transcription? Trends Genet 13:444–449

Kimura H, Shiota K (2003) Methyl-CpG-binding Protein, MeCP2, Is a Target Molecule for Maintenance DNA Methyltransferase, Dnmt1*. J Biol Chem 278:4806–4812. https://doi.org/https:// doi.org/ 10.1074/jbc.M209923200

Kardos, M., Taylor, H.R., Ellegren, H., Luikart, G., Allendorf, F.W., 2016. Genomics advances the study of inbreeding depression in the wild. Evol. Appl. 9, 1205–1218. https://doi.org/10.1111/eva.12414

Kim, O.T.P., Nguyen, P.T., Shoguchi, E., Hisata, K., Vo, T.T.B., Inoue, J., Shinzato, C., Le, B.T.N., Nishitsuji, K., Kanda, M., Nguyen, V.H., Nong, H. V., Satoh, N., 2018. A draft genome of the striped catfish, Pangasianodon hypophthalmus, for comparative analysis of genes relevant to development and a resource for aquaculture improvement 06 Biological Sciences 0604 Genetics. BMC Genomics 19, 1–16. https://doi.org/10.1186/s12864-018-5079-x

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

Krishna, G., Gopikrishna, G., Gopal, C., Jahageerdar, S., Ravichandran, P., Kannappan, S., Pillai, S.M., Paulpandi, S., Kiran, R.P., Saraswati, R., Venugopal, G., Kumar, D., Gitterle, T., Lozano, C., Rye, M., Hayes, B., 2011. Genetic parameters for growth and survival in Penaeus monodon cultured in India. Aquaculture 318, 74–78. https://doi.org/10.1016/j.aquaculture.2011.04.028

Kulski, J.K., 2012. Next-Generation Sequencing — An Overview of the History, Tools, and “Omic” Applications, in: Intech. p. 13. https://doi.org/http://dx.doi.org/10.5772/61964

Kumar, P., Janmejay, P., 2021. Advances in Fisheries Biotechnology, Advances in Fisheries Biotechnology. https://doi.org/10.1007/978-981-16-3215-0

Kwak, S.M., Myung, S.K., Lee, Y.J., Seo, H.G., 2012. Efficacy of omega-3 fatty acid supplements (eicosapentaenoic acid and docosahexaenoic acid) in the secondary prevention of cardiovascular disease: A meta-analysis of randomized, double-blind, placebo-controlled trials. Arch. Intern. Med. 172, 686–694. https://doi.org/10.1001/archinternmed.2012.262

Kruuk, L. E. B., & Hill, W. G. (2008). Introduction. Evolutionary dynamics of wild populations: The use of long-term pedigree data. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1635), 593–596. doi: 10.1098/rspb.2007.1689

Kulski, J. K. (2012). Next-Generation Sequencing — An Overview of the History, Tools, and “Omic” Applications. In *Intech* (p. 13). doi: http://dx.doi.org/10.5772/61964

**-L-**

Lacal, I., Ventura, R., 2018. Epigenetic Inheritance: Concepts, Mechanisms and Perspectives. Front. Mol. Neurosci. 11, 1–22. https://doi.org/10.3389/fnmol.2018.00292

Lam, N.C., Huong, D.T.T.H., Son, N.V., Phuong, T.N., 2011. Effects of salinity on physiology and growth of striped catfish (*Pangasisanodon hypophthalmus)*. Can Tho Univ. J. Sci. 17a, 60–69.

Lapègue, S., Harrang, E., Heurtebise, S., Flahauw, E., Donnadieu, C., Gayral, P., Ballenghien, M., Genestout, L., Barbotte, L., Mahla, R., Haffray, P., Klopp, C., 2014. Development of SNP-genotyping arrays in two shellfish species. Mol. Ecol. Resour. 14, 820–830. https://doi.org/10.1111/1755-0998.12230

Le Dang, H., Li, E., Bruwer, J., Nuberg, I., 2014. Farmers' perceptions of climate variability and barriers to adaptation: Lessons learned from an exploratory study in Vietnam. Mitig. Adapt. Strateg. Glob. Chang. 19, 531–548. https://doi.org/10.1007/s11027-012-9447-6

Legendre, M., Slembrouck, J., Subagja, J., Hari Kristanto, A., 2000. Ovulation rate, latency period and ova viability after GnRH-or hCG-induced breeding in the Asian catfish Pangasius hypophthalmus (*Siluriformes, Pangasiidae*). Aquat. Living Resour. 13, 145–151. https://doi.org/10.1016/S0990-7440(00)00148-0

Liem, P.T., Tam, B.M., Phuong, N.T., Duc, P.M., 2017. Developing and operating a demonstration grow-out farm for sustainable and advanced pangasius aquaculture in vietnam (vidatec-tender).

Lundh, A., 2020. Blue Biotechnology : Its role in the future of food.

Legault L-M, Bertrand-Lehouillier V, McGraw S (2018) Pre-implantation alcohol exposure and developmental programming of FASD: an epigenetic perspective. Biochem cell Biol 96:117–130

Li E, Beard C, Jaenisch R (1993) Role for DNA methylation in genomic imprinting. Nature 366:362–365. https://doi.org/10.1038/366362a0

Li S, He F, Wen H, et al (2017) Low salinity affects cellularity, DNA methylation, and mRNA expression of igf1 in the liver of half smooth tongue sole (*Cynoglossus semilaevis*). Fish Physiol Biochem 43:1587–1602

Leeds, T.D., Silverstein, J.T., Weber, G.M., Vallejo, R.L., Palti, Y., Rexroad, C.E., Evenhuis, J., Hadidi, S., Welch, T.J., Wiens, G.D., 2010. Response to selection for bacterial cold water disease resistance in rainbow trout. J. Anim. Sci. 88, 1936–1946. https://doi.org/10.2527/jas.2009-2538

Le Boucher, R., Dupont-Nivet, M., Vandeputte, M., Kerneïs, T., Goardon, L., Labbé, L., Chatain, B., Bothaire, M.J., Larroquet, L., Médale, F., Quillet, E., 2012. Selection for Adaptation to Dietary Shifts: Towards Sustainable Breeding of Carnivorous Fish. PLoS One 7, 3–9. https://doi.org/10.1371/ journ

Liu, S., Palti, Y., Gao, G., Rexroad, C.E., 2016. Development and validation of a SNP panel for parentage assignment in rainbow trout. Aquaculture 452, 178–182. https://doi.org/10.1016/j.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., Hvidsten, T.R., Leong, J.S., Minkley, D.R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R.A., et al., 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533, 200–205. https://doi.org/10.1038/nature17164

Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M., Pevzner, P.A., 2016. Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl. Acad. Sci. U. S. A. 113, E8396–E8405. https://doi.org/10.1073/pnas.1604560113

Liu, Z., Liu, S., Yao, J., Bao, L., Zhang, J., Li, Y., Jiang, C., Sun, L., Wang, R., Zhang, Y., Zhou, T., Zeng, Q., Fu, Q., Gao, S., Li, N., Koren, S., Jiang, Y., Kaltenboeck, L., Liu, H., Armbruster, J., Xie, Y., Kirby, M.L., Tian, Y., Flanagan, M.E., Mu, W., Waldbieser, G.C., 2016. The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. Nat. Commun. 7, 1–13. https://doi.org/10.1038/ncomms11757

Le Luyer, J., Laporte, M., Beacham, T.D., Kaukinen, K.H., Withler, R.E., Leong, J.S., Rondeau, E.B., Koop, B.F., Bernatchez, L., 2017. Parallel epigenetic modifications induced by hatchery rearing in a Pacific salmon. Proc. Natl. Acad. Sci. U. S. A. 114, 12964–12969. https://doi.org/10.1073/pnas. 1711229114

Lefevre, S., Huong, D.T.T., Ha, N.T.K., Wang, T., Phuong, N.T., Bayley, M., 2011. A telemetry study of swimming depth and oxygen level in a Pangasius pond in the Mekong Delta. Aquaculture 315, 410–413. https://doi.org/10.1016/j.aquaculture.2011.02.030

Lee, C. Y., Chiu, Y. C., Wang, L. B., Kuo, Y. L., Chuang, E. Y., Lai, L. C., & Tsai, M. H. (2013). Common applications of next-generation sequencing technologies in genomic research. *Translational Cancer Research*, *2*(1), 33–45. doi: 10.3978/j.issn.2218-676X.2013.02.09

Lewontin, R.C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49,** 49–67

Li, H., Wu, K., Ruan, C., Pan, J., Wang, Y., Long, H., 2019. Cost-reduction strategies in massive genomics experiments. Mar. Life Sci. Technol. 1, 15–21. https://doi.org/10.1007/s42995-019-00013-2

Li, Y., Sidore, C., Kang, H.M., Boehnke, M., Abecasis, G.R., 2011. Low-coverage sequencing: Implications for design of complex trait association studies. Genome Res. 21, 940–951. https://doi.org/10.1101/gr.117259.110

Lien, S., Gidskehaug, L., Moen, T., Hayes, B.J., Berg, P.R., Davidson, W.S., Omholt, S.W., Kent, M.P., 2011. A dense SNP-based linkage map for Atlantic salmon (Salmo salar) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. BMC Genomics 12, 1–10. https://doi.org/10.1186/1471-2164-12-615

Lind, C.E., Ponzoni, R.W., Nguyen, N.H., Khaw, H.L., 2012. Selective breeding in fish and conservation of genetic resources for aquaculture. Reprod. Domest. Anim. 47, 255–263.

Lischer, H.E.L., Shimizu, K.K., 2017. Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics 18, 1–12. https://doi.org/10.1186/s12859-017-1911-6.

Liu, F., Sun, F., Xia, J.H., Li, J., Fu, G.H., Lin, G., Tu, R.J., Wan, Z.Y., Quek, D., Yue, G.H., 2014. A genome scan revealed significant associations of growth traits with a major QTL and GHR2 in tilapia. Sci. Rep. 4, 1–9. https://doi.org/10.1038/srep07256

Liu, P., Xia, J.H., Lin, G., Sun, F., Liu, F., Lim, H.S., Pang, H.Y., Yue, G.H., 2012. Molecular Parentage Analysis Is Essential in Breeding Asian Seabass. PLoS One 7, 1–6.

Liu, S., Palti, Y., Gao, G., Rexroad, C.E., 2016. Development and validation of a SNP panel for parentage assignment in rainbow trout. Aquaculture 452, 178–182. https://doi.org/10.1016/j.aquaculture.2015.11.001

Liu, Z.J., Cordes, J.F., 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture 238, 1–37. https://doi.org/10.1016/j.aquaculture.2004.05.027

Logsdon, G.A., Vollger, M.R., Eichler, E.E., 2020. Long-read human genome sequencing and its applications. Nat. Rev. Genet. 21, 597–614. https://doi.org/10.1038/s41576-020-0236-x

Lu, J.Y., Shao, W., Chang, L., Yin, Y., Li, T., Zhang, H., Hong, Y., Percharde, M., Guo, L., Wu, Z., Liu, L., Liu, W., Yan, P., Ramalho-Santos, M., Sun, Y., Shen, X., 2020. Genomic Repeats Categorize Genes with Distinct Functions for Orchestrated Regulation. Cell Rep. 30, 3296-3311.e5. https://doi.org/10.1016/j.celrep.2020.02.048

Ludwig, A., Belfiore, N.M., Pitra, C., Svirsky, V., Jenneckens, I., 2001. Genome duplication events and functional reduction of ploidy levels in sturgeon (Acipenser, Huso and Scaphirhynchus). Genetics 158, 1203–1215. https://doi.org/10.1093/genetics/158.3.1203

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Liu, S., Zhou, Z., Lu, J., Sun, F., Wang, S., Liu, H., … Liu, Z. (2011). Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. *BMC Genomics*, *12*(53), 1471–2164. doi: 10.1186/1471-2164-12-53

Liu, W., Yang, T., Zhao, W. and Chase, G.A. (2007), Accounting for Genotyping Errors in Tagging SNP Selection. Annals of Human Genetics, 71: 467-479.

**-M-**

Marshall, T. C., Slate, J., Kruuk, L. E. B., & Pemberton, J. M. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, *7*(5), 639–655. doi: 10.1046/j.1365-294x.1998.00374.x

Macqueen, D.J., Primmer, C.R., Houston, R.D., Nowak, B.F., Naish, K.A., Nichols, K.M., Ólafsson, K., Omholt, S.W., Palti, Y., Plastow, G.S., Rexroad, C.E., Rise, M.L., Ritchie, R.J., Sandve, S.R., Schulte, P.M., Tello, A., Vidal, R., Vik, J.O., Wargelius, A., Yáñez, J.M., 2017. Functional Annotation of All Salmonid Genomes (FAASG): An international initiative supporting future salmonid research, conservation and aquaculture. BMC Genomics 18, 1–9. https://doi.org/10.1186/s12864-017-3862-8

Mair, G.C., 2007. Genetics and breeding in seed supply for inland aquaculture pp. 519-547. In: M.G. Bondad-Reantaso (ed.). Assessment of freshwater fish seed resources for sustainable aquaculture. FAO Fish. Tech. Pap. No. 501. Rome, FAO. 628p. 519–547.

Marandel, L., Lepais, O., Arbenoits, E., Véron, V., Dias, K., Zion, M., Panserat, S., 2016. Remodelling of the hepatic epigenetic landscape of glucose-intolerant rainbow trout (Oncorhynchus mykiss) by nutritional status and dietary carbohydrates. Sci. Rep. 6, 1–12. https://doi.org/10.1038/-srep32187

MARD, 2013. Decision No. No. 2760/QĐ-BNN-TCTS dated on November 22, 2013 by the Minister of Agriculture and Rural Development on re-structuring the aquaculture and fisheries to enhance the values and sustainable development (in Vietnamese).

Marklund, S., Ellegren, H., Eriksson, S., Sandberg, K., Andersson, L., Marklund, S., Ellegren, H., Eriksson, S., Sandberg, K., Andersson, L., 2009. Parentage testing and linkage analysis in the horse using a set of highly polymorphic microsatellites. Anim. Genet. 25, 19–23. https://doi.org/10.1111/j.-1365-2052.1994.tb00050.x

Malhis, N., Jones, S.J.M., 2010. High quality SNP calling using Illumina data at shallow coverage. Bioinformatics 26, 1029–1035. https://doi.org/10.1093/bioinformatics/btq092

Mackay, P., Russell, M., 2011. Socialist Republic of Viet Nam : Climate Change Impact and Adaptation Study in the Mekong Delta ( Cofinanced by the Climate Change Fund and the Government of Climate Change Impact and Adaptation Study in the Mekong Delta- Part A:Ca Mau Atlas 49.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., Church, G.M., 2013. RNA-guided human genome engineering via Cas9. Science (80-. ). 339, 823–826. https://doi.org/10.1126/science.1232033

Mathieu, L.M., Le Phu, V., Nicolas, G., 2021. What Is the Future of the Lower Mekong Basin Struggling against Human Activities? A Review, in: River Deltas - Recent Advances [Working Title]. IntechOpen, p. 13. https://doi.org/10.5772/intechopen.95010

Manly, B. F. J. (2007) Randomization, Bootstrap, and Monte Carlo Methods in Biology (3rd ed.), London: Chapman & Hall.

Mehrotra, S., Goyal, V., 2014. Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. Genomics, Proteomics Bioinforma. 12, 164–171. https://doi.org/10.101

Moen, T., Torgersen, J., Santi, N., Davidson, W.S., Baranski, M., Ødegård, J., Kjøglum, S., Velle, B., Kent, M., Lubieniecki, K.P., Isdal, E., Lien, S., 2015. Epithelial cadherin determines resistance to infectious pancreatic necrosis virus in Atlantic salmon. Genetics 200, 1313–1326. https://doi.org/10.

Molina-Mora, J.A., Campos-Sánchez, R., Rodríguez, C., Shi, L., García, F., 2020. High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 Pseudomonas aeruginosa genome: Benchmark of hybrid and non-hybrid assemblers. Sci. Rep. 10, 1–16. https://doi.org/10.1038/s41598-020-58319-6

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

Magdeldin S, Uchida K, Hirano T, et al (2007) Effects of environmental salinity on somatic growth and growth hormone/insulin-like growth factor-I axis in juvenile tilapia *Oreochromis mossambicus*. Fish Sci 73:1025–1034

Marsit CJ (2015) Influence of environmental exposure on human epigenetic regulation. J Exp Biol 218:71–79

Mayer E, Provine WB (1998) The evolutionary synthesis perspectives on the unification of biology, with a new preface

McCormick SD (1995) 11 hormonal control of gill Na+, K+-ATPase and chloride cell function. In: Fish physiology. Elsevier, pp 285–315

McLean E, Donaldson EM (1993) The role of growth hormone in the growth of poikilotherms. Endocrinol growth, Dev Metab Vertebr 43–71

Mhanni AA, McGowan RA (2004) Global changes in genomic methylation levels during early development of the zebrafish embryo. Dev Genes Evol 214:412–417. https://doi.org/10.1007/s00427-004-0418-0

Moghadam HK, Johnsen H, Robinson N, et al (2017) Impacts of early life stress on the methylome and transcriptome of Atlantic salmon. Sci Rep 7:1–11

Moore LD, Le T, Fan G (2013) DNA methylation and its basic function. Neuropsychopharma cology 38:23–38

Morey L, Santanach A, Di Croce L (2015) Pluripotency and epigenetic factors in mouse embryonic stem cell fate regulation. Mol Cell Biol 35:2716–2728

Moriyama S, Takahashi A, Hirano T, Kawauchi H (1990) Salmon growth hormone is transported into the circulation of rainbow trout, *Oncorhynchus mykiss*, after intestinal administration. J Comp Physiol B 160:251–257

Moriyama S, Yamamoto H, Sugimoto S, et al (1993) Oral administration of recombinant salmon growth hormone to rainbow trout, Oncorhynchus mykiss. Aquaculture 112:99–106

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., Gnirke, A., Jaenisch, R., Lander, E.S., 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454, 766–770.

Men, L.T., Thanh, V.C., Hirata, Y., Yamasaki, S., 2005. Evaluation of the genetic diversities and the nutritional values of the Tra (Pangasius hypophthalmus) and the Basa (Pangasius bocourti) catfish cultivated in the Mekong River Delta of Vietnam. Asian-Australasian J. Anim. Sci. 18, 671–676.

Mérot, C., Oomen, R.A., Tigano, A., Wellenreuther, M., 2020. A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. Trends Ecol. Evol. 35, 561–572.

Metzger, D.C.H., Schulte, P.M., 2016. Epigenomics in marine fishes. Mar. Genomics 30, 43–54. https://doi.org/10.1016/j.margen.2016.01.004

Moav, R., Wohlfarth, G., 1976. Two-way selection for growth rate in the common carp (*Cyprinus carpio L.*). Genetics 82, 83–101. https://doi.org/10.1093/genetics/82.1.83.

Moghadam, H., Mørkøre, T., Robinson, N., 2015. Epigenetics-potential for programming fish for aquaculture? J. Mar. Sci. Eng. 3, 175–192. https://doi.org/10.3390/jmse3020175

Morán, P., Marco-Rius, F., Megías, M., Covelo-Soto, L., Pérez-Figueroa, A., 2013. Environmental induced methylation changes associated with seawater adaptation in brown trout. Aquaculture 392–395, 77–83. https://doi.org/10.1016/j.aquaculture.2013.02.006

Morán, P., Pérez-Figueroa, A., 2011. Methylation changes associated with early maturation stages in the Atlantic salmon. BMC Genet. 12. https://doi.org/10.1186/1471-2156-12-86

Morga, B., Renault, T., Faury, N., Arzul, I., 2012. New insights in flat oyster Ostrea edulis resistance against the parasite Bonamia ostreae. Fish Shellfish Immunol. 32, 958–968.

Morgan, H.D., Sutherland, H.G.E., Martin, D.I.K., Whitelaw, E., 1999. Epigenetic inheritance at the agouti locus in the mouse. Nat. Genet. 23, 314–318. https://doi.org/10.1038/15490

Morris, A.J., Pollott, G.E., 1997. Comparison of selection based on phenotype, selection index and best linear unbiased prediction using data from a closed broiler line. Br. Poult. Sci. 38, 249–254. https://doi.org/10.1080/00071669708417981

Meagher, T. R., & Thompson, E. (1986). The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theoretical Population Biology*, *29*(1), 87–106. doi: 10.1016/0040-5809(86)90006-7

**-N-**

Nguyen, A.L., Dang, V.H., Bosma, R.H., Verreth, J.A.J., Leemans, R., De Silva, S.S., 2014. Simulated Impacts of Climate Change on Current Farming Locations of Striped Catfish (*Pangasianodon hypophthalmus; Sauvage*) in the Mekong Delta, Vietnam. Ambio 43, 1059–1068.

