SPECIES-SPECIFIC DNA MARKERS FOR IMPROVING THE GENETIC MANAGEMENT OF TILAPIA

A thesis submitted for the degree
of Doctor of Philosophy in
Aquaculture

by

Mochamad Syaifudin

September 2015

INSTITUTE OF AQUACULTURE

UNIVERSITY OF
STIRLING

2

DECLARATION

I hereby declare that the work and the results presented in this thesis have been carried out by