Nguyen, H.T.P., 2015. Effects of temperature and salinity on growth performance in cultured Tra Catfish (*Pangasianodon hypophthalmus*) in Vietnam (Doctoral dissertation). Science and Engineering Faculty, Queensland University of Technology, Brisbane, Australia. 18–22.

Nguyen, H.T.P., Do, T.T.H., Mather, P.B., Hurwood, D.A., 2014. Experimental assessment of the effects of sublethal salinities on growth performance and stress in cultured tra catfish (Pangasianodon hypophthalmus). Fish Physiol. Biochem. 40, 1839–1848. https://doi.org/10.1007/s10695-014-9972-1

Nguyen, N.H., 2016. Genetic improvement for important farmed aquaculture species with a reference to carp, tilapia and prawns in Asia: achievements, lessons and challenges. Fish Fish. 17, 483–506. https://doi.org/10.1111/faf.12122

Nguyen, T.K.H., Nguyen, T.E., Nguyen, M.N., Yasuaki, T., Nguyen, T.P., Do, T.T.H., 2021. Effects of salinity on growth performance, survival rate, digestive enzyme activities and physiological parameters of striped catfish (*Pangasianodon hypophthalmus*) at larval stage. Can Tho Univ. J. Sci. 13, 1–9. https://doi.org/10.22144/ctu.jen.2021.011

Nguyen, T.T.T., 2009. Patterns of use and exchange of genetic resources of the striped catfish *Pangasianodonâƒhypophthalmus* (Sauvage 1878). Rev. Aquac. 1, 224–231. https://doi.org/10.1111/j.1753-5131.2009.01016.x

Nguyen, T.T.T., Hayes, B.J., Ingram, B.A., 2014. Genetic parameters and response to selection in blue mussel (*Mytilus galloprovincialis*) using a SNP-based pedigree. Aquaculture 420–421, 295–301. https://doi.org/10.1016/j.aquaculture.2013.11.021

Nguyen, P.T., Bui, T.M., Nguyen, T.A., De Silva, S., 2013. Developments in hatchery technology for striped catfish (*Pangasianodon hypophthalmus*). Advances in Aquaculture Hatchery Technology. Woodhead Publ. Ltd. 498–518. https://doi.org/10.1533/9780857097460.3.498

Nguyen, A.L., Truong, M.H., Verreth, J.A.J., Leemans, R., Bosma, R.H., De Silva, S.S., 2015. Exploring the climate change concerns of striped catfish producers in the Mekong Delta, Vietnam. Springerplus 4, 1–8. https://doi.org/10.1186/s40064-015-0822-0

Nguyen, A.L., Vu, N.K., 2020. Mitigating the Impacts of Salinity Intrusion due to Climate Change on the Tra Catfish Farming in Coastal Provinces of the Mekong Delta , Viet Nam. Fish People 19, 32–37.

Nguyen, L.A., Pham, T.B.V., Bosma, R., Verreth, J., Leemans, R., De Silva, S., Lansink, A.O., 2018. Impact of Climate Change on the Technical Efficiency of Striped Catfish, *Pangasianodon hypophthalmus*, Farming in the Mekong Delta, Vietnam. J. World Aquac. Soc. 49, 570–581.

Nguyen, L.A., Verreth, J.A.J., Leemans, R., Bosma, R., De Silva, S., 2016. A decision tree analysis to support potential climate change adaptations of striped catfish (*Pangasianodon hypophthalmus Sauvage*) farming in the Mekong Delta, Vietnam. Tropicultura 34, 105–115.

Nguyen, N.H., Sonesson, A.K., Houston, R.D., Moghadam, H., 2022. Editorial: Applications of Modern Genetics and Genomic Technologies to Enhance Aquaculture Breeding. Front. Genet. 13, 10–13. https://doi.org/10.3389/fgene.2022.898857

Nhan, D.K., Be, N. V., Trung, N.H., 2007. Water use and competition in the Mekong Delta, Vietnam. Lit. Anal. Chanllenges to Sustain. Dev. Mekong Delta Reg. Natl. Policy Issues Res. Needs 143–188.

Ninh, N.H., Ponzoni, R.W., Nguyen, N.H., Woolliams, J.A., Taggart, J.B., McAndrew, B.J., Penman, D.J., 2013. A comparison of communal and separate rearing of families in selective breeding of common carp (*Cyprinus carpio*): Responses to selection. Aquaculture 408, 152–159.

Ninh, N.H., Thoa, N.P., Knibb, W., Nguyen, N.H., 2014. Selection for enhanced growth performance of Nile tilapia (*Oreochromis niloticus*) in brackish water (15-20ppt) in Vietnam. Aquaculture 428–429, 1–6. https://doi.org/10.1016/j.aquaculture.2014.02.024

Nath, S., Shaw, D.E., White, M.A., 2021. Improved contiguity of the threespine stickleback genome using long-read sequencing. G3 Genes, Genomes, Genet. 11. https://doi.org/10.1093/g3journal/jkab007

Nguyen, T.P., Dang, D.T.H., 2010. Striped catfish aquaculture in Vietnam: A decade of unprecedented development. In : S.S. De Silva and F.B. Davy (eds.). Success Stories in Asian Aquaculture. Springer Sci. Media B.V 131–147. https://doi.org/10.1007/978-90-481-3087-0_7

Nowak, R.M., Jastrzębski, J.P., Kuśmirek, W., Sałamatin, R., Rydzanicz, M., Sobczyk-Kopcioł, A., Sulima-Celińska, A., Paukszto, Ł., Makowczenko, K.G., Płoski, R., Tkach, V. V., Basałaj, K., Młocicki, D., 2019. Hybrid de novo whole-genome assembly and annotation of the model tapeworm Hymenolepis diminuta. Sci. Data 6, 1–14. https://doi.org/10.1038/s41597-019-0311-3

Nurk, S., Walenz, B.P., Rhie, A., Vollger, M.R., Logsdon, G.A., Grothe, R., Miga, K.H., Eichler, E.E., Phillippy, A.M., Koren, S., 2020. HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 30, 1291–1305.

Nan X, Ng H-H, Johnson CA, et al (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature 393:386–389.

Nguyen AL, Dang VH, Bosma RH, et al (2014a) Simulated impacts of climate change on current farming locations of striped catfish (*Pangasianodon hypophthalmus*; Sauvage) in the Mekong Delta, Vietnam. Ambio 43:1059–1068. https://doi.org/10.1007/s13280-014-0519-6

Nguyen PTH, Do HTT, Mather PB, Hurwood DA (2014b) Experimental assessment of the effects of sublethal salinities on growth performance and stress in cultured tra catfish (*Pangasianodon hypophthalmus*). Fish Physiol Biochem 40:1839–1848. https://doi.org/10.1007/s10695-014-9972-1

Norouzitallab P, Baruah K, Vanrompay D, Bossier P (2019) Can epigenetics translate environmental cues into phenotypes? Sci Total Environ 647:1281–1293

Norman, J.D., Danzmann, R.G., Glebe, B., Ferguson, M.M., 2011. The genetic basis of salinity tolerance traits in Arctic charr (Salvelinus alpinus). BMC Genet. 12, 1–12. https://doi.org/10.1186/1471-2156-12-81

Norris, A.T., Bradley, D.G., Cunningham, E.P., 2000. Parentage and relatedness determination in farmed Atlantic salmon (*Salmo salar*) using microsatellite markers. Aquaculture 182, 73–83. https://doi.org/10.1016/S0044-8486(99)00247-1

Nielsen, R., Paul, J.S., Albrechtsen, A. & Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451.

**-O-**

Okada, Y., Momozawa, Y., Sakaue, S., Kanai, M., Ishigaki, K., Akiyama, M., Kishikawa, T., Arai, Y., Sasaki, T., Kosaki, K., Suematsu, M., Matsuda, K., Yamamoto, K., Kubo, M., Hirose, N., Kamatani, Y., 2018. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. Nat. Commun. 9, 1–10. https://doi.org/10.1038/s41467-018-03274-0

Ohbo K, Tomizawa S (2015) Epigenetic regulation in stem cell development, cell fate conversion, and reprogramming. Biomol Concepts 6:1–9

Okano M, Bell DW, Haber DA, Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99:247–257

Olson, C.E., Roberts, S.B., 2014. Genome-wide profiling of DNA methylation and gene expression in Crassostrea gigas male gametes. Front. Physiol. 5 JUN, 1–7. https://doi.org/10.3389/fphys.2014.00224

Ong-Abdullah, M., Ordway, J.M., Jiang, N., Ooi, S.E., Kok, S.Y., Sarpan, N., Azimi, N., Hashim, A.T., Ishak, Z., Rosli, S.K., Malike, F.A., Bakar, N.A.A., Marjuni, M., Abdullah, N., Yaakub, Z., Amiruddin, M.D., Nookiah, R., Singh, R., Low, E.T.L., X., Wong, C.K., Choo, C.N., Wong, W.C., Kwan, Y.Y., Alwee, S.S.R.S., Sambanthamurthi, R., Martienssen, R.A., 2015. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature 525, 533–537.

Olesen, I., Bentsen, H.B., Phillips, M., Ponzoni, R.W., 2015. Can the global adoption of genetically improved farmed fish increase beyond 10%, and how? J. Mar. Sci. Eng. 3, 240–266.

Olesen, I., Myhr, A.I., Rosendal, G.K., 2011. Sustainable Aquaculture: Are We Getting There? Ethical Perspectives on Salmon Farming. J. Agric. Environ. Ethics 24, 381–408.

Orban, E., Nevigato, T., Di Lena, G., Masci, M., Casini, I., Gambelli, L., Caproni, R., 2008. New trends in the seafood market. Sutchi catfish (*Pangasius hypophthalmus*) fillets from Vietnam: Nutritional quality and safety aspects. Food Chem. 110, 383–389. https://doi.org/10.1016/j.foodchem.2008.02.014

**-P-**

Parker, H., Carr, L., Syeda, S., Bryant, D., & Strefford, J. C. (2019). Characterization of Somatically-Acquired Copy Number Alterations in Chronic Lymphocytic Leukaemia Using Shallow Whole Genome Sequencing. In *Methods in Molecular Biology* (pp. 327–353). Humana Press, New York, NY. doi: 10.1007/978-1-4939-8876-1_23

Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet.* 44:631–5.

Palti, Y., Gao, G., Liu, S., Kent, M.P., Lien, S., Miller, M.R., Rexroad, C.E., Moen, T., 2015. The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. Mol. Ecol. Resour. 15, 662–672. https://doi.org/10.1111/1755-0998.12337

Panserat, S., Marandel, L., Geurden, I., Veron, V., Dias, K., Plagnes-Juan, E., Pegourié, G., Arbenoits, E., Santigosa, E., Weber, G., Verlhac Trichet, V., 2017. Muscle catabolic capacities and global hepatic epigenome are modified in juvenile rainbow trout fed different vitamin levels at first feeding. Aquaculture 468, 515–523. https://doi.org/10.1016/j.aquaculture.2016.11.021

Parker, L.M., O'Connor, W.A., Raftos, D.A., Pörtner, H.O., Ross, P.M., 2015. Persistence of positive carryover effects in the oyster, Saccostrea glomerata, following transgenerational exposure to ocean acidification. PLoS One 10, 1–19. https://doi.org/10.1371/journal.pone.0132276

Parker, L.M., Ross, P.M., O'Connor, W.A., Borysko, L., Raftos, D.A., Pörtner, H.O., 2012. Adult exposure influences offspring response to ocean acidification in oysters. Glob. Chang. Biol. 18, 82–92. https://doi.org/10.1111/j.1365-2486.2011.02520.x

Pang, M., Fu, B., Yu, X., Liu, H., Wang, X., Yin, Z., Xie, S., Tong, J., 2017. Quantitative trait loci mapping for feed conversion efficiency in crucian carp (*Carassius auratus*). Sci. Rep. 7, 1–11. https://doi.org/10.1038/s41598-017-17269-2

Phan, L.T., Bui, T.M., Nguyen, T.T.T., Gooley, G.J., Ingram, B.A., Nguyen, H. V., Nguyen, P.T., De Silva, S.S., 2009. Current status of farming practices of striped catfish, *Pangasianodon hypophthalmus* in the Mekong Delta, Vietnam. Aquaculture 296, 227–236. https://doi.org/10.1016/j.aquaculture.2009.08.0

Purcell, K.M., Hitch, A.T., Klerks, P.L., Leberg, P.L., 2008. Adaptation as a potential response to sea-level rise: a genetic basis for salinity tolerance in populations of a coastal marsh fish. Evol. Appl. 1, 155–160. https://doi.org/10.1111/j.1752-4571.2007.00001.x

Pollard, M.O., Gurdasani, D., Mentzer, A.J., Porter, T., Sandhu, M.S., 2018. Long reads: Their purpose and place. Hum. Mol. Genet. 27, R234–R241. https://doi.org/10.1093/hmg/ddy177

Pilsner JR, Lazarus AL, Nam D-H, et al (2010) Mercury-associated DNA hypomethylation in polar bear brains via the LUminometric Methylation Assay: a sensitive method to study epigenetics in wildlife. Mol Ecol 19:307–314. https://doi.org/10.1111/j.1365-294X.2009.04452.x

Potok ME, Nix DA, Parnell TJ, Cairns BR (2013) Reprogramming the maternal zebrafish genome after fertilization to match the paternal methylation pattern. Cell 153:759–772

Phuc, N.T.H., Mather, P.B., Hurwood, D.A., 2017. Effects of sublethal salinity and temperature levels and their interaction on growth performance and hematological and hormonal levels in tra catfish (Pangasianodon hypophthalmus). Aquac. Int. 25, 1057–1071. https://doi.org/10.1007/s10499-016-0097-7

Pemberton, J.M., 2008. Wild pedigrees: The way forward. Proc. R. Soc. B Biol. Sci. 275, 613–621. https://doi.org/10.1098/rspb.2007.1531

Phadphon, P., Amontailak, T., Kotchantuek, N., Srithawong, S., Kutanan, W., Suwannapoom, C., 2019. Genetic Diversity of the Endangered Mekong Giant Catfish, Striped Catfish, and Their Hybrids From Thailand. Trop. Conserv. Sci. 12. https://doi.org/10.1177/1940082919869487

Pham, T.L., Bui, M.T., Dao, M.H., 2021. Current states of seed production of striped catfish (Pangasianodon hypophthalmus) in Mekong Delta. ODA Proj. Report, Can Tho Univ.

Plough, L. V., 2016. Genetic load in marine animals: A review. Curr. Zool. 62, 567–579. https://doi.org/10.1093/cz/zow096

Preston, J., VanZeeland, A., Peiffer, D.A., 2021. Innovation at Illumina: The road to the $600 human genome. Nat. Portf.Qin, D., 2019. Next-generation sequencing and its clinical application. Cancer Biol. Med. 16, 4–10. https://doi.org/10.20892/j.issn.2095-3941.2018.0055

Pfeiffer, F., Gröber, C., Blank, M. *et al.* (2018) Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci Rep* **8,** 10950. https://doi.org/10.1038/s41598-018-29325

Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: Causes, consequences and solutions. *Nature Reviews Genetics*, *6*(11), 847–859. doi: 10.1038/nrg1707

**-Q-**

**-R-**

Raman, L., Dheedene, A., De Smet, M., Van Dorpe, J., Menten, B., 2019. WisecondorX: improved copy number detection for routine shallow whole-genome sequencing. Nucleic Acids Res. 47, 1605–1614. https://doi.org/10.1093/nar/gky1263

Raul, W.P., Belen, O.A., G.Ponniah, A., 2006. Development of Aquatic Animal Genetic Improvement and Dissemination Programs: Current Status and Action Plans. WorldFish Center.

Renaud, F.G., Kuenzer, C., Tri, V.K., Garschagen, M., Diez, J.R., Nhan, D.K., Kraas, F., Moder, F., 2012. The Mekong Delta System: Interdisciplinary analyses of a River Delta.

Ronco, F., Matschiner, M., Böhne, A., Boila, A., Büscher, H.H., El Taher, A., Indermaur, A., Malinsky, M., Ricci, V., Kahmen, A., Jentoft, S., Salzburger, W., 2021. Drivers and dynamics of a massive adaptive radiation in cichlid fishes. Nature 589, 76–81. https://doi.org/10.1038/s41586-020-2930-4

Riley LG, Richman NH 3rd, Hirano T, Gordon Grau E (2002) Activation of the growth hormone/insulin-like growth factor axis by treatment with 17 alpha-methyltestosterone and seawater rearing in the tilapia, *Oreochromis mossambicus*. Gen Comp Endocrinol 127:285–292.

Rezk, M.A., Smitherman, R.O., Williams, J.C., Nichols, A., Kucuktas, H., Dunham, R.A., 2003. Response to three generations of selection for increased body weight in channel catfish, *Ictalurus punctatus*, grown in earthen ponds. Aquaculture 228, 69–79. https://doi.org/10.1016/S0044-8486(03)00216-3

Ross SE, Bogdanovic O (2021) Generation and Molecular Characterization of Transient tet1/2/3 ZebrafishZebrafish Knockouts BT - TET Proteins and DNA Demethylation: Methods and Protocols. In: Bogdanovic O, Vermeulen M (eds). Springer US, New York, NY, pp 281–318

Rougier N, Bourc'his D, Gomes DM, et al (1998) Chromosome methylation patterns during mammalian preimplantation development. Genes Dev 12:2108–2113

Rebollo, R., Romanish, M.T., Mager, D.L., 2012. Transposable elements: An abundant and natural source of regulatory sequences for host genes. Annu. Rev. Genet. 46, 21–42.

Rengmark, A.H., Slettan, A., Lee, W.J., Lie, Lingaas, F., 2007. Identification and mapping of genes associated with salt tolerance in tilapia. J. Fish Biol. 71, 409–422. https://doi.org/10.1111/j.1095-8649.2007.01664.x

Rey, O., Danchin, E., Mirouze, M., Loot, C., Blanchet, S., 2016. Adaptation to Global Change: A Transposable Element-Epigenetics Perspective. Trends Ecol. Evol. 31, 514–526.

Rye, M., Lillevik, K.M., Gjerde, B., 1990. Survival in early life of Atlantic salmon and rainbow trout: estimates of heritabilities and genetic correlations. Aquaculture 89, 209–216.

Rye, M., Refstie, T., 1995. Phenotypic and genetic parameters of body size traits in Atlantic salmon Salmo Salar L. Aquac. Res. 26, 875–885. https://doi.org/10.1111/j.1365-2109.1995.tb00882.x

Riester, M., Stadler, P.F., Klemm, K., 2009. FRANz: Reconstruction of wild multi-generation pedigrees. Bioinformatics 25, 2134–2139. https://doi.org/10.1093/bioinformatics/btp064

Riviere, G., He, Y., Tecchio, S., Crowell, E., Gras, M., Sourdaine, P., Guo, X., Favrel, P., 2017. Dynamics of DNA methylomes underlie oyster development. PLoS Genet. 13, 1–16.

Rodgers, A.B., Morgan, C.P., Leu, N.A., Bale, T.L., 2015. Transgenerational epigenetic programming via sperm microRNA recapitulates effects of paternal stress. Proc. Natl. Acad. Sci. U. S. A. 112, 13699–13704. https://doi.org/10.1073/pnas.1508347112

**-S-**

Saillant, E., Dupont-Nivet, M., Haffray, P., Chatain, B., 2006. Estimates of heritability and genotype–environment interactions for body weight in sea bass (*Dicentrarchus labrax L.*) raised under communal rearing conditions. Aquaculture 254, 139–147. https://doi.org/10.1016/j.aquaculture.-2005.10.

Sang, N. Van, Klemetsdal, G., Ødegård, J., Gjøen, H.M., 2012. Genetic parameters of economically important traits recorded at a given age in striped catfish (*Pangasianodon hypophthalmus*). Aquaculture 344–349, 82–89. https://doi.org/10.1016/j.aquaculture.2012.03.013.

Sang, N. Van, Nguyen, T., Sang, V., Gjøen, H.M., Gunnar, P., Prof, K., Thomassen, M., 2010. Genetic studies on improvement of striped catfish (Pangasianodon hypophthalmus) for economically important traits Genetic studies on improvement of striped catfish (Pangasianodon hypophthalmus) for economically important traits Genetiske studier av økonom.

Santos, F., Hendrich, B., Reik, W., Dean, W., 2002. Dynamic reprogramming of DNA methylation in the early mouse embryo. Dev. Biol. 241, 172–182. https://doi.org/10.1006/dbio.2001.0501

SAS Institute (2009) SAS Proprietary Software Version 9.3. SAS Institute, Cary, NC.

Sae-Lim, P., Kause, A., Mulder, H.A., Olesen, I., 2017. Breeding and genetics symposium: Climate change and selective breeding in aquaculture. J. Anim. Sci. 95, 1801–1812.

Shikano, T., Chiyokubo, T., Taniguchi, N., 2001. Effect of inbreeding on salinity tolerance in the guppy (*Poecilia reticulata*). Aquaculture 202, 45–55. https://doi.org/10.1016/S0044-8486(01)00568-3

Smajgl, A., Toan, T.Q., Nhan, D.K., Ward, J., Trung, N.H., Tri, L.Q., Tri, V.P.D., Vu, P.T., 2015. Responding to rising sea levels in the Mekong Delta. Nat. Clim. Chang. 5, 167–174.

Son, V.N., Phuong, N.T., Hai, T.N., Yakupitiyage, A., 2011. Production and economic efficiencies of intensive black tiger prawn (Penaeus monodon) culture during different cropping seasons in the Mekong Delta, Vietnam. Aquac. Int. 19, 555–566. https://doi.org/10.1007/s10499-010-9371-2

Su, G.-S., Liljedahl, L.-E., Gall, G.A.E., 1996. Effects of inbreeding on growth and reproductive traits in rainbow trout (*Oncorhynchus mykiss*). Aquaculture 142, 139–148. https://doi.org/10.1016/0044-8486(96)01255-0

Shikano, T., Chiyokubo, T., Taniguchi, N., 2001. Temporal changes in allele frequency, genetic variation and inbreeding depression in small populations of the guppy, Poecilia reticulata. Heredity (Edinb). 86, 153–160. https://doi.org/10.1046/j.1365-2540.2001.00792.x

Shikano, T., Fujio, Y., 1998a. Maternal Effect on Salinity Tolerance in Newborn Guppy *Poecilia reticulata*. Fish. Sci. 64, 52–56. https://doi.org/10.2331/fishsci.64.52

Shikano, T., Fujio, Y., 1998b. Strain Differences in Seawater Adaptability in Newborn Guppy *Poecilia reticulata.* Fish. Sci. 64, 987–988. https://doi.org/10.2331/fishsci.64.987

Shoemaker, C.A., Lozano, C.A., LaFrentz, B.R., García, J.C., Soto, E., Xu, D.H., Beck, B.H., Rye, M., 2017. Additive genetic variation in resistance of Nile tilapia (*Oreochromis niloticus*) to *Streptococcus iniae* and S. *agalactiae* capsular type Ib: Is genetic resistance correlated? Aquaculture 468, 193–198. https://doi.org/10.1016/j.aquaculture.2016.10.022

Storset, A., Strand, C., Wetten, M., Kjøglum, S., Ramstad, A., 2007. Response to selection for resistance against infectious pancreatic necrosis in Atlantic salmon (*Salmo salar* L.). Aquaculture 272, 62–68. https://doi.org/10.1016/j.aquaculture.2007.08.011

Sukhavachana, S., Poompuang, S., Onming, S., Luengnaruemitchai, A., 2019. Heritability estimates and selection response for resistance to *Streptococcus agalactiae* in red tilapia *Oreochromis spp*. Aquaculture 502, 384–390. https://doi.org/10.1016/j.aquaculture.2018.12.075

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., Waterston, R.H., 2017. DNA sequencing at 40: Past, present and future. Nature 550.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Sreeputhorn, K., Mangumphan, K., Muanphet, B., Tanomtong, A., Supiwong, W., Kaewmad, P., 2017. The first report on chromosome analysis of F1 hybrid catfish: Mekong giant catfish (Pangasianodon gigas) × striped catfish (*Pangasianodon hypophthalmus*) and spot pangasius (*Pangasius larnaudii*) × *Pangasianodon hypophthalmus* (*Siluriformes, Pangasiidae*). Cytologia (Tokyo). 82, 457–463.

Star, B., Nederbragt, A.J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T.F., Rounge, T.B., Paulsen, J., Solbakken, M.H., Sharma, A., Wetten, O.F., Lanzén, A., Winer, R., Knight, J., Vogel, J.H., Aken, B., Andersen, Ø., Lagesen, K., Previti, C., Karlsen, B.O., Moum, T., Skage, M., Berg, P.R., Gjøen, T., Kuhl, H., Thorsen, J., Malde, K., Reinhardt, R., Du, L., Johansen, S.D., Searle, S., Lien, S., Nilsen, F., Jonassen, I., Omholt, S.W., Stenseth, N.C., Jakobsen, K.S., 2011. The genome sequence of Atlantic cod reveals a unique immune system. Nature 477, 207–210. https://doi.org/10.1038/nature10342

Sharma, P., Al-Dossary, O., Alsubaie, B., Al-Mssallem, I., Nath, O., Mitter, N., Rodrigues Alves Margarido, G., Topp, B., Murigneux, V., Kharabian Masouleh, A., Furtado, A., Henry, R.J., 2021. Improvements in the sequencing and assembly of plant genomes. Gigabyte 2021, 1–10.

Sakamoto T, McCormick SD (2006) Prolactin and growth hormone in fish osmoregulation. Gen Comp Endocrinol 147:24–30

Santos F, Hendrich B, Reik W, Dean W (2002) Dynamic reprogramming of DNA methylation in the early mouse embryo. Dev Biol 241:172–182

Schmitz M, Douxfils J, Mandiki SNM, et al (2016) Chronic hyperosmotic stress interferes with immune homeostasis in striped catfish (*Pangasianodon hypophthalmus*, S.) and leads to excessive inflammatory response during bacterial infection. Fish Shellfish Immunol 55:550–558.

Si Y, He F, Wen H, et al (2021) Effects of low salinity on epigenetic changes of growth hormone and growth hormone receptor in half smooth tongue sole (*Cynoglossus semilaevis*). Reprod Breed 1:11–21. https://doi.org/https://doi.org/10.1016/j.repbre.2021.01.001

Skinner MK (2015) Environmental epigenetics and a unified theory of the molecular aspects of evolution: a neo-Lamarckian concept that facilitates neo-Darwinian evolution. Genome Biol Evol 7:1296–1302

Smith J, Sen S, Weeks RJ, et al (2020) Promoter DNA hypermethylation and paradoxical gene activation. Trends in cancer 6:392–406

Song J, Teplova M, Ishibe-Murakami S, Patel DJ (2012) Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. Science (80- ) 335:709–712

Suebsong W, Poompuang S, Srisapoome P, et al (2019) Selection response for Streptococcus agalactiae resistance in Nile tilapia *Oreochromis niloticus*. J Fish Dis 42:1553–1562

Swales AKE, Spears N (2005) Genomic imprinting and reproduction. Reproduction 130:389–399. https://doi.org/10.1530/rep.1.00395

Star, B., Nederbragt, A.J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T.F., Rounge, T.B., Paulsen, J., Solbakken, M.H., Sharma, A., Wetten, O.F., Lanzén, A., Winer, R., Knight, J., Vogel, J.H., Aken, B., Andersen, Ø., Lagesen, K., Tooming-Klunderud, A., Edvardsen, R.B., Tina, K.G., Espelund, M., Nepal, C., Previti, C., Karlsen, B.O., Moum, T., Jakobsen, K.S., 2011. The genome sequence of Atlantic cod reveals a unique immune system. Nature 477, 207–210. https://doi.org/10.1038/nature10342

Schmitz, M., Mandiki, S.N.M., Douxfils, J., Ziv, T., Admon, A., Kestemont, P., 2016. Synergic stress in striped catfish (*Pangasianodon hypophthalmus, S.*) exposed to chronic salinity and bacterial infection: Effects on kidney protein expression profile. J. Proteomics 142, 91–101.

Sébastien, L., 2012. Relationships between growth, physiological and behavioural traits of the European abalone Haliotis tuberculata in the perspective of a selective breeding plan implementation 1–50.

Sebesvari, Z., Le, T.T.H., Fabrice, G.R., 2011. Climate Change Adaptation and Agrichemicals in the Mekong Delta, Vietnam. In : M.A. Stewart and P.A. Coclanis (eds.), Environmental Change and Agricultural Sustainability in the Mekong Delta, Advances in Global Change Research 45. Springer Sci. Media B.V 219–239. https://doi.org/10.1007/978-94-007-0934-8

Seisenberger, S., Peat, J.R., Hore, T.A., Santos, F., Dean, W., Reik, W., 2013. Reprogramming DNA methylation in the mammalian life cycle: Building and breaking epigenetic barriers. Philos. Trans. R. Soc. B Biol. Sci. 368, 1–11. https://doi.org/10.1098/rstb.2011.0330

Sellars, M.J., Dierens, L., Mcwilliam, S., Little, B., Murphy, B., Coman, G.J., Barendse, W., Henshall, J., 2014. Comparison of microsatellite and SNP DNA markers for pedigree assignment in Black Tiger shrimp, Penaeus monodon. Aquac. Res. 45, 417–426.

Senaldi, L., Smith-Raska, M., 2020. Evidence for germline non-genetic inheritance of human phenotypes and diseases. Clin. Epigenetics 12, 1–12. https://doi.org/10.1186/s13148-020-00929-y

Shao, C., Li, Q., Chen, S., Zhang, P., Lian, J., Hu, Q., Sun, B., Jin, L., Liu, S., Wang, Z., Zhao, H., Jin, Z., Liang, Z., Li, Y., Zheng, Q., Zhang, Y., Wang, J., Zhang, G., 2014. Epigenetic modification and inheritance in sexual reversal of fish. Genome Res. 24, 604–615. https://doi.org/10.1101/gr.162172.113.

Simmen, M.W., Leitgeb, S., Charlton, J., Jones, S.J.M., Harris, B.R., Clark, V.H., Bird, A., 1999. Nonmethylated Transposable Elements and Methylated Genes in a Chordate Genome. Science (80-. ). 283, 1164–1167. https://doi.org/10.1126/science.283.5405.1164.

Sodeland, M., Gaarder, M., Moen, T., Thomassen, M., Kjøglum, S., Kent, M., Lien, S., 2013. Genome-wide association testing reveals quantitative trait loci for fillet texture and fat content in Atlantic salmon. Aquaculture 408–409, 169–174. https://doi.org/10.1016/j.aquaculture.2013.05.029

Sopha, L., Bun, N.P., Kong, H., Joergensen, J. V, 2002. Status of the Mekong Pangasianodon hypophthalmus resources, with special reference to the stock shared between Cambodia and Viet Nam. MRC Tech. Pap. 1–29.

Soubry, A., Hoyo, C., Jirtle, R.L., Murphy, S.K., 2014. A paternal environmental legacy: Evidence for epigenetic inheritance through the male germ line. BioEssays 36, 359–371. https://doi.org/10.1002/bies.201300113

Sourinejad, I., Kalbassi, M.R., Pino-Querido, A., Vera, M., Bouza, C., Martinez, P., 2011. Parentage assignment of progeny in mixed milt fertilization of caspian brown trout salmo trutta caspius

using microsatellite DNA markers: Implications for conservation. African J. Biotechnol. 10, 5084–5090. https://doi.org/10.5897/AJB10.664

Strömqvist, M., Tooke, N., Brunström, B., 2010. DNA methylation levels in the 5' flanking region of the vitellogenin I gene in liver and brain of adult zebrafish (*Danio rerio*)-Sex and tissue differences and effects of 17α-ethinylestradiol exposure. Aquat. Toxicol. 98, 275–281. https://doi.org/10.1016/j.aquatox

Strucken, E.M., Lee, S.H., Lee, H.K., Song, K.D., Gibson, J.P., Gondro, C., 2016. How many markers are enough? Factors influencing parentage testing in different livestock populations. J. Anim. Breed. Genet. 133, 13–23. https://doi.org/10.1111/jbg.12179

Suebsong, W., Poompuang, S., Srisapoome, P., Koonawootrittriron, S., Luengnaruemitchai, A., Johansen, H., Rye, M., 2019. Selection response for Streptococcus agalactiae resistance in Nile tilapia Oreochromis niloticus. J. Fish Dis. 42, 1553–1562. https://doi.org/10.1111/jfd.13074

Sobel, E., Papp, J. C., & Lange, K. (2002). Detection and Integration of Genotyping Errors in Statistical Genetics. *Am. J. Hum. Genet*, *70*, 496–508.

Strucken, E. M., Lee, S. H., Lee, H. K., Song, K. D., Gibson, J. P., & Gondro, C. (2016). How many markers are enough? Factors influencing parentage testing in different livestock populations. *Journal of Animal Breeding and Genetics*, *133*(1), 13–23. doi: 10.1111/jbg.12179

**-T-**

Taggart, J.B., 2006. PROGRAM NOTE: FAP: an exclusion-based parental assignment program with enhanced predictive functions. Mol. Ecol. Notes 7, 412–415.

Takei, Y., Balment, R.J., 2009. Chapter 8 The Neuroendocrine Regulation of Fluid Intake and Fluid Balance. pp. 365–419. https://doi.org/10.1016/S1546-5098(09)28008-3

Taugbøl, A., Arntsen, T., Østbye, K., Vøllestad, L.A., 2014. Small changes in gene expression of targeted osmoregulatory genes when exposing marine and freshwater threespine stickleback (gasterosteus aculeatus) to abrupt salinity transfers. PLoS One 9, e106894. https://doi.org/10.1371/journal.pone.0106894

Tayamen, M.M., Abella, T.A., Reyes, R.A., Danting, M.J.C., Mendoza, A.M., Marquez, E.B., Salguet, A.C., Apaga, M.M., González, R.C., 2010. Development of tilapia for saline waters in the Philippines. Proc 6th Int Symp Tilapia Aquac. Boliv. R.B., Mair G.C., Fitzsimmons K. (eds.), ICLARM, Manila, Philipp. 463-478 pp.

Tamura K, Stecher G, Kumar S (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol 38:3022–3027. https://doi.org/10.1093/molbev/msab120

Tan L, Shi YG (2012) Tet family proteins and 5-hydroxymethylcytosine in development and disease. Development 139:1895–1902. https://doi.org/10.1242/dev.070771

Tan, M.H., Austin, C.M., Hammer, M.P., Lee, Y.P., Croft, L.J., Gan, H.M., 2018. Finding Nemo: Hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (Amphiprion ocellaris) genome assembly. Gigascience 7, 1–6. https://doi.org/10.1093/gigascience/gix137

Tang, B., Wang, Zhongkai, Liu, Q., Wang, Zhengfei, Ren, Y., Guo, H., Qi, T., Li, Yuetian, Zhang, H., Jiang, S., Ge, B., Xuan, F., Sun, Y., She, S., Yam Chan, T., Sha, Z., Jiang, H., Li, H., Jiang, W., Qin, Y., Wang, K., Qiu, Q., Wang, W., Li, X., Ng, N.K., Zhang, D., Li, Yongxin, 2021. Chromosome-level genome assembly of Paralithodes platypus provides insights into evolution and adaptation of king crabs. Mol. Ecol. Resour. 21, 511–525. https://doi.org/10.1111/1755-0998.13266

Tarailo-Graovac, M., Chen, N., 2009. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Curr. Protoc. Bioinforma. 25. https://doi.org/10.1002/0471250953.bi0410s25

Tahapari, E., Darmawan, J., Suharyanto, 2018. Genetic improvement of growth trait in siamese catfish (*Pangasianodon hypophthalmus* (Sauvage, 1878)) through family selection. AACL Bioflux 11, 1648–1657.

Taniguchi, N., Perez-Enriquez, R., Estu, N., 2003. DNA markers as a tool for genetic management of brood stock for aquaculture, in: Aquatic Genomics. Springer Japan, Tokyo, pp. 417–429. https://doi.org/10.1007/978-4-431-65938-9_38

Tartila, S.S.Q., Jusadi, D., Setiawati, M., Fauzi, I.A., 2021. Evaluation of dietary supplementation with cinnamon products on growth, blood composition, liver structure, and meat quality of striped catfish *(Pangasianodon hypophthalmus*). Aquac. Int. 29, 2243–2257. https://doi.org/10.1007/s10499-021-00746

Therkildsen, N.O., Palumbi, S.R., 2017. Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Mol. Ecol. Resour. 17, 194–208. https://doi.org/10.1111/1755-0998.12593

Tuan, L.A., Chinvanno, S., 2011. Climate Change in the Mekong River Delta and Key Concerns on Future Climate Threats. pp. 207–217. https://doi.org/10.1007/978-94-007-0934-8_12

Thanh, N.M ., Jung, H., Lyons, R.E., Viet, N., Thi, V., Thu, M., Mather, P., 2014. A transcriptomic analysis of striped catfish (*Pangasianodon hypophthalmus*) in response to salinity adaptation: De novo assembly, gene annotation and marker discovery. Comp. Biochem. Physiol. Part D 10, 52–63.

Thi, T., Trieu, N., Phong, N.T., 2015. The impact of climate change on salinity intrusion and Pangasius ( *Pangasianodon Hypophthalmus* ) farming. Aquac. Int. 23, 523–534.

Thoa, N.P., Ninh, N.H., Knibb, W., Nguyen, N.H., 2016. Does selection in a challenging environment produce Nile tilapia genotypes that can thrive in a range of production systems? Sci. Rep. 6, 1–11. https://doi.org/10.1038/srep21486

Thodesen, J., Grisdale-Helland, B., Helland, S.J., Gjerde, B., 1999. Feed intake, growth and feed utilization of offspring from wild and selected Atlantic salmon (*Salmo salar*). Aquaculture 180, 237–246. https://doi.org/10.1016/S0044-8486(99)00204-5

Torda, G., Donelson, J.M., Aranda, M., Barshis, D.J., Bay, L., Berumen, M.L., Bourne, D.G., Cantin, N., Foret, S., Matz, M., Miller, D.J., Moya, A., Putnam, H.M., Ravasi, T., Van Oppen, M.J.H., Thurber, R.V., Vidal-Dupiol, J., Voolstra, C.R., Watson, S.A., Whitelaw, E., Willis, B.L., Munday, P.L., 2017. Rapid adaptive responses to climate change in corals. Nat. Clim. Chang. 7, 627–636.

Tran, D.L., Olesen, I., Ødegård, J., Kolstad, K., Nguyen, C.D., 2008. Genotype by environment interaction for harvest body weight and survival of Nile Tilapia (*Oreochromis niloticus*) in brackish and fresh water ponds. The Proceedings of 8th International Symposium on Tilapia in Aquaculture (Egypt). 231–239.

Thiel, T., Michalek, W., Varshney, R., Graner, A., 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor. Appl. Genet. 106, 411–422. https://doi.org/10.1007/s00122-002-1031-0

Turek-Plewa J, Jagodzinski PP (2005) The role of mammalian DNA methyltransferases in the regulation of gene expression. Cell Mol Biol Lett 10:631

Teichert-Coddington, D.R., Smitherman, R.O., 1988. Lack of Response by Tilapia nilotica to Mass Selection for Rapid Early Growth. Trans. Am. Fish. Soc. 117, 297–300. https://doi.org/10.1577/1548-8659(1988)117<0297:LORBNT>2.3.CO;2

Terova, G., Díaz, N., Rimoldi, S., Ceccotti, C., Gliozheni, E., Piferrer, F., 2016. Effects of sodium butyrate treatment on histone modifications and the expression of genes related to epigenetic regulatory mechanisms and immune response in European Sea Bass (*Dicentrarchus Labrax*) fed a plant-based diet. PLoS One 11, 1–20. https://doi.org/10.1371/journal.pone.0160332

Thao, N.L., Ha, N.T.K., Phuong, N.T., Phuc, N.T.H., 2013. Effects of salinity of growth and cortisol of striped catfish (*Pangasisanodon hypophthalmus*). Can Tho Univ. J. Sci. 25, 1–10.

Thodesen, J., Rye, M., Wang, Y.-X., Bentsen, H.B., Gjedrem, T., 2012. Genetic improvement of tilapias in China: Genetic parameters and selection responses in fillet traits of Nile tilapia (*Oreochromis niloticus*) after six generations of multi-trait selection for growth and fillet yield. Aquaculture 366–367, 67–75. https://doi.org/10.1016/j.aquaculture.2012.08.028.

Tine, M., Kuhl, H., Gagnaire, P.A., Louro, B., Desmarais, E., Martins, R.S.T., Hecht, J., Knaust, F., Belkhir, K., Klages, S., Dieterich, R., Stueber, K., Piferrer, F., Guinand, B., Bierne, N., Volckaert, F.A.M., Bargelloni, L., Power, D.M., Bonhomme, F., Canario, A.V.M., Reinhardt, R., 2014. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nat. Commun. 5, 5770. https://doi.org/10.1038/ncomms6770

Tortereau, F., Moreno, C.R., Tosser-Klopp, G., Servin, B., Raoul, J., 2017. Development of a SNP panel dedicated to parentage assignment in French sheep populations. BMC Genet. 18, 1–11.

Tran, T.X., 1994. Some biological characteristics and artificial reproduction of river catfish (Pangasius micronemus Bleeker) in the South Vietnam. Paper presented at the International Workshop on the biological bases for aquaculture of Siluriformes, May 24–27,Montpellier,.

Trọng, T.Q., van Bers, N., Crooijmans, R., Dibbits, B., Komen, H., 2013. A comparison of microsatellites and SNPs in parental assignment in the GIFT strain of Nile tilapia (Oreochromis niloticus): The power of exclusion. Aquaculture 388–391, 14–23.

Tuan, N.., Phuong, N.T., Liem, P.T., Thuong, N.V., 2003. Results of the study on Pangasius catfishes and their future development. J. Mekong Fish. (in Vietnamese). Agric. Publ. house. Ho Chi Minh city. Viet Nam. 129–134.

Tucker, K., 2009. Methylated Cytosine and the Brain: A New Base for Neuroscience. Stud. Health Technol. Inform. 150, 762–766. https://doi.org/10.3233/978-1-60750-044-5-762

Thrasher, D. J., Butcher, B. G., Campagna, L., Webster, M. S., & Lovette, I. J. (2018). Double-digest RAD sequencing outperforms microsatellite loci at assigning paternity and estimating relatedness: A proof of concept in a highly promiscuous bird. *Molecular Ecology Resources*, *18*(5), 953–965. doi: 10.1111/1755-0998.12771

Trọng, T. Q., van Bers, N., Crooijmans, R., Dibbits, B., & Komen, H. (2013). A comparison of microsatellites and SNPs in parental assignment in the GIFT strain of Nile tilapia (*Oreochromis niloticus)*: The power of exclusion. *Aquaculture*, *388–391*, 14–23. doi: 10.1016/J.Aquaculture. 2013.01.004

**-U-**

Ulloa, P.E., Rincón, G., Islas-Trejo, A., Araneda, C., Iturra, P., Neira, R., Medrano, J.F., 2015. RNA Sequencing to Study Gene Expression and SNP Variations Associated with Growth in Zebrafish Fed a Plant Protein-Based Diet. Mar. Biotechnol. 17, 353–363. https://doi.org/10.1007/s10126-015-9624-1

Uren Webster TM, Rodriguez-Barreto D, Martin SAM, et al (2018) Contrasting effects of acute and chronic stress on the transcriptome, epigenome, and immune response of Atlantic salmon. Epigenetics 13:1191–1207

**-V-**

Van Raden, P., Sun, C., & O'Connell, J. (2015). Fast imputation using medium or low-covergae sequence data. BMC Genetics, 16, 82. doi: 10.1186/s12863-015-0243-7

Van Roy, N., Van Der Linden, M., Menten, B., Dheedene, A., Vandeputte, C., Van Dorpe, J., … De Preter, K. (2017). Shallow whole genome sequencing on circulating cell-free DNA allows reliable noninvasive copy-number profiling in neuroblastoma patients. *Clinical Cancer Research*, *23*(20), 6305–6315. doi: 10.1158/1078-0432.CCR-17-0675

Vandeputte, M., & Haffray, P. (2014). Parentage assignment with genomic markers: A major advance for understanding and exploiting genetic variation of quantitative traits in farmed aquatic animals. *Frontiers in Genetics*, *5*(DEC), 1–8. doi: 10.3389/fgene.2014.00432

Van Beek, D., 2012. Detection of Copy Number Variation using Shallow Whole Genome Sequencing Data to replace Array-Comparative, Master Thesis, Delft University of Technology.

Van der Eecken, K., Van der Linden, M., Raman, L., Creytens, D., Dedeurwaerdere, F., De Winne, K., Ferdinande, L., Lammens, M., Menten, B., Rottiers, I., Van Gaever, B., Van den Broecke, C., Van de Vijver, K., Van Roy, N., Verbeke, S., Van Dorpe, J., 2022. Shallow whole-genome sequencing: a useful, easy to apply molecular technique for CNA detection on FFPE tumor tissue—a glioma-driven study. Virchows Arch. https://doi.org/10.1007/s00428-022-03268-w

Van Otterdijk, S.D., Michels, K.B., 2016. Transgenerational epigenetic inheritance in mammals: How good is the evidence? FASEB J. 30, 2457–2465. https://doi.org/10.1096/fj.201500083

Vandeputte, M., Dupont-Nivet, M., Haffray, P., Chavanne, H., Cenadelli, S., Parati, K., Vidal, M.O., Vergnet, A., Chatain, B., 2009. Response to domestication and selection for growth in the European sea bass (*Dicentrarchus labrax*) in separate and mixed tanks. Aquaculture 286, 20–27.

Vandeputte, M., Kocour, M., Mauger, S., Dupont-Nivet, M., De Guerry, D., Rodina, M., Gela, D., Vallod, D., Chevassus, B., Linhart, O., 2004. Heritability estimates for growth-related traits using microsatellite parentage assignment in juvenile common carp (*Cyprinus carpio L.*). Aquaculture 235, 223–236. https://doi.org/10.1016/j.aquaculture.2003.12.019

Vandeputte, M., MAUGER, S., DUPONT-NIVET, M., 2006. An evaluation of allowing for mismatches as a way to manage genotyping errors in parentage assignment by exclusion. Mol. Ecol. Notes 6, 265–267. https://doi.org/10.1111/j.1471-8286.2005.01167.x

Vandeputte, M., Rossignol, M.-N., Pincent, C., 2011. From theory to practice: Empirical evaluation of the assignment power of marker sets for pedigree analysis in fish breeding. Aquaculture 314, 80–86. https://doi.org/10.1016/j.aquaculture.2011.01.043

Varsamos, S., Nebel, C., Charmantier, G., 2005. Ontogeny of osmoregulation in postembryonic fish: A review. Comp. Biochem. Physiol. - A Mol. Integr. Physiol. 141, 401–429.

VASEP, 2021. Vietnam Association of Seafood Exporters and Producers (VASEP). http://vasep. com.vn.

Vu, N.T., Ha, T.T.T., Thuy, V.T.B., Trang, V.T., Nguyen, N.H., 2020. Population genomic analyses of wild and farmed striped catfish *pangasianodon hypophthalmus* in the lower Mekong river. J. Mar. Sci. Eng. 8. https://doi.org/10.3390/JMSE8060471

Vu, N.T., Sang, N. Van, Trong, T.Q., Duy, N.H., Dang, N.T., Nguyen, N.H., 2019. Breeding for improved resistance to Edwardsiella ictaluri in striped catfish (*Pangasianodon hypophthalmus*): Quantitative genetic parameters. J. Fish Dis. 42, 1409–1417. https://doi.org/10.1111/jfd.13067

Viet, N.T., Jung, H., Thanh, N.M., Hurwood, D., Mather, P., 2016. Marine Genomics Evaluation of potential candidate genes involved in salinity tolerance in striped catfish ( *Pangasianodon hypophthalmus* ) using an RNA-Seq approach. Mar. Genomics 25, 75–88.

Vu, N.T., Trong, T.Q., Nguyen, N.H., 2017. Effects of selection for fast growth on survival rate during grow-out phase in giant freshwater prawn (*Macrobrachium rosenbergii*). BMC Genet. 18.

Vougiouklakis T, Nakamura Y, Saloura V (2017) Critical roles of protein methyltransferases and demethylases in the regulation of embryonic stem cell fate. Epigenetics 12:1015–1027

Vollger, M.R., Logsdon, G.A., Audano, P.A., Sulovari, A., Porubsky, D., Peluso, P., Wenger, A.M., Concepcion, G.T., Kronenberg, Z.N., Munson, K.M., Baker, C., Sanders, A.D., Spierings, D.C.J., Lansdorp, P.M., Surti, U., Hunkapiller, M.W., Eichler, E.E., 2020. Improved assembly and variant detection

of a haploid human genome using single-molecule, high-fidelity long reads. Ann. Hum. Genet. 84, 125–140. https://doi.org/10.1111/ahg.12364

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

VASEP, 2019. Vietnam Association of Seafood Exporters and Producers (VASEP). http://vasep.com.vn.

VAWR, 2017. Reports on salinity intrusion in 2016 and warning for agriculture in Mekong Delta, Southern institute of water resource research. Vietnam academy for water resources (VAWR). www.siwrr.org.vn.

Villanueva, B., Verspoor, E., Visscher, P.M., 2002. Parental assignment in fish using microsatellite genetic markers with finite numbers of parents and offspring. Anim. Genet. 33, 33–41.

Visscher, P.M., Hill, W.G., Wray, N.R., 2008. Heritability in the genomics era — concepts and misconceptions. Nat. Rev. Genet. 9, 255–266. https://doi.org/10.1038/nrg2322

Vu, N.T., Van Sang, N., Phuc, T.H., Vuong, N.T., Nguyen, N.H., 2019. Genetic evaluation of a 15-year selection program for high growth in striped catfish *Pangasianodon hypophthalmus*. Aquaculture 509, 221–226. https://doi.org/10.1016/j.aquaculture.2019.05.034

**-W-**

Wang, J. (2012). Computationally efficient sibship and parentage assignment from multilocus marker data. *Genetics*, *191*(1), 183–194. doi: 10.1534/genetics.111.138149

Wang, J. (2019). Pedigree reconstruction from poor quality genotype data. *Heredity*, *122*(6), 719–728. doi: 10.1038/s41437-018-0178-7

Wassmann, R., Hien, N.X., Hoanh, C.T., Tuong, T.P., 2004. Sea level rise affecting the Vietnamese Mekong Delta: Water elevation in the flood season and implications for rice production. Clim. Change 66, 89–107. https://doi.org/10.1023/B:CLIM.0000043144.69736.b7

Wen, M., Pan, Q., Jouanno, E., Montfort, J., Zahm, M., Cabau, C., Klopp, C., Iampietro, C., Roques, C., Bouchez, O., Castinel, A., Donnadieu, C., Parrinello, H., Poncet, C., Belmonte, E., Gautier, V., Avarre, J., Dugue, R., Gustiano, R., Hà, T.T.T., Campet, M., Sriphairoj, K., Ribolli, J., de Almeida, F.L., Desvignes, T., Postlethwait, J.H., Floi Bucao, C., Robinson-Rechavi, M., Bobe, J., Herpin, A., Guiguen, Y., 2022. An ancient truncated duplication of the anti-Mullerian hormone receptor type 2 gene is a potential

conserved master sex determinant in the Pangasiidae catfish family. Mol. Ecol. Resour. 1–18. https://doi.org/10.1111/1755-0998.13620

White, I., 2002. Water management in the Mekong Delta: changes, conflicts and opportunities; Technical documents in hydrology; Vol.:61; 2002.

Wong, L.L., Razali, S.A., Deris, Z.M., Danish-Daniel, M., Tan, M.P., Nor, S.A.M., Ma, H., Min, W., Yantao, L., Asaduzzaman, M., Sung, Y.Y., Liu, Z., Sorgeloos, P., Van de Peer, Y., Afiqah-Aleng, N., 2022. Application of second-generation sequencing (SGS) and third generation sequencing (TGS) in aquaculture breeding program. Aquaculture 548, 737633.

Wendelaar-Bonga, S.E. (2011) Hormonal response to stress. In: Farrel AP (ed) Encyclopedia of fish: fish physiology from genome to environment. Academic Press, p 1515–1523

Whalen, A., Gorjanc, G., & Hickey, J. M. (2019). Parentage assignment with genotyping-by-sequencing data. *Journal of Animal Breeding and Genetics*, *136*(2), 102–112. doi: 10.1111/jbg.12370

Whalen, A., Ros-Freixedes, R., Wilson, D. L., Gorjanc, G., & Hickey, J. M. (2018). Hybrid peeling for fast and accurate calling, phasing, and imputation with sequence data of any coverage in pedigrees. *Genetics Selection Evolution*, *50*(1), 1–15. doi: 10.1186/s12711-018-0438-2

Wang, C.M., Lo, L.C., Zhu, Z.Y., Lin, G., Feng, F., Li, J., Yang, W.T., Tan, J., Chou, R., Lim, H.S., Orban, L., Yue, G.H., 2008a. Estimating reproductive success of brooders and heritability of growth traits in Asian sea bass (*Lates calcarifer*) using microsatellites. Aquac. Res. 39, 1612–1619.

Wang, J., Santure, A.W., 2009. Parentage and sibship inference from multilocus genotype data under polygamy. Genetics 181, 1579–1594. https://doi.org/10.1534/genetics.108.100214

Wang, Jun, Zhang, Guofan, Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H., Xiong, Z., Que, H., Xie, Y., Holland, P.W.H., Paps, J., Zhu, Y., et al,. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490, 49–54.

Wang, L., Wu, J., Liu, X., Di, D., Liang, Y., Feng, Y., Zhang, S., Li, B., Qi, X.G., 2019. A high-quality genome assembly for the endangered golden snub-nosed monkey (Rhinopithecus roxellana). Gigascience 8, 1–11. https://doi.org/10.1093/gigascience/giz098

Wang, X., 2016. A Genome-Wide Association Study for Low Oxygen Tolerance in Catfish using the 250K SNP Array. Plant and Animal Genome XXIV Conference, Plant and Animal Genome. https://doi.org/https://pag.confex.com/pag/xxiv/webprogram/Paper22097.html

Wang, Y., Wang, C., Zhang, J., Chen, Y., Zuo, Z., 2009. DNA hypomethylation induced by tributyltin, triphenyltin, and a mixture of these in Sebastiscus marmoratus liver. Aquat. Toxicol. 95, 93–98. https://doi.org/10.1016/j.aquatox.2009.06.008

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq a revolutionary tool for transcriptomics 10, 57–63. https://doi.org/10.1038/nrg2484.RNA-Seq

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M., 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One 9, e112963.

Walter RB, Li H-Y, Intano GW, et al (2002) Absence of global genomic cytosine methylation pattern erasure during medaka (*Oryzias latipes*) early embryo development. Comp Biochem Physiol Part B Biochem Mol Biol 133:597–607

Wang L, Zhang J, Duan J, et al (2014) Programming and inheritance of parental DNA methylomes in mammals. Cell 157:979–991

Wang X, Bhandari RK (2019) DNA methylation dynamics during epigenetic reprogramming of medaka embryo. Epigenetics 14:611–622. https://doi.org/10.1080/15592294.2019.1605816

Wu X, Zhang Y (2017) TET-mediated active DNA demethylation: mechanism, function and beyond. Nat Rev Genet 18:517–534. https://doi.org/10.1038/nrg.2017.33

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019b. Highly-accurate long-read sequencing improves variant detection and assembly of a human genome. bioRxiv. https://doi.org/10.1101/519025

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019a. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 37, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9.

Weaver, I.C.G., Cervoni, N., Champagne, F.A., D'Alessio, A.C., Sharma, S., Seckl, J.R., Dymov, S., Szyf, M., Meaney, M.J., 2004. Epigenetic programming by maternal behavior. Nat. Neurosci. 7, 847–854. https://doi.org/10.1038/nn1276

Whibley, A., Kelley, J.L., Narum, S.R., 2021. The changing face of genome assemblies: Guidance on achieving high-quality reference genomes. Mol. Ecol. Resour. 21, 641–652.

Wright, J.M., Bentzen, P., 1994. Microsatellites: genetic markers for the future. Rev. Fish Biol. Fish. 4, 384–388. https://doi.org/10.1007/BF00042912

**-X-**

Xiaofei, Y., Hendrik-Jan, M., John W.M, B., Langquing, L., Martien A.M, G., Komen, H., 2021. Genomic selection and introgression signatures resulting from adaptation to high salinity: A case study on an Indonesian farmed tilapia (*Sukamandi*) strain. Proc. Aquac. Eur. 21, Maderia, Port. Oct. 4-7,2021 42.

Xu, P., Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., Xu, J., Zheng, X., Ren, L., Wang, G., Zhang, Y., Huo, L., Zhao, Z., Cao, D., Lu, C., Li, C., Zhou, Y., Liu, Z., Fan, Z., Shan, G., Li, X., Wu, S., Song, Lipu, Hou, G., Jiang, Y., Jeney, Z., Yu, D., Wang, L., Shao, C., Song, Lai, Sun, J., Ji, P., Wang, Jian, Li, Q., Xu, L., Sun, F., Feng, J., Wang, C., Wang, S., 2014. Genome sequence and genetic diversity of the common carp, Cyprinus carpio. Nat. Genet. 46, 1212–1219. https://doi.org/10.1038/ng.3098

**-Y-**

Yamamoto, T., Murashita, K., Matsunari, H., Oku, H., Furuita, H., Okamoto, H., Amano, S., Suzuki, N., 2016. Amago salmon Oncorhynchus masou ishikawae juveniles selectively bred for growth on a low fishmeal diet exhibit a good response to the low fishmeal diet due largely to an increased feed intake with a particular preference for the diet. Aquaculture 465, 380–386.

Yamamoto, T., Okamoto, H., Furuita, H., Murashita, K., Matsunari, H., Iwashita, Y., Amano, S., Suzuki, N., 2014. Growth performance and physiological condition of F1 amago salmon Oncorhynchus masou ishikawae juveniles obtained from broodstock with selective breeding for growth on a low fish-meal diet. Fish. Sci. 80, 569–579. https://doi.org/10.1007/s12562-014-0707-0

Yáñez, J.M., Newman, S., Houston, R.D., 2015. Genomics in aquaculture to better understand species biology and accelerate genetic progress. Front. Genet. 6, 1–3. https://doi.org/10.3389/fgene.-2015.00128

Yang, X., Zhao, X., Qu, S., Jia, P., Wang, B., Gao, S., Xu, T., Zhang, W., Huang, J., Ye, K., 2022. Haplotype-resolved Chinese male genome assembly based on high-fidelity sequencing. Fundam. Res. https://doi.org/10.1016/j.fmre.2022.02.005

Yada T, Muto K, Azuma T, Ikuta K (2004) Effects of prolactin and growth hormone on plasma levels of lysozyme and ceruloplasmin in rainbow trout. Comp Biochem Physiol C Toxicol Pharmacol 139:57–63. https://doi.org/10.1016/j.cca.2004.09.003

Yoshizaki, G., Yazawa, R., 2019. Application of surrogate broodstock technology in aquaculture. Fish. Sci. 85, 429–437. https://doi.org/10.1007/s12562-019-01299-y

Yue, G.H., Wang, L., 2017. Current status of genome sequencing and its applications in aquaculture. Aquaculture 468, 337–347. https://doi.org/10.1016/j.aquaculture.2016.10.036

Yamamoto T, Murashita K, Matsunari H, et al (2016) Amago salmon Oncorhynchus masou ishikawae juveniles selectively bred for growth on a low fishmeal diet exhibit a good response to the low fishmeal diet due largely to an increased feed intake with a particular preference for the diet. Aquaculture 465:380–386

Ye J, Coulouris G, Zaretskaya I, et al (2012) Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 13:134. https://doi.org/10.1186/1471-2105-13-134

Ye, B., Wan, Z., Wang, L., Pang, H., Wen, Y., Liu, H., Liang, B., Lim, H.S., Jiang, J., Yue, G., 2017. Heritability of growth traits in the Asian seabass (Lates calcarifer). Aquac. Fish. 2, 112–118. https://doi.org/10.1016/j.aaf.2017.06.001

Yu, A., Lepère, G., Jay, F., Wang, J., Bapaume, L., Wang, Y., Abraham, A.L., Penterman, J., Fischer, R.L., Voinnet, O., Navarro, L., 2013. Dynamics and biological relevance of DNA demethylation in Arabidopsis antibacterial defense. Proc. Natl. Acad. Sci. U. S. A. 110, 2389–2394.

Yue, G.H., 2014. Recent advances of genome mapping and marker-assisted selection in aquaculture. Fish Fish. 15, 376–396. https://doi.org/10.1111/faf.12020

Yue, G.H., Wang, L., 2017. Current status of genome sequencing and its applications in aquaculture. Aquaculture 468, 337–347. https://doi.org/10.1016/j.aquaculture.2016.10.036

Yue, G.H., Xia, J.H., 2014a. Practical Considerations of Molecular Parentage Analysis in Fish. J. World Aquac. Soc. 45, 89–103. https://doi.org/10.1111/jwas.12107

Yue, G.H., Xia, J.H., 2014b. Practical Considerations of Molecular Parentage Analysis in Fish 45, 89–103. https://doi.org/10.1111/jwas.12107.

**-Z-**

Zak, T., Deshev, R., Benet-Perlberg, A., Naor, A., Magen, I., Shapira, Y., Ponzoni, R.W., Hulata, G., 2014. Genetic improvement of Israeli blue (Jordan) tilapia, Oreochromis aureus (Steindachner), through selective breeding for harvest weight. Aquac. Res. 45, 546–557. https://doi.org/10.1111/are.12072

Zhang, J., Cao, F., Liu, J., Yuan, R., 2017. Genetic parameters for growth and survival traits in *Litopenaeus vannamei* at different ages. Aquac. Int. 25, 1901–1911. https://doi.org/10.1007/s10499-017-0163-9

Zenger, K.R., Khatkar, M.S., Jones, D.B., Khalilisamani, N., Jerry, D.R., Raadsma, H.W., 2019. Genomic selection in aquaculture: Application, limitations and opportunities with special reference to marine shrimp and pearl oysters. Front. Genet. 10. https://doi.org/10.3389/fgene.2018.00693

Ziv, G., Baran, E., Nam, S., Rodríguez-Iturbe, I., Levin, S.A., 2012. Trading-off fish biodiversity, food security, and hydropower in the Mekong River Basin. Proc. Natl. Acad. Sci. U. S. A. 109, 5609–5614. https://doi.org/10.1073/pnas.1201423109

Zeng Y, Chen T (2019) DNA Methylation Reprogramming during Mammalian Development. Genes 10

Zhang W, Feng G, Wang L, et al (2018) MeCP2 deficiency promotes cell reprogramming by stimulating IGF1/AKT/mTOR signaling and activating ribosomal protein-mediated cell cycle gene translation. J Mol Cell Biol 10:515–526. https://doi.org/10.1093/jmcb/mjy018

Zhang X, Li H, Qiu Q, et al (2014) 2, 4-Dichlorophenol induces global DNA hypermethylation through the increase of S-adenosylmethionine and the upregulation of DNMTs mRNA in the liver of goldfish *Carassius auratus*. Comp Biochem Physiol Part C Toxicol Pharmacol 160:54–59

Zhou M, Zhao Z, Zhao J, et al (2021) Gene expression profiling of DNA methyltransferase genes in *Siniperca chuatsi* based on transcriptome sequencing. J Fish Biol 99:1755–1760.

Zimin, A. V., Salzberg, S.L., 2022. The SAMBA tool uses long reads to improve the contiguity of genome assemblies. PLoS Comput. Biol. 18, 1–11. https://doi.org/10.1371/journal.pcbi.1009860.

Zemach, A., McDaniel, I.E., Silva, P., Zilberman, D., 2010. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. Science (80-. ). 328, 916–919. https://doi.org/10.1126/-science.1186366

Zhang, C., Hoshida, Y., Sadler, K.C., 2016. Comparative epigenomic profiling of the DNA methylome in mouse and Zebrafish uncovers high interspecies divergence. Front. Genet. 7, 1–15. https://doi.org/10.3389/fgene.2016.00110

Zhang, Xiaojun, Yuan, J., Sun, Y., Li, S., Gao, Y., Yu, Y., Liu, C., Wang, Q., Lv, X., Zhang, Xiaoxi, Ma, K.Y., Wang, X., Lin, W., Wang, Long, Zhu, X., Zhang, C., Zhang, J., Jin, S., Yu, K., Kong,J., Xu, P., Chen, J., Zhang, H., Sorgeloos, P., Sagi, A., Alcivar-Warren, A., Liu, Z., Wang, Lei, Ruan, J., Chu, K.H., Liu, B., Li, F., Xiang, J., 2019. Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. Nat. Commun. 10. https://doi.org/10.1038/s41467-018-08197-4

Zhao, L., Schöne, B.R., Mertz-Kraus, R., Yang, F., 2017. Sodium provides unique insights into transgenerational effects of ocean acidification on bivalve shell formation. Sci. Total Environ. 577, 360–366. https://doi.org/10.1016/j.scitotenv.2016.10.200

Zhou, H., Ma, T.Y., Zhang, R., Xu, Q.Z., Shen, F., Qin, Y.J., Xu, W., Wang, Y., Li, Y.J., 2016. Analysis of different ploidy and parent–offspring genomic DNA methylation in the loach: Misgurnus anguillicaudatus. Int. J. Mol. Sci. 17. https://doi.org/10.3390/ijms17081299.

# Appendices

# Appendices

## 1. Identification of parentage using shallow sequencing data (manual) of Chapter 3

### 1. Context

Obtaining parentage information based on genomic data is useful in situations where a set of potential parents is available for an individual. For example, obtaining relationships between individuals in ecological studies can help to better understand the population dynamics of invasive species. Or identifiying the parents of fish cultured in a breeding experiment is necessary to optimize the genetic values of the selected individuals while preserving the genetic diversity and controlling inbreeding. More generally, parentage identification can be useful in situations where the parents cannot be directly observed. This is for example the case for many wild animals and for aquaculture species, where the small size of the progeny at birth and the necessity to pool families for efficiency reasons and for a fair comparison of the breeders make posterior identification of the parents necessary.

« **shallowped** » is a software that is designed to identify the closest relatives of an offspring in a set of potential relatives using data obtained from a (shallow) sequencing experiment and to infer the relationship (if any) between this offspring and its closest parents in the dataset. The software implements the likelihood approach described in the accompanying paper “Parentage assignment with shallow whole genome sequencing data”. The program reads parameters (see below), computes a likelihood in order to infer the most likely parents, generates the data used to infer the needed distributions used to assess the significance of the relationships, collects the results and presents the inferences in a simple form. In brief, the program evaluates the likelihood that a couple is related to a tested individual for all possible couples from the set of parents. The likelihood of the couple with the highest likelihood is then compared to various distributions to infer the exact relationship of the putative parents with the tested individual.

### 2. Installation

The program is a standalone program (*shallowped.f95*), and the compilation is straightforward. The program has only been tested on Windows (10 professional) and on Linux Ubuntu, but, due to the absence of specific features and dependencies, we don’t expect problems on other platforms (for example, recent Mac can use a GNU fortran compiler 1 similar to the one used for Ubuntu).

#### 2.1 Linux Ubuntu

Provided the (free) GNU fortran has been previously properly installed, the f95 compiler should be available from any directory in your account (Figure 1).

```
fred@fmvpa-stat51:~$ f95 --version
GNU Fortran (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0
Copyright (C) 2019 Free Software Foundation, Inc.
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.

fred@fmvpa-stat51:~$
```

Figure 1: Ubuntu command showing the versions of the compiler and of the operating system.

The installation of **shallowped** is as follows:

1. Create a directory to save the source program. For example:

   **mkdir ~/SHALLOWPED**

will create a directory named **SHALLOWPED** in your home directory.

2. Dowload the program into this new directory. If you downloaded into another directory, you can move it to the newly created directory. For example, to move the program from your home directory to the newly created directory, you can type:

   **mv ~/*shallowped.f95*  ~/SHALLOWPED**

This should have moved the program file to the **~/SHALLOWPED** directory.

3. Go to the ~/SHALLOWPED directory:

   **cd ~/SHALLOWPED**

4. Compile the program:

   **f95 *shallowped.f95* -o shallowped**

This should generate an executable file shallowped in the current directory.

5. If you want the executable file to be available from any directory in your account, you can either:

- Type the full path of the executable. In our example, we can start the program from any directory by typing:

**~/SHALLOWPED/shallowped**

Move the executable file to one of the directories present in your PATH vari- able. The list (separated using ":") of these directories containing executables can be obtained by typing:

**echo** $**PATH**

Typically, ~/bin will be one of those directories and the command to move the executable to that directory is:

**mv ~/SHALLOWPED/shallowped ~/bin**

- Create a link to the executable file in one of the directories contained in PATH. For example:

**ln -s ~/SHALLOWPED/shallowped ~/bin/shallowped** will create a symbolic link to the executable. Since ~/bin is in PATH, this will make the executable available from everywhere.

**2.2 Windows**

Free versions of fortran 95 compilers are available for private use (for a more professional use, a license is need). For example, FTN95 is distributed with a GUI making it easy to visualize, modify and compile the program if needed. Using such software, the compilation is very simple:

1. Download the source (***shallowped.f95***) into a target directory (for example ~**MyDocs****SHALLOWPED**),
2. Open the graphical user interface (named Plato for FTN95),
3. Open the file (for example, ~**MyDocs****SHALLOWPED*****shallowped.f95***) us- ing the *File* menu,
4. Use the *Build* menu to compile and build the file (Figure 2),
5. That's all! You should have a **shallowped.exe** file in the directory containing the source file (in our example, ~**MyDocs****SHALLOWPED****shallowped.exe**). You may want to move that file to a more appropriate place if needed.

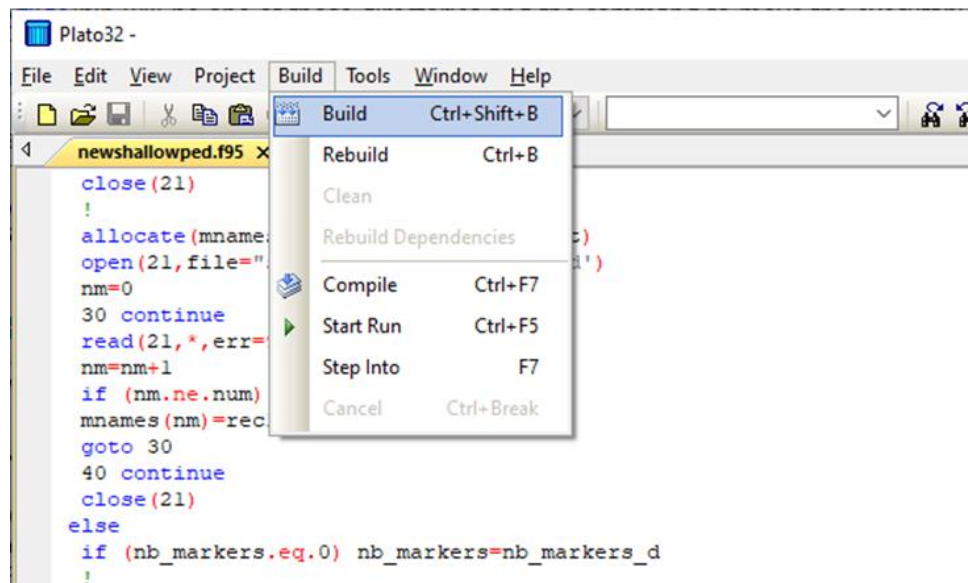


Figure 2: Building the executable in the Plato GUI interface.

## 3. Data preparation

The data is provided to the program using 4 files. One file (shallow.prm) contains parameters to be used by the program. Since defaults are provided for these parameters, the file needs only be used when the values of the targeted parameters have to be modified. The other 3 files (shallow.ind, shallow.gen, shallow.mrk) contain the data. All these files will be described in the next sections.

### 3.1 Individuals file **shallow.ind**

This mandatory file is used to give the identifications of the individuals considered in the experiment along with their status. Each line contains the information on one individual. The format of the file is:

<number> <name> <status>

where <number> is the individual number, ranging from 1 to NI, with NI the total number of individuals in the analysis, <name> the name of the corresponding individ- uals, and <status> the status of the individual, which is either *F (*for a father), *M* (for a mother) or *O* (for an offspring). The goal of the program is to reconstruct triplets of *F, M* and *O* individuals where *F* and *M* are *O*'s closest relatives, and to infer the most likely relation for these closest relatives (father, grandfather, sibs...).

### 3.2 Reads file **shallow.gen**

This mandatory file contains the information extracted from the shallow sequencing experiment. It is assumed that polymorphic loci have previously been identified in the genome (in the current version, shallowped uses only biallelic markers, such as SNP for example, although the approach could easily be extended to polyallelic markers), and that the number of reads for each locus and each individual have been obtained before using shallowped. Consequently, each line contains the information for one marker and one individual. The format is:

<number_i> <number_m> <reads_1> <reads_2>

where <number_i> is the number corresponding to the individual (the number provided in the individuals file), <number_m> is the number corresponding to the marker (the number provided in the markers file), and <reads_1> and <reads_2> are the number of reads for each of the 2 alleles for this marker and this individual. When no information is available for a combination individual*marker, the line can be omitted from the file (in other words, the default is "<number_i> <number_m> 0 0" for each combination).

### 3.3 Markers file shallow.mrk

This file is used to allow providing a list of markers names. The format is again very simple:

<number> <name>

where <number> is the marker number, ranging from 1 to NM, with NM the total number of markers in the analysis, and <name> the name of the corresponding markers (names are limited to 30 characters). If this file is not provided, the parameter NMRK should be given in the parameters file (the markers name would then be M1, M2, ...). If neither the file and the NMRK parameter are provided, the default number of markers will be used (which might not correspond to the genomic data).

### 3.4 Parameters file shallow.prm

Although all parameters have default values, this optional file can be used to provide different values to the parameters in a run. Each line of the file has the same format:

<keyword> <value>

where <keyword> is a four characters' keyword specifying the parameter and <value> is the value to be attributed to the corresponding parameter. In the following description of the parameters, many options are used to specify the number of genotypes that will be generated to obtain the distribution of the

genotypes probabilities under various hypotheses. These distributions are written P [*G*(*i*)/*d*(*j*), *d*(*k*)]. This notation means "the distribution of *i*'s genotypes, given genomic data on *j* and *k*, where *i*, *j* and *k* are either *f* (father), *m* (mother), *o* (offspring), *pgf* (paternal grandfather), *mgm* (maternal grandmother), *ffs* (full-sib of the father), *mfs* (full-sib of the mother), *unf* (unrelated to the father) or *unm* (unrelated to the mother). The alphabetic list of the possible parameters is as follows:

1. FREQ < *mode* > (Default: 1)

where < *mode* > is either 1, 2 or 3. When < *mode* >= 1, the frequencies are based on counts (the number of copies of an allele is incremented each time the corre- sponding number of reads is positive for a potential parent). When < *mode* >= 2, the frequencies are estimated using the iterative procedure described in the paper. And when < *mode* >= 3, th frequencies are read from **shallow.freqs** (an error occurs if the file is not available).

2. LIKE < mode > (Default : 1)

where < mode > is either 1 or 2. When < mode >= 1, the computation of the log-likelihoods use the formula based on the posterior probabilities $\frac{\prod_{i=1}^{n} P_k(i)}{\sum_{k=1}^{h} \prod_{i=1}^{n} P_k(i)}$ . When < mode >= 2, the computation uses $\prod_{k=1}^{n} \frac{Pk(i)}{\sum_{k=1}^{h} Pk(i)}$, where n and h are the number of individuals and the number of hypotheses, respectively. < mode >= 2 can be useful in situations where < mode >= 1 cannot be computed due to underflows.

3. NFAT < nb_fathers > (Default : 200)

where < nb_fathers > is the number of fathers that will be simulated to generate the father likelihoods distribution, assuming that the "best" mother is the real mother, and using information from that mother and from the offspring. Symbolically : P[G(f)|d (m = M), d(o)].

4. NFSF < nb_fathers_fullsibs > (Default: 200)

where < nb_fathers fullsibs > is the number of father's fullsibs that will be simulated to generate the fathers' fullsibs likelihoods distribution, assuming that the "best" mother is the real mother, and using information from that mother and from the offspring. Symbolically : P[G(ffs)|d (m = M), d(o)].

5. NFSM < nb_mothers_fullsibs > (Default: 200)

where < nb_mothers_fullsibs > is the number of mother's fullsibs that will be simulated to generate the mothers' fullsibs likelihoods distribution, assuming that the "best" father is the real father, and using information from that father and from the offspring. Symbolically : P[G(mfs)|d (f = F), d(o)].

6. NHSF < nb_fathers halfsibs > (Default: 200)

where < nb_fathers_halfsibs > is the number of father's halfsibs that will be simulated to generate the fathers' halfsibs likelihoods distribution, assuming that the "best" mother is the real mother, and using information from that mother and from the offspring. Symbolically : P[G(hsf)|d (m = M), d(o)].

7. NHSM < nb_mothers_halfsibs > (Default: 200)

where < nb_mothers_halfsibs > is the number of mother's halfsibs that will be simulated to generate the mothers' halfsibs likelihoods distribution, assuming that the "best" father is the real father, and using information from that father and from the offspring. Symbolically : P[G(hsm)|d (f = F), d(o)].

8. NMGM < nb_maternal_grandmothers > (Default: 200)

where < nb_maternal_grandmothers > is the number of maternal grandmothers that will be simulated to generate the maternal grandmother likelihoods distribution, assuming that the "best" father is the real father, and using information from that father and from the offspring. Symbolically : P[G(mgm)|d (f = F), d(o)].

9. NMOT < nb_fathers > (Default : 200)

where < nb_mothers > is the number of mothers that will be simulated to generate the mother likelihoods distribution, assuming that the "best" father is the real father, and using information from that father and from the offspring. Symbolically : P[G(m)|d (f = F), d(o)].

10. NMRK < nb_markers > (Default : 0)

where < nb_markers > is the number of markers in the analysis. When set to 0 (the default), the number of markers will be read from the **shallow.mrk** file. This option is only useful when the **shallow.mrk** file is not present. The specified number of markers should match the number of markers used in **shallow.gen**.

11. NPGF < nb_paternal_grandfathers > (Default: 200)

where < nb_paternal_grandf athers > is the number of paternal grandfathers that will be simulated to generate the paternal grandfather likelihoods distribution, assuming that the "best" mother is the real mother, and using information from that mother and from the offspring. Symbolically : P[G(pgf)|d (m = M), d(o)].

12. NUNF < nb_unrelated_to-father > (Default: 200)

where < nb-unrelated_to_father > is the number of unrelated individuals that will be simulated to generate the unrelated individual likelihoods distribution. Symbolically : P[G(unf)|d (m = M), d(o)].

13. NUNM < nb_unrelated_to_mother > (Default: 200)

where < nb_unrelated_to_mother > is the number of unrelated individuals that will be simulated to generate the unrelated individual likelihoods distribution. Symbolically : P[G(unm)|d (f = F), d(o)].

14. SEQE <probability> (Default: 0.01)

    where <probability> is the probability of a sequencing error at a given marker, changing an allele to another allele. This default value is quite high if, for example, used markers are a set of SNP selected on the quality of the reads underlying these SNP (making sequencing errors less likely).

If the **shallow.prm** file is not provided, all default values are used for the parameters.

**4. Running the program**

When data files are ready, the program can be run by simply typing in a shell (linux - Figure 3) or in a command box (Windows):

**shallowped**

The program will then perform the computations described in the companion paper:

1. Read the **shallow.mrk** file (if present) and display the number of markers that will be used for the analysis.
2. Read the **shallow.ind** file and display the number of individuals (and their status) that will be used for the analysis.
3. Read the **shallow.gen** file and report the number of combinations markers*individuals for which one or several reads are available. In addition, the minimal and maximal individual's covers will be output.
4. Allelic frequencies in the parental generation are estimated (using either the itera- tive algorithm suggested in the paper, reads counts or a provided frequencies file, see option FREQ), and reported in the **shallow.freqs** file.
5. Obtain the most likely (offspring - father - mother) triplet using the likelihood approach detailed in the paper. This will be performed for each offspring in the **shallow.ind** file. All computed likelihoods are reported in the **shallow.lik** file.
6. Check the significance under various null hypotheses (see below).
7. Report the likelihood distributions under the various tested null hypotheses in the file **shallow.lik.hyp**.
8. Report the p-values associated to all these hypotheses in the **shallow.fam** file. In addition, the most likely relatives for each individual are provided in **shallow.top**, and the significant relatives are provided in **shallow.ped**.

Based on these results, it is possible to statistically infer the exact relationship between the "best couples" and the tested offspring.

```
shallow.gen  shallow.ind  shallow.mrk
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL3$ shallowped
Starting SHALLOWPED...
--> Markers # =   2000
--> Individuals # =     41
---> # fathers =     20
---> # mothers =     20
---> # offspring =     1
--> # of reads:   410324
--> Covers range from   4.922 to   5.118
--> P[d|g] computed
--> Markers frequencies computed using counts
===> O = 41          F = 20          M = 21          (lik =       -0.00*      -0.00)
--> Best families computed
--> Simulation for offspring    1/    1
Leaving SHALLOWPED...
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL3$ ls
shallow.fam  shallow.frq  shallow.gen  shallow.ind  shallow.lik  shallow.lik.hyp
shallow.mrk  shallow.ped  shallow.top
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL3$ _
```

Figure 3: Executing SHALLOWPED in a linux shell: the first **ls** command lists the files in the current directory before running the program. Note that **shallow.prm** is not present: all default values will be used for the parameters. The **shallowped** command runs the program and displays some information on the run (number ofmarkers, number of individuals, tested hypotheses, ...). The second **ls** command lists the content of the directory after running the program. Additional files (**shallow.fam**, **shallow.frq**, **shallow.lik**, **shallow.lik.hyp**) will be described below.

**5. An example**

To allow readers to perform tests on simulated data, we have written a perl script **mksimul.pl** that is able to generate the needed input files in the correct format. The script is provided in the appendix of this manual. A minimal help is available by typing **mksimul.pl -h** (Figure 4)

```
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$ perl mksimul.pl -h
Usage: [perl] mksimul.pl [options]
 Possible options:
 -NBM nm (number of markers; default: 10000)
 -MEAN m (Poisson mean for reads; default: 5)
 -PED pedig (real pedigree; default: pedigree)
   Format: Indiv Father Mother Status [Cover]
           Status is O|F|M|X (X are not kept in the results)
           Cover is optional. When used, it supersedes MEAN
 -ALPHA Y|N (report AlphaAssign files; default: N)
 -SHALLOW Y|N (report Shallowped files; default: Y)
 -SEQE se (sequencing error rate; default: 0.01)
 -H (this help)

fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$
```

Figure 4: Running **mksimul.pl** with the '-h' option provides a little help on the use of the script. We will demonstrate the use of the script on a small example. We will consider the pedigree depicted in Figure 5.

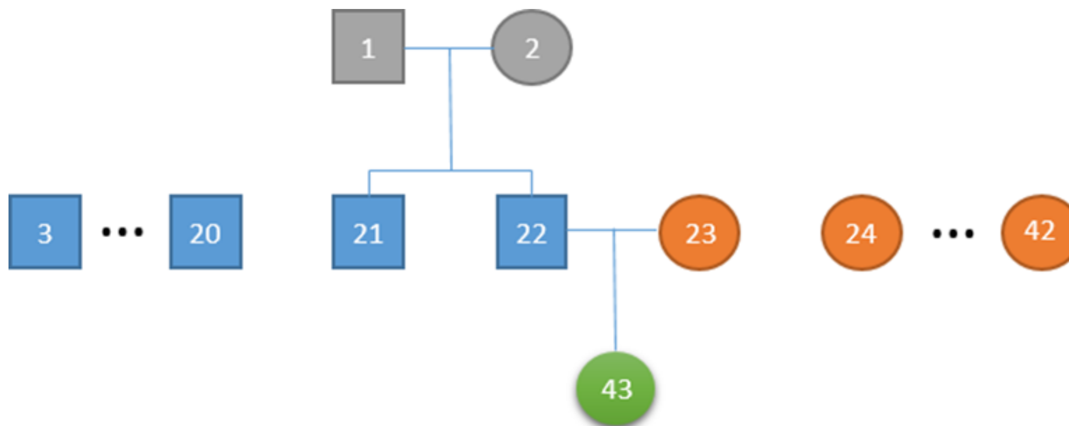


Figure 5: Example of pedigree used in the simulation program. The two grand-parents (1 and 2) will be used to generate the data, but will not be reported in the data files. The blue individuals (3-22) make up the set of possible fathers, and the orange ones (23-42) the set of possible mothers of the offspring (43). Note that in that simuation, a full-sibs of the "real" father will be present in the list of potential fathers.

We will use the default parameters for the options -NBM (i.e. 10,000 markers), -MEAN (i.e. an average of 5 reads/genotype), -PED (i.e. the simulated pedigree will be stored in a file named **pedigree**), -

SHALLOW (i.e. report the files needed to run **shallowped**) and -SEQE (i.e. use an sequencing error rate of 0.01). We will generate the files needed to run **AlphaAssign** (the software implementing the approach described in (Whalen et al., 2019)) by specifying the option "-ALPHA Y" in order to compare the results obtained with the 2 methods. Let's first take a look at the **pedigree** file:

```
1 0  0  X
2 0  0  X
3 0  0  F
4 0  0  F
5 0  0  F
- - -
20 0 0 F
21 1 2 F
22 1 2 F
23 0 0 M
24 0 0 M
- - -
41 0 0 M
42 0 0 M
43 22 23 O
```

As mentioned in the help screen, individuals with a 'X' status (i.e. the grandparents 1 and 2) will not be reported, individuals with a 'F' status will make up the list of potential fathers (i.e. individuals 3-22), individuals with a 'M' status will make up the list of potential mothers (i.e. individuals 23-42), and the only offspring (status 'O') is **individual 43.**

After creating this file, we run the simulation script by typing (Figure 6):

perl **mksimul.pl -alpha y**

```
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$ perl $s/mksimul.pl -alpha y
-> Reading the pedigree
--> 43 individuals in the pedigree
---> 20 fathers, 20 mothers, 1 offspring
--> Reporting the needed files for AlphaAssign
--> Reporting the needed file for shallowped
-> Creating markers file for shallowped
--> Done !
-> Creating the genotypes
--> Using 10000 markers
--> Reporting the genotypes of all 43 pedigreed individuals to genotypes file
-> Creating the reads from the genotypes
--> Using a Poisson(5) distribution
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$ ls
fathers.aa  pedigree  shallow.gen        shallow.ind
mothers.aa  reads.aa  shallow.genotypes  shallow.mrk
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$
```

Figure 6: Running the simulation script. Note that although 43 individuals were present in the **pedigree** file, only data for 41 are reported in the results because the grandparent status was set to 'X'. All files needed by **shallowped** (i.e. **shallow.ind**, **shallow.mrk** and **shallow.gen**) and by **AlphaAssign** (i.e. **fathers.aa**, **mothers.aa** and **reads.aa**) have been generated. The file **shallow.genotypes** provides the genotypes of the indi- viduals for information, but will not be used by the programs.

The first and last lines of the file **shallow.ind** are as follows (the file contains 41 lines):

```
1  3 F
2  4 F
3  5 F
4  6 F
5  7 F
- - -
37  39  M
38  40  M
39  41  M
40  42  M
41  43  O
```

As explained above, the first field in each line is the internal code of the individual, the second field is the external name (the one used in the pedigree 5), and the third field corresponds to the status ('F' for potential fathers, 'M' for potential mothers and 'O' for offspring. Note that the individuals with a 'X' status in the initial pedigree have not been reported). The first and last lines of the file **shallow.mrk** are as follows (the file contains 10,000 lines):

```
1  M1
2  M2
3  M3
4  M4
5  M5
- - -
9996   M9996
9997   M9997
9998   M9998
9999   M9999
10000  M10000
```

The first field is the internal number of the marker, and the second is the external name of the marker. In this case, with this simulated dataset, the markers names have been generated automatically. The third file (**shallow.gen**) contains the information on the reads for each marker and each individual. The first and last lines of this file (which contains 410,000 lines...) are:

```
1 1 2 3
1 2 4 0
1 3 0 6
1 4 0 4
1 5 2 5
...
41 9996 0 2
41 9997 0 2
41 9998 5 1
41 9999 3 4
41 10000 4 0
```

Now that all needed files are ready, we can run **shallowped**. Since no **shallow.prm** file has been provided, the program will use the default values for the parameters. The run will obtain the various ditributions needed to test various hypotheses (Figure 7)

```
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$ shallowped
Starting SHALLOWPED...
--> Markers # =  10000
--> Individuals # =     41
---> # fathers =     20
---> # mothers =     20
---> # offspring =      1
--> # of reads:  2049593
--> Covers range from   4.954 to   5.045
--> P[d|g] computed
--> Markers frequencies computed using counts
===> O = 43          F = 22          M = 23          (lik =       -0.00*     -0.00)
--> Best families computed
--> Simulation for offspring    1/    1
Leaving SHALLOWPED...
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$ ls
fathers.aa   reads.aa     shallow.gen        shallow.lik      shallow.ped
mothers.aa   shallow.fam  shallow.genotypes  shallow.lik.hyp  shallow.top
pedigree     shallow.frq  shallow.ind        shallow.mrk
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$ _
```

Figure 7: Output of the run. Note that the program obtained the correct family for individual 43.

As in the previous run, new files have been generated. The file **shallow.lik** contains the likelihoods for all tested parental combinations for each offspring (see Figure 8). The file **shallow.lik.hyp** contains the likelihood distributions under the tested null hypotheses (see Figure 9).

```
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$ head -5 shallow.lik
 43  3 23 -2454.26369407466700068 -228.137313330275515455
 43  3 24 -2572.88917262442555511 -2443.67886667176207993
 43  3 25 -2662.01136842680716654 -2527.82791587855172111
 43  3 26 -2550.13959086928662146 -2520.41982631046994356
 43  3 27 -2557.01302540925826179 -2489.45036333009193186
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$ tail -5 shallow.lik
 43  2 38 -220.917755975692330264 -2444.75838394338643411
 43  2 39 -277.810563427364088124 -2489.36832901454909006
 43  2 40 -245.549373980524665489 -2375.61092398751861765
 43  2 41 -242.101250448211330266 -2578.13303968244508724
 43  2 42 -243.919963759885456603 -2331.55500516080792295
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$ _
```

Figure 8: First and last lines of the file **shallow.lik**: each line contains the likelihood value computed for one couple of potential parents.

```
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$ head -5 shallow.lik.hyp
   1 FAT    1      -0.000
   1 FAT    2      -0.000
   1 FAT    3      -0.000
   1 FAT    4      -0.000
   1 FAT    5      -0.000
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$ tail -5 shallow.lik.hyp
   1 UNM  196  -2477.520
   1 UNM  197  -2374.251
   1 UNM  198  -2328.380
   1 UNM  199  -2419.735
   1 UNM  200  -2364.267
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$
```

Figure 9: First and last lines of the file **shallow.lik.hyp**: each line shows a likelihood value computed using the distribution generated under the tested null hypotheses. In this example, the first 5 lines are the 5 first likelihood values obtained using the 'FAT' null hypothesis, and the last 5 lines are the last 5 likelihood values obtaiend using the 'UNF" null hypothesis.

The file **shallow.fam** contains the best trio with the corresponding likelihood and the p-values for the various tested null hypotheses (Figure 10)

```
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$ cat shallow.fam
Offspring  Best_m     Best_f       log(L_m)   log(L_f)   FAT    PGF    FSF    HSF    UNF
 MOT    MGM    FSM    HSM    UNM
43         22         23                -0.00      -0.00 1.0000 0.0000 0.0000 0.0000 0.0000
1.0000 0.0000 0.0000 0.0000 0.0000
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$
```

Figure 10: The file **shallow.fam contains** the results of the run: the best parents, with various p-values related to the tested null hypotheses. In this example, the only hypotheses that are not rejected are FAT and MOT. In other words, the results are compatible with the facts that the actual father of 43 is 22 and the actual mother of 43 is 23, which matches the simulated pedigree.

In more details, the first 3 fields show the offspring (43) the best parents (father 22, mother 23) (check Figure 5). The next fields are the likelihoods corresponding to that trio (the first is the likelihood for the father assuming that the mother is correct, and the second the likelihood for the mother assuming that te father is correct), and thus the "best" likelihoods. The next 10 fields are p-values, corresponding to 10 distinct null hypotheses:

1. The first p-value corresponds to the null hypothesis "Assuming that the best female is the mother of the tested offspring, the best male is the father of the tested offspring" (FAT). The p-value is equal to 1 because the posterior probability of that situation is $1 - \epsilon$, where $\epsilon$ is lower than the computer precision... This hypothesis is thus accepted.
2. The next 4 p-values correspond to the null hypotheses "Assuming that the best fe- male is the mother of the tested offspring, the best male is the paternal grandfather (PGF) of the tested offspring, a full-sib (FSF) or a half-sib (HSF) of the father, or unrelated to the father (UNF)". The p-values are all equal to 0: no likelihood value obtained under these null hypotheses exceeded the best likelihood. These hypotheses are thus rejected.
3. The next p-value corresponds to the null hypothesis "Assuming that the best male is the father of the tested offspring, the best female is the mother of the tested offspring" (MOT). The p-value is equal to 1 because the posterior probability of that situation is $1 - \epsilon$, where $\epsilon$ is lower than the computer precision... This hypothesis is thus accepted.
4. The last 4 p-values correspond to the null hypotheses "Assuming that the best male is the father of the tested offspring, the best female is the maternal grandmother of the tested offspring (MGM), a full-sib (FSM) or a half-sib (HSM) of te mother, or unrelated to the mother (UNM)". The p-values are all equal to 0: no likelihood value obtained under these null hypotheses exceeded the best likelihood. These hypotheses are thus rejected.

The Figure 11 shows another view of these results.

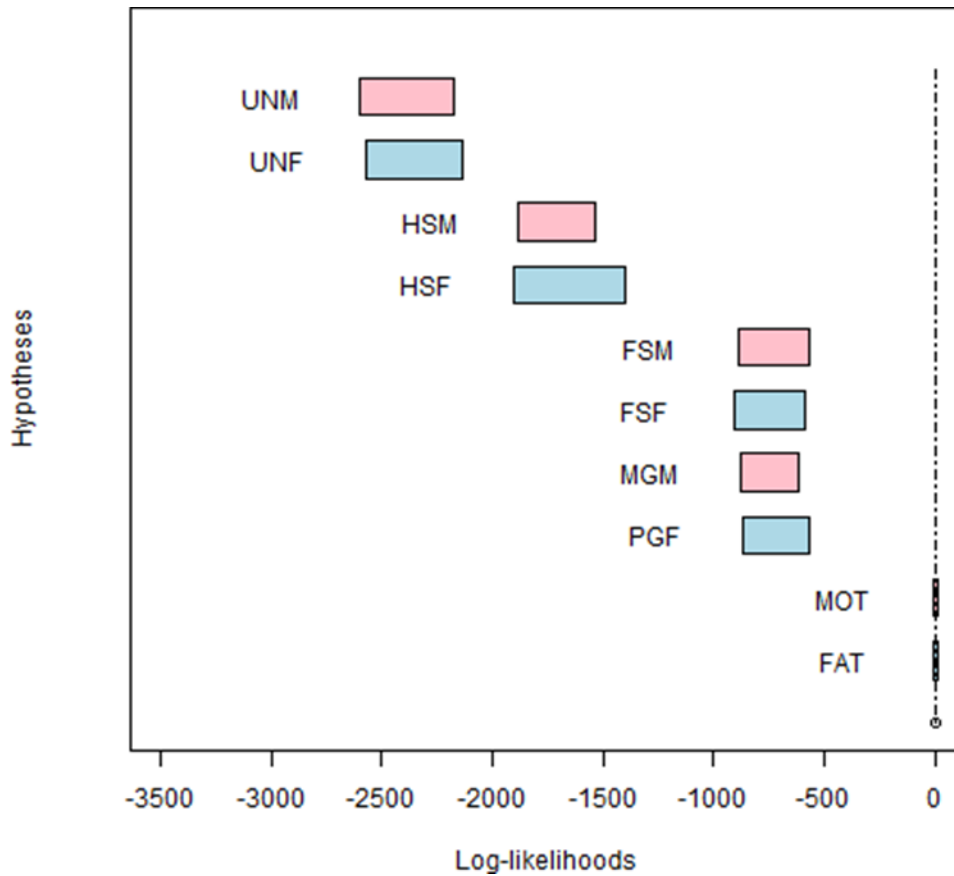


Figure 11: Schematic representation of the ranges of the empirical distributions corre- sponding to the tested null hypotheses (see the definitions in the text). The dashed vertical line corresponds to the best likelihood value: this result suggests that we have obtained the real parents

These results support the hypotheses that the identified male (22) and female (23) are the parents of the tested offspring (43). As an additional example, we can simulate a more difficult situation: we use the same pedigree as in Figure 5, but we discard the real father (individual 22) from the analysis. Consequently, we have a full sib of the real father, but not the real father in the potential male parents. To simulate this situation, we only need to change one line in the pedigree file. The line:

**22 1 2 F**

Becomes:

**22 1 2 X**

Running mksimul.pl -alpha y followed by shallowped leads to results shown in Fig- ures 12 and 13.

```
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$ shallowped
Starting SHALLOWPED...
--> Markers # =  10000
--> Individuals # =     40
---> # fathers =     19
---> # mothers =     20
---> # offspring =      1
--> # of reads:  1999890
--> Covers range from   4.933 to   5.043
--> P[d|g] computed
--> Markers frequencies computed using counts
===> O = 43         F = 21         M = 23         (lik =    -850.91*     -0.00)
--> Best families computed
--> Simulation for offspring    1/   1
Leaving SHALLOWPED...
fred@fmvpa-stat51:~/PANGAGEN/SHALLOWPED/DATA/MANUAL1$
```

Figure 12: Screen output of the run of shallowped on simulated data including a full-sib of the father, but not the father. Note that the program identifies the full sib of the father (21) as the best male parent. The best female parent is still the mother.

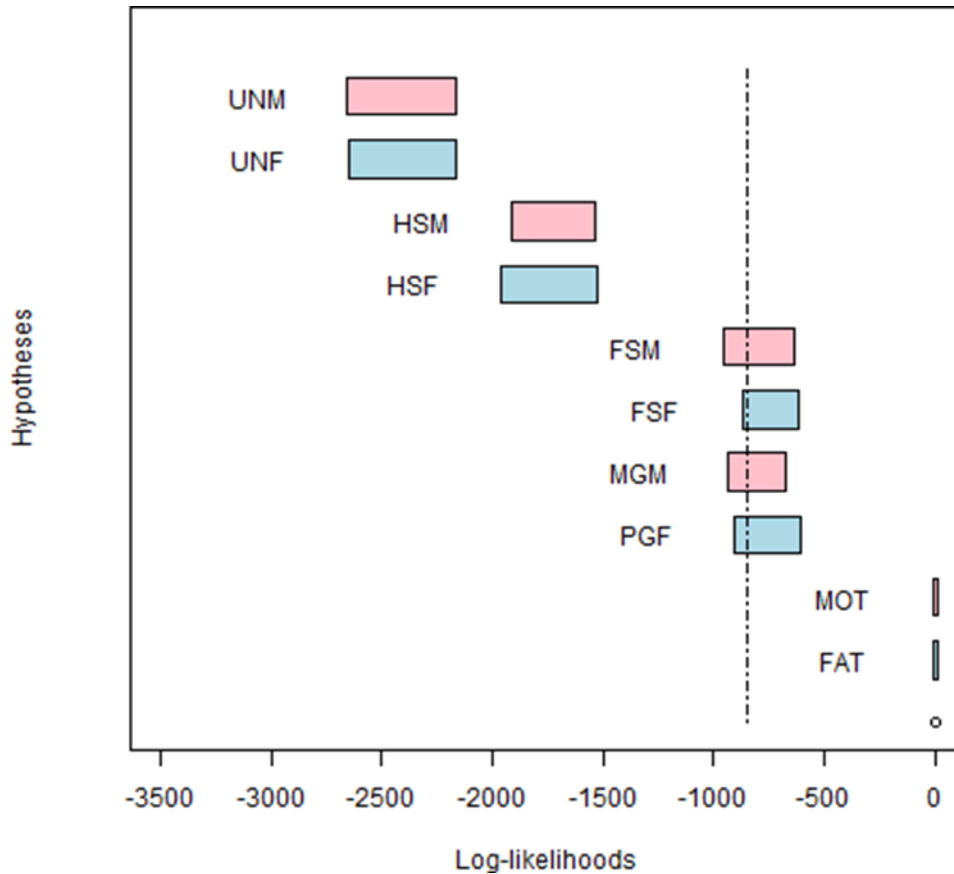


Figure 13: Schematic representation of the ranges of the empirical distributions corre- sponding to the tested null hypotheses (see the definitions in the text). The dashed vertical line corresponds to the best likelihood value. The interpretation of the results presented in Figure 13 is as follows:

1. The best result is less likely than those obtained when both parents have been identified (hypotheses FAT and MOT).
2. For the "best" male (blue rectangles in Figure 13), the best likelihood is compatible with two possibilities:
   (a) assuming that the "best" female is the mother, the "best" male is the paternal grandfather of the offspring (PGF) or,
   (b) assuming that the "best" female is the mother, the "best" male is a full-sib of the father (FFS) (i.e. the correct situation)

3.For the "best" female (red rectangles in Figure 13), the best likelihood is also compatible with two situations:

(a) assuming that the “best” male is the father, the “best” female is the maternal grandmother (MGM), or

(b) assuming that the “best” male is the father, the “best” female is a full-sib of the mother (FSM).

Since the assumption on the male parent is incorrect, the inference on the female is not correct. This result suggests testing additional hypotheses (assuming that the father is a full-sib of the father to derive the distribution for the mother, for example). In comparison, runs of AlphaAssign report the correct parents for the first example, and only the mother for the second.

# 6 Appendix

## 6.1 **mksimul.pl** script

The following perl script can be used to generate simulated data. The programs use various parameters (for which defaults exist) to create the desired data files for **shal- lowped** (and, eventually, for **AlphaAssign**). Limited help is available by typing *“perl mksimul.pl -h”*. Examples are also provided in this manual.

---

```
#---------------------------- # Parameters
#----------------------------
$pedig=$pedig_d="pedigree";
$nbm=$nbm_d=10000;
$mean=$mean_d=5;
$seqe=$seqe_d=0.01;
$alpha=$alpha_d="N";
$shallow=$shallow_d="Y"
; while ($option=shift)
{
 $o=uc($option);
 if ($o eq'-NBM') {
  $nbm=shift;
 }
 elsif ($o eq'-MEAN') {
  $mean=shift;
 }
 elsif ($o eq'-SEQE') {
  $seqe=shift;
 }
 elsif ($o eq'-PED') {
  $pedig=shift;
 }
 elsif ($o eq'-ALPHA') {
  $alpha=shift;
  $alpha=uc($alpha);
```

```
}
elsif ($o eq'-SHALLOW') {
 $shallow=shift;
 $shallow=uc($shallow);
}
elsif ($o eq'-H') {
 print "Usage: [perl] mksimul.pl
 [options]\n";print " Possible
 options:\n";
 print "-NBM nm (number of markers; default: $nbm_d) \n";
 print "-MEAN m (Poisson mean for reads; default: $mean_d)
 \n"; print "-PED  pedig (real  pedigree; default:
 $pedig_d)\n"; print "      Format: Indiv Father Mother Status
 [Cover]\n";
 print "             Status is O|F|M|X (X are not kept in the results)
 \n"; print "        Cover is optional. When used, it supersedes MEAN\n";
 print " -ALPHA Y|N (report AlphaAssign files; default: $alpha_d)
 \n"; print " -SHALLOW Y|N (report Shallowped files; default:
 $shallow_d) \n";print " -SEQE se (sequencing error rate; default:
 $seqe_d) \n";
 print " -H (this
 help)\n"; die "\n";
}
else {
 die "Error: invalid option $option\n";
};
}; #___# Read pedigree
#-----------------------------------------
print "-> Reading the pedigree\n";
open(PED,"$pedig") || die "Error: pedigree $pedig cannot be found\n";
@ped= ();
%ped= ();
$ni=$ns=$nd=$no=$nind=
0;while (<PED>) {
 $ni
 ++;
 cho
 p;
 s/^
 \s+
 //;
 @x=split(/\s+/);
 $ped[$ni] [1] =$x [0];
 $ped[$ni] [2] =$x [1];
 $ped[$ni] [3] =$x [2];
 $ped[$ni] [4] =$x [3]; # Status (X: remove, F: sire, M: dam, O: offspring)
 $ns++ if ($x [3] eq 'F');
 $nd++ if ($x [3] eq 'M');
 $no++ if ($x [3] eq 'O');
 $ped[$ni] [5] =0;    # Genotyped (1) or not yet (0)
 $nind++ unless ($x [3] eq 'X');
 $ped[$ni] [6] = ($x [3] eq 'X')?0: $nind; # Code for shallowped
```

```
 $ped[$ni] [7] = ($x [4] =="")? $mean: $x [4];
 $ped {$x [0]} =$ni;
};
print "--> $ni individuals in the pedigree\n";
print "---> $ns fathers, $nd mothers, $no
offspring\n";#
if (uc($alpha) ne 'N') {
 print "--> Reporting the needed files for AlphaAssign\n";
 open(SIRE,">fathers. aa");
 open(DAM,">mothers.
 aa"); for ($i=1;
 $i<=$ni; $i++) {
  next unless ($ped[$i] [4] eq 'O');
  print SIRE "$ped[$i]
  [1]"; for ($j=1;
  $j<=$ni; $j++) {print
  SIRE " $ped[$j][1]"
  if ($ped[$j][4] eq
  'F');
   }; print SIRE "\n";
  print DAM "$ped[$i]
  [1]"; for ($j=1;
  $j<=$ni; $j++) {
   print DAM "$ped[$j] [1]" if ($ped[$j] [4] eq 'M');
  };
  print DAM "\n";
 };
 close(DAM
 );
 close(SIRE
 );
};
if (uc($shallow) ne 'N') {
 print "--> Reporting the needed file for
 shallowped\n";open(IND,">shallow.ind");
 $nind=0;
 for ($i=1; $i<=$ni; $i++) {
  next if ($ped[$i] [4] eq 'X');
  $nind++;
  print IND "$nind $ped[$i] [1] $ped[$i] [4] \n";
 };
 close(IND);
};
#-----------------------------------------
# Markers file (for
shallowped)#----------------
if (uc($shallow) eq 'Y') {
 print "-> Creating markers file for
 shallowped\n";open(MRK,">shallow.mrk");
 for ($m=1; $m<=$nbm;
  $m++) {printMRK "$m
  M$m\n";
```

```
 };
 close(MRK);
 print "--> Done! \n";
};
#_________________________________ # Genotypes
#_________________________________
print "-> Creating the
genotypes\n";print "--> Using
$nbm markers\n";
$all_genotyped=0;
while ($all_genotyped==0) {
 $all_genotyped=1;
 for ($i=1; $i<=$ni;
  $i++) {next if
  ($ped[$i] [5] ==3);#
  Check paternal
  haplo
  if ($ped[$i] [5]! =1) {
   if ($ped[$i] [2] ==0) {
    # Generate random alleles# print “DBG Building paternal haplotype of $ped[$i] [1]
    randomly\n"; for ($m=1; $m<=$nbm; $m++) {$geno[$i] [$m] [1] =(rand()<0.5)?1:2;
    };
    $ped[$i] [5] +=1;
   }
   else {
    $is=$ped{$ped[$i][2]};
    if ($ped[$is] [5]! =3) {
     # Sire not completely genotyped
     $all_genotyped=0;
    }
    else {
     # Generate alleles from sire genotype
     # print “DBG Building paternal haplotype of $ped[$i] [1] from
         $ped[$is] [1] \n";
     for ($m=1; $m<=$nbm; $m++) {
      $geno[$i] [$m] [1] =$geno[$is] [$m] [(rand () <0.5)?1:2];
     };
     $ped[$i] [5] +=1;
    };
   };
  };
  # Check maternal
  haploif ($ped[$i]
  [5]<2) {
   if ($ped[$i] [3]==0) {
    # Generate random alleles
    # print “DBG Building maternal haplotype of $ped[$i] [1] randomly\n";
    for ($m=1; $m<=$nbm; $m++) {
     $geno[$i][$m][2]=(rand()<0.5)?1:2;
    };
    $ped[$i][5]+=2;
   }
```

```
   else {
    $id=$ped{$ped[$i] [3]};
    if ($ped[$id] [5]! =3) {
     # Dam not completely genotyped
     $all_genotyped=0;
    }
    else {
     # Generate alleles  from dam genotype
     #  print “DBG Building maternal haplotype of $ped[$i] [1] from
          $ped[$id] [1] \n";
     for ($m=1; $m<=$nbm; $m++) {
       $geno[$i] [$m] [2] =$geno[$id] [$m] [(rand () <0.5)?1:2];
      };
     $ped[$i] [5] +=2;
    };
   };
  };
  $all_genotyped=0 unless ($ped[$i] [5] ==3);
 };
};
print "--> Reporting the genotypes of all $ni pedigreed individuals
    togenotypes file\n";
open(GEN,">shallow.
genotypes"); for ( $i=1;
$i<=$ni; $i++) {
 print GEN “$ped[$i]
 [1]"; for ($m=1;
 $m<=$nbm; $m++) {
  print GEN “$geno[$i] [$m] [1] “.”$geno[$i] [$m] [2]";
 };
 print GEN "\n";
};
close(GEN);
#___________________________________ # Reads
#___________________________________
print "-> Creating the reads from the
genotypes\n";print "--> Using a
Poisson($mean) distribution\n";@reads= ();
for ($i=1; $i<=$ni; $i++) {
 next if ($ped[$i] [4] eq
 ’X’); for ($m=1;
 $m<=$nbm; $m++) {
  $g=$geno[$i] [$m] [1] +$geno[$i] [$m] [2];
  $reads[$i] [$m] [1] =$reads[$i] [$m] [2] =0;
  $nr=&rpois($ped[$i
  ] [7]);if ($g==2) {
   $reads[$i] [$m] [1] =&rbinom($nr,1-$seqe);
   $reads[$i] [$m] [2] =$nr-$reads[$i] [$m] [1];
  }
  elsif ($g==4) {
   $reads[$i] [$m] [1] =&rbinom ($nr, $seqe);
   $reads[$i] [$m] [2] =$nr-$reads[$i] [$m] [1];
```

```
  }
  else {
   $reads[$i] [$m] [1] =&rbinom($nr,0.5);
   $reads[$i] [$m] [2] =$nr-$reads[$i] [$m] [1];
  };
 };
};
if (uc($alpha) eq 'Y') {
 open(SEQ,">reads.
 aa"); for ($i=1;
 $i<=$ni; $i++) {
  next if ($ped[$i] [4] eq 'X');
  print SEQ "$ped[$i]
  [1]"; for ($m=1;
  $m<=$nbm; $m++) {
   print SEQ "$reads[$i] [$m] [1]";
  };
  print SEQ "\n";
  print SEQ "$ped[$i]
  [1]"; for ($m=1;
  $m<=$nbm; $m++) {
   print SEQ "$reads[$i] [$m] [2]";
  };
  print SEQ "\n";};
 close(SEQ);
};
if  (uc($shallow)  eq
 'Y')
 {open(SEQ,">shallo
 w.gen");  for  ($i=1;
 $i<=$ni; $i++)  {
  next if ($ped[$i] [4] eq
  'X'); for ($m=1;
  $m<=$nbm; $m++) {
   print SEQ "$ped[$i] [6] $m $reads[$i] [$m] [1] $reads[$i] [$m] [2] \n";
  };
 };
 close(SEQ);
};

exit;

sub rpois {
 my ($m)  = @_;
 my ($k, $pk, $r, $pc);
 $k=0;
 $pk=exp(-$m);
 $r=rand ();
 $pc=$pk;
 while ($r>$pc) {
  $pk=$m*$pk/($k+1);
  $k++;
```

```
  $pc+=$pk;
 };
 $k;
};
sub rbinom {
 my ($n, $p)  = @_;
 my ( $r,$d,$pr);
 $r=rand () ;
 $d=0 ;
 $pr=(1-$p) **$n ;
 $pc=$pr;
 while ($r>$pc) {
  $pr*=$p/(1-$p) *($n-$d) /($d+1) ;
  $d++;
  $pc+=$pr;
 };
 $d;
};
```

## 2. Supplementary material for Chapter 4

The husbandry of striped catfish in the various stages of selection program

### 2.1 Production of the first generation (G1)

#### 2.1.1 **The base population (G0) selection**

Broodstocks of catfish have been selected from 3 different sources (An Giang, Vinh Long and Can Tho province, with 10 males and 10 females in each province), with average weights ranging from 5 to 7 kg. Principally, the selected broodstocks had to be healthy, without visible injury or abnormal signs. The females had to have a big belly, thin abdominal skin, swollen, reddish genitals and the males had to release milt on application of a gentle pressure on the abdomen. Due to the difficulty to evaluate externally the level of ovary maturation, an intra-ovarian biopsy with a flexible catheter has been performed to check the oocyte development. The well-matured females should have an oocyte diameter of 1.0-1.1 mm.

A piece of fin (~1 $cm^2$) from each fish (i.e. for a total of 60 broodstocks) has been collected and preserved in 95% ethanol for genetic analyses.

#### 2.1.2 Induced breeding and crossing.

Spawning has been induced by injection with human chorionic gonadotrophin (HCG). HCG was injected in females at doses of 500 - 5000 UI/kg at a time and injected up to four times before the fish were finally induced. The males received only one injection of 1000 UI/kg at the last time of injection to the females (Table 1).

Table 1 Hormone (HCG) dose and timing of hormone injection used to induced spawning striped catfish

| Female | | | Male | | |
|---|---|---|---|---|---|
| Injection No. | Time (hr) | Dose (UI/kg) | Injection No. | Time (hr) | Dose (UI/kg) |
| 1st | 0 | 500 | | | |
| 2nd | 24 | 500 | | | |
| 3rd | 48 | 1000 | 1st | 48 | 1000 |
| 4th | 56 | 3000 | | | |
| Combined | | 5000 | | | |

Ovulation occurred 8 hours after the last injection. Eggs were collected as soon as possible by stripping. Milt was collected by pressing gently on the abdomen of the fish and using a syringe, and then the collected milts were stored in an immobilization solution (10 g Tris-HCL in 1 L of 9ppt water). The milt was diluted five times in this solution and stored at 4-5 $^{O}$C before using. A dry fertilization process was used, where eggs and milt were mixed gently. 29 females (~ 3000 eggs of each female) were crossed with 30 males to create 870 families (one female of Can Tho province could not be used due to low quality of her eggs). Fertilization solution (3 g urea and 4 g salt in 1 L of water) was added to the mixture of eggs and milt to trigger fertilization after 4 min. The fertilized eggs were then transferred into boxes (one box per family) for incubation. The fertilized eggs started to hatch 24 hours after fertilization. Fertilization rates varied from 83-90 % and the hatching rate was 61-73 % (Fig. 1).

Figure1 Incubating 300 families in 300 boxes (10 males * 30 females). This procedure has been repeated in the 3 selected hatcheries, leading to 900 boxes (i.e. full-sib families).

### 2.1.3. Seedstock rearing

The larvae were transferred to rearing earthen ponds within 15 hours after hatching and fed live feed. From each family, 2000 good quality larvae (with no abnormal signs, uniform size, swimming actively and responding to external stimuli quickly) were selected for nursing in two earthen ponds (leading to 1.740.000 larvae in total).

The nursing ponds (1.600 $m^2$ for each pond) were prepared a week before stocking to encourage the growth of live food. Preparing the ponds included the removal of the bottom sludge, liming (10 kg/100 $m^2$), killing unwanted organisms by applying saponin products (1 kg for 300 $m^3$) of water and drying the pond bottom for four days. The water supplied into ponds was screened through a fine mesh to prevent the entrance of eggs and larvae of undesirable organisms, until the level of water in pond reached 1.2 m. To encourage the growth of natural food, 10 kg of super benthos (one kind of fertilizer) were added into each pond 24 hours after supplying water. Moreover, beneficial bacterial products were added into ponds at a rate of 400 g/1000 $m^3$ together with 3 kg of live food (Moina). Larvae were transported from hatcheries to nursing ponds in oxygenated bags in the early morning (8:00 am) to avoid direct exposure to sunlight. The larvae were acclimated to pond water by keeping bags in the pond for 20 min before releasing.

Figure 2: Rearing ponds (freshwater) for 47 days

Water quality and larval behavior were checked early every morning. Fry were fed with commercial pelleted feeds (Proconco Company) containing 28-35 % crude protein according to size. The feeding rate varied from 6 – 10 % of body weight with two or three daily feedings. Additionally, nutrient supplements such as Vitamin C and mix of mineral were regularly used during the nursing period.

## 2.2. Process of 3 selection round in RAS system

| Weight (g) | Salinity (ppt) | Dph | Aquaculture system | Operation |
|---|---|---|---|---|
| 5.4 | 0 | 0<br>47 | Rearing in earthen ponds | 30♂ X 30♀ (from 3 provinces)<br>2000 larvae x 870 families (29 ♀)<br>(1,740,000 larvae in total)<br>Survival rate :1.38%<br>21,550 fry |
| 115 | 10 | 148 | Rearing in RAS | 21,000 fry were acclimated to targeted salinity (10ppt)<br>Survival rate :80.8 %<br>16,988 fish |
| 115 | 10 | 148 | Rearing in RAS | **Selection steps** 50% \| 50%><br>**1st selection**<br>Random: 5,618 ind ; 115± 39g<br>Selected (~ 50 % best): 4,132ind; 157±36 g |
| ~ 300 | 10 | 237 | Rearing in RAS | 89 days<br>~60% ~40% \| 50% ~50% ><br>**2nd selection**<br>Random: 2,854 ind ; – 188± 89 g<br>Selected (~ 50% best): 1,426 ind ; 374±84 g |
| ~ 1000 | 10 | 340 | Rearing in RAS | 103 days<br>~60% ~40% \| ~50% ~50% ><br>**3rd selection**<br>Random: 470 ind ; 793±229.9 g<br>Selected (~ 50% best): 576 ind ; 1,380± 174.9g<br>(PIT tag + DNA) (PIT tag + DNA) |
| ~ 1000 | 5 | 340 | Rearing in RAS | 470 fish \| 576 fish<br>Conditioning for breeding |

Figure 3: Procedure of selection experiment (Values were express as mean ± SD)

## 2.3 Production of the G2 generation

### 2.3.1 Selecting parents for production of the G2 generation

We produced a G2 generation using random couples from each of the three G1 groups (i.e. from the selected group (10♂ X 10♀), the random group (10♂ X 10♀), and the freshwater group (8♂ X 8♀). The pedigree information of selected parents was identified by our program Shallowped, only crossing between parents with different families. The process for selecting broodstock, injecting HCG, breeding, and fertilization was presented in Figure 3. We pooled the same number of fertilized eggs per family within each group and incubated them together. Fertilized eggs hatch 22-24 h after fertilization. The results of characteristic of broodstock, egg and larva from three fish groups were presented in Table 2.

Figure 4: Process of selecting broodstock, injecting HCG, breeding and fertilizing

Table 2: Characteristic of broodstock, egg and larva from three fish groups

| Characteristics | Freshwater | Selected | Random |
|---|---|---|---|
| Weight of female (kg) | 4.04±0.42[a] | 5.27±0.88[b] | 4.51±0.57[ab] |
| Productivity (egg/kg of female) | 161,051±65,359[a] | 178,039±35,053[a] | 175.556±40,716[a] |
| Egg diameter (mm) | 1.016±0.025[a] | 1.042±0.016[a] | 1.052±0.029[a] |
| Yolk volume ($mm^3$) | 0.44±0.051[a] | 0.47±0.044[a] | 0.48±0.047[a] |
| Total length of larva (mm) | 3.208±0.044[a] | 3.213 ±0.07[a] | 3.357±0.111[a] |

**** Data was expressed as mean ± standard error. Different in letters in line was significantly different ($p< 0.05$)*

**2.3.2 Rearing larva of three fish groups**

We subsequently reared the three groups in identical freshwater environments under optimal conditions for the development of striped catfish. Larvae hatched after 29–34 h and were transferred to three 2 $m^3$ outdoor composite tanks for nursing. The larvae from each group were nursed separately in composite tanks with the same density of 5 individuals/L. The larvae were fed Rotifer, Artemia nauplii, and Moina in the first 5 days. After that period, a high protein powder and flake diet (40 -35 % protein) was used in combination with Moina until 4–5 weeks of age. The The technical process for rearing larva of striped catfish was presented in Table 3. Subsequently, fries (0.6–0.8 g of body weight) from each group were transferred to 10 m3 outdoor composite tanks, where they were raised until fingerling stage with an average body weight of 20 g.

Table 3: The technical process for rearing larva of striped catfish (*Pangasianodon hypophthalmus*)

From larva to fry (5 ) g , stocking density 5 ind/L, tempature of water 28-30°C, DO ≥ 4ppm, pH 7.5-8.0

| Dph (ngày) | Algae (L/$m^3$) | Super-stock (g/$m^3$) | Moina (ind/mL) | Rotifer (con/mL) | (Probio) (g/$m^3$) | Vimelec (g/$m^3$) | Commercial feed (powder: 40% P, flake: 35%P, size of feed -mm) | | | | | | Siphong | Exchange water (% ) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Powder (g/$m^3$) | Flake (g/$m^3$) | Pellet-0.4 (% body weight) | Pellet-0.6 (% body weight) | Pellet-0.8 (% body weight) | Pellet-1 (% body weight) | | |
| 0 | 40 | 4 | 0.5 | - | - | - | - | - | - | - | - | - | - | - |
| 1 | 20 | 4 | 0.5 | - | - | - | - | - | - | - | - | - | - | - |
| 2 | 20 | 4 | 1 | 3 | - | - | - | - | - | - | - | - | - | - |
| 3 | 40 | 4 | 1 | 4 | - | - | - | - | - | - | - | - | - | - |
| 4 | 40 | 4 | 2 | 5 | 1 | - | - | - | - | - | - | - | - | - |
| 5 | 30 | 4 | 2 | 2 | 1 | - | - | - | - | - | - | - | - | - |
| 6 | - | 4 | 4 | - | 1 | - | - | - | - | - | - | - | - | - |
| 7 | - | - | 4 | - | 1 | - | 5 | - | - | - | - | - | Siphong | 10 |
| 8-10 | - | - | 5-5 | - | 1-1 | - | 10-20 | - | - | - | - | - | Siphong | 15 |
| 11-15 | - | - | 6-3 | - | 1-1 | 1-1 | 25-35 | - | - | - | - | - | Siphong | 20 |
| 16-20 | - | - | 3-2 | - | 1-1 | 1-1 | 30-10 | 10-30 | - | - | - | - | Siphong | 20 |
| 21-25 | - | - | 2-1 | - | 1-1 | 1-1 | 5-0 | 30-40 | - | - | - | - | Siphong | 30 |
| 26-30 | - | - | - | - | 1-1 | 1-1 | - | 30-5 | 2-7 | - | - | - | - | 3 |
| 31-35 | - | - | - | - | 1-1 | 1-1 | - | - | 8-5 | 2-5 | - | - | - | 3 |
| 36-40 | - | - | - | - | - | - | - | - | - | 4-2 | 6-8 | - | - | 3 |
| 41-45 | - | - | - | - | - | - | - | - | - | - | 2-1 | 8-9 | - | 3 |

1 *Vimelec (Commercial product of Vemidem company): contain mineral and complex Vitamin*

### 2.4 Results on parentage assigment of striped catfish

| # offspring | MAG1 | MAG10 | MAG2 | MAG3 | MAG4 | MAG5 | MAG6 | MAG7 | MAG8 | MAG9 | MCT1 | MCT10 | MCT2 | MCT3 | MCT4 | MCT5 | MCT6 | MCT7 | MCT8 | MCT9 | MVL1 | MVL10 | MVL2 | MVL3 | MVL4 | MVL5 | MVL6 | MVL7 | MVL8 | MVL9 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FAG1 | 4 | 4 | | | | 2 | 2 | | 2 | | 7 | 1 | 1 | 8 | 1 | 7 | | 1 | 1 | 2 | 2 | 1 | 4 | 9 | | 4 | | 7 | 1 | 5 | 76 |
| FAG10 | 1 | 1 | | | | 1 | 2 | | 1 | | 5 | | 1 | 1 | 1 | 5 | | | 1 | | | 2 | 3 | 1 | | 5 | | 2 | | 2 | 35 |
| FAG3 | | 6 | | | 5 | 1 | 12 | 1 | | 1 | 2 | 6 | 1 | 3 | 2 | 4 | 4 | 2 | 2 | | | 16 | | 1 | | 1 | | 7 | 2 | 2 | 81 |
| FAG4 | 1 | 1 | | | 1 | 1 | 3 | | 2 | 2 | 2 | 3 | | 4 | 1 | 1 | | | 2 | 1 | 1 | 1 | 3 | 2 | 1 | | | 2 | 1 | | 36 |
| FAG5 | 1 | 1 | 1 | | | 1 | 1 | | | | | 3 | 1 | 7 | | 9 | 1 | | 2 | 3 | | | 1 | | 3 | 1 | | | | 7 | 43 |
| FAG6 | | 2 | | 1 | | | 2 | | | | | | 2 | 4 | | 8 | 2 | 1 | | | 1 | 2 | | | | | | | | 1 | 26 |
| FAG7 | | 2 | | | | | 2 | | | | | 1 | | | | 2 | | 1 | 1 | 1 | | | | 6 | | | | | | | 16 |
| FAG8 | | 12 | | | 1 | | 1 | | | 3 | 2 | 3 | 2 | 9 | 2 | 20 | | | 9 | 2 | | 6 | 1 | 7 | 1 | 3 | | 3 | 2 | | 89 |
| FAG9 | 1 | 3 | 1 | | 1 | | 1 | | 3 | | 2 | | 4 | | | 1 | | | 2 | 2 | 1 | | 2 | 1 | | | | 1 | 1 | 1 | 28 |
| FCT1 | 1 | | | | | | | | | | | | | 1 | | | | 1 | | | 1 | | | | | | 1 | 1 | | | 6 |
| FCT10 | | | | | 1 | | 1 | | | | | | | | | 2 | 1 | | | 1 | 1 | | | | | 2 | | 1 | | | 10 |
| FCT3 | | 4 | | | 1 | | 1 | | | 1 | | | | | 1 | | 1 | | | 1 | | | | | | | 1 | 1 | | | 12 |
| FCT4 | 1 | | | | | 1 | | | | | | | | | | | | | | | | | | | | | | | | 1 | 3 |
| FCT6 | | | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | 1 |
| FCT7 | | | | | | | | | | | 1 | | | | | | | | | | | | | | | | | | | | 1 |
| FCT8 | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | | | | | 1 |
| FVL1 | | 2 | | | | 1 | | | 1 | | | | | | | | | | | | | | | | | | | | | | 4 |
| FVL3 | | | | | | | | | | | | | | | | | | | | | | | | 1 | | | | | | | 1 |
| FVL6 | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | | | 1 |
| Total | 10 | 38 | 2 | 1 | 10 | 8 | 28 | 1 | 9 | 7 | 21 | 17 | 12 | 37 | 9 | 59 | 10 | 6 | 20 | 13 | 7 | 28 | 14 | 28 | 6 | 16 | 2 | 25 | 7 | 19 | 470 |

Figure 5: The example results of parentage assignment of striped catfish in selection program. The lines and columns headers correspond to the male (M) and female (F) breeders, respectively, and include the origin of the fish (AG: fish from An Giang province; CT: fish from Can Tho province; VL: fish from Vinh Long province).

2.5. **Summary results from water quality management of all experiments**

Table 4: Water quality parameters for all experiments from each developmental stage of striped catfish

| Index | DO (mg $L^{-1}$) | pH | Temperature ($^0$C) | TAN(mg $L^{-1}$) | N-$NO_2^-$(mg $L^{-1}$) |
|---|---|---|---|---|---|
| Embryonic stage | 5.1 to 5.6 | 8.2 to 8.6 | 26.3 to 28.4 | NA | NA |
| Larval to fry stages | 4.8 to 5.5 | 7.2 to 8.1 | 27.2 to 28.7 | NA | NA |
| Fry to fingerling stage | 5.0 to 6.5 | 7.1 to 8.0 | 27.1 to 30.5 | < 0.5 | < 0.3 |
| Adult stage | 27.3 to 30.8 | 7.1 to 8.5 | 3.9 to 5.2 | < 1.5 | < 0.8 |

Dissolved oxygen (DO), pH, temperature was measured daily a multiple-parameter water quality meter (DKK-TOA, WQC-24, Japan). Total ammonia nitrogen (TAN) and N-$NO_2^-$ were measured weekly using a Multiparameter Photometer (Hana Hi 83300). NA: not available.

**3. Supplementary materials of Chapter 5.**

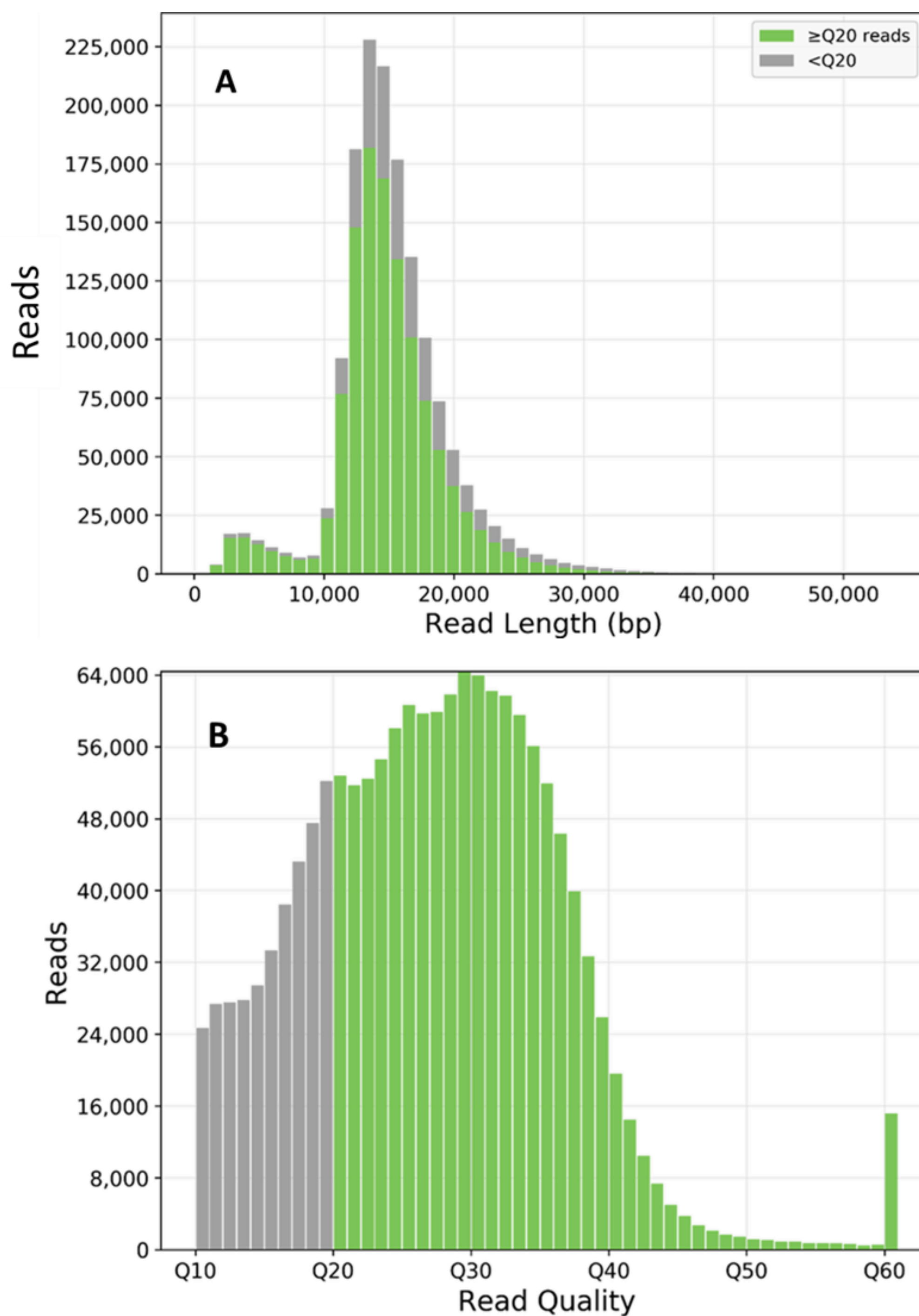


Figure S1: Distribution of read length (A) and read quality (B). Used HiFi sequences (Q ≥ 20) are shown in green.

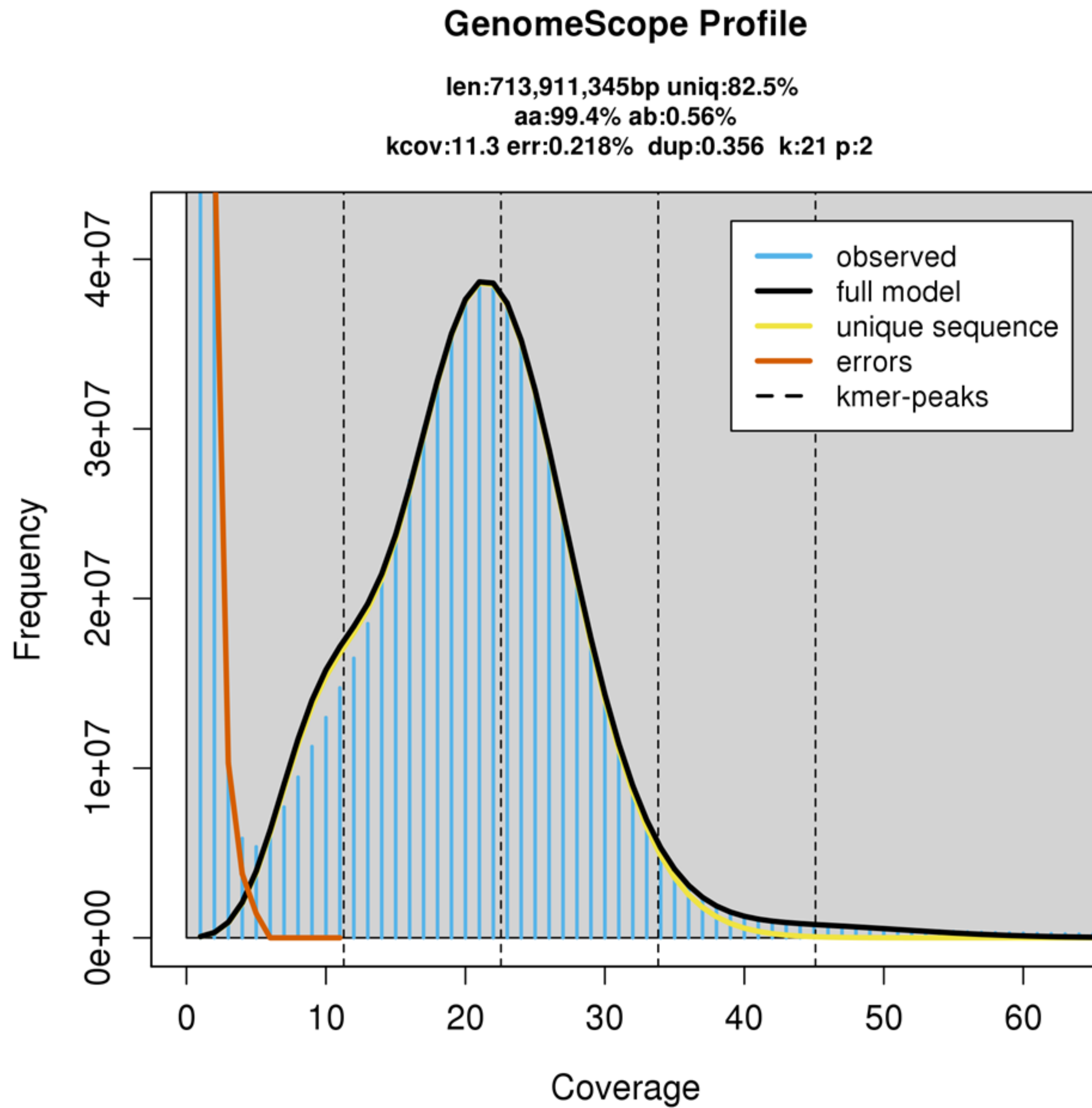


Figure S2: A *k-mer* analysis of the genome sequencing reads for the striped catfish using GenomeScope

Table S1 Repetitive elements in the striped catfish genome

| TE Class | Family | Sub-family | De novo | | | Homology | | |
|---|---|---|---|---|---|---|---|---|
| | | | Count | Length (bp) | Percentage | Count | Length (bp) | Percentage |
| Class I (Retroelements) | Non-LTR elements | SINEs: | 20,832 | 31,13,943 | 0.39 | 50584 | 6755978 | 0.86 |
| | | Penelope | 1,757 | 615,424 | 0.08 | 6402 | 1467082 | 0.19 |
| | | LINEs: | 98,550 | 30,509,760 | 3.87 | 121,145 | 29,469,338 | 3.74 |
| | | CRE/SLACS | 0 | 0 | 0.00 | 0 | 0 | 0.00 |
| | | L2/CR1/Rex | 76,961 | 24,726,066 | 3.14 | 77,327 | 20,030,762 | 2.54 |
| | | R1/LOA/Jockey | 1,567 | 82,3571 | 0.10 | 1,918 | 412,240 | 0.05 |
| | | R2/R4/NeSL | 1,219 | 376,738 | 0.05 | 649 | 270,429 | 0.03 |
| | | RTE/Bov-B | 8,119 | 1,813,291 | 0.23 | 19,762 | 3,853,265 | 0.49 |
| | | L1/CIN4 | 3,687 | 935,167 | 0.12 | 3,221 | 1,146,312 | 0.15 |
| | LTR elements | BEL/Pao | 2,006 | 781,888 | 0.10 | 1,030 | 399,154 | 0.05 |
| | | Ty1/Copia | 0 | 0 | 0.0 | 1,642 | 513,400 | 0.07 |
| | | Gypsy/DIRS1 | 42,978 | 14,309,769 | 1.82 | 45,254 | 15,858,477 | 2.01 |
| | | Retroviral | 2,639 | 924,326 | 0.12 | 6,469 | 1,185,074 | 0.15 |
| | | Total LTR elements: | 87,491 | 23,554,747 | 2.99 | 79,867 | 21,805,013 | 2.77 |
| Total Retroelements (a): | | | 206,873 | 57,178,450 | 7.25 | 251,596 | 58,030,329 | 7.36 |
| Class II (DNA transposons) | | hobo-Activator | 99,579 | 17,622,531 | 2.24 | 122,119 | 19,242,281 | 2.44 |
| | | Tc1-IS630-Pogo | 193,506 | 53,485,756 | 6.78 | 230,862 | 59,455,614 | 7.54 |
| | | En-Spm | 0 | 0 | 0.00 | 0 | 0 | 0.00 |
| | | MuDR-IS905 | 0 | 0 | 0.00 | 0 | 0 | 0.00 |
| | | PiggyBac | 2,069 | 67,2161 | 0.09 | 4,871 | 838,840 | 0.11 |
| | | Tourist/Harbinger | 20,401 | 4,866,723 | 0.62 | 29,947 | 5,121,870 | 0.65 |
| | | Other (Mirage, P-element, Transib) | 1,608 | 321,938 | 0.04 | 1,660 | 460,487 | 0.06 |
| Total DNA transposons (b): | | | 442,919 | 101,790,693 | 12.91 | 576,633 | 115,072,113 | 14.60 |
| Rolling-circles (c) | | | 11,569 | 2,998,325 | 0.38 | 18437 | 3986467 | 0.51 |
| Unclassified (d): | | | 694,670 | 112,360,030 | 14.25 | 680699 | 125,718,307 | 15.95 |
| Total interspersed repeats (a+b+c+d) | | | | 271,329,173 | 34.42 | | 298,820,749 | 37.90 |
| Small RNA: | | | 3,756 | 2,166,067 | 0.27 | 5,071 | 1,558,812 | 0.20 |
| Satellites: | | | 1,817 | 497,961 | 0.06 | 11,591 | 2,308,540 | 0.29 |
| Simple repeats: | | | 573,469 | 28,230,949 | 3.58 | 520,916 | 22,917,129 | 2.91 |
| Low complexity: | | | 45,375 | 3,133,402 | 0.40 | 40,587 | 2,517,469 | 0.32 |
| Total bases masked: | | | | 308,355,877 | 39.11 | | 331,921,288 | 42.10 |

Table S2 Overview of simple sequence repeats (SSRs)

| Cutoffs and Range | | | | |
|---|---|---|---|---|
| SSR Type | Set of repeating bases | Repetition number for the set | Example | Total Length |
| Mono nucleotide Repeats (p1) | 1 | >= 10 bases | AAAAAAAAAAAA | >=10 t0 Any length |
| Di nucleotide Repeats (p2) | 2 | >= 6 Pairs | CACACACACACACA | >=12 to Any length |
| Tri nucleotide Repeats (p3) | 3 | >= 5 Sets | AACAACAACAACAAC | >=15 to Any length |
| Tetra nucleotide Repeats (p4) | 4 | >= 5 Sets | AAACAAACAAACAAACAAAC | >=20 to Any length |
| Penta nucleotide Repeats (p5) | 5 | >= 5 Sets | AAATGAAATGAAATGAAATGAAATG | >=25 to Any length |
| Hexa nucleotide Repeats (p6) | 6 | >= 5 Sets | AAATATAAATATAAATATAAATATAAATAT | >=30 to Any length |

| Cutoffs and Range<br>Statistics | Count |
|---|---|
| Total number of sequences examined | 381 |
| Total size of examined sequences (bp) | 788,355,903 |
| Total number of identified SSRs | 960,574 |
| Number of SSR containing sequences | 377 |
| Number of sequences containing more than 1 SSR | 376 |
| Number of SSRs present in compound formation | 320,010 |
| p1 | 315,319 |
| p2 | 510,826 |
| p3 | 73,996 |
| p4 | 53,670 |
| p5 | 4,726 |
| p6 | 2,037 |

**Link for assess to a new reference genome of striped catfish on NCBI** :
https://www.ncbi.nlm.nih.gov/assembly/GCA_024500225.1#/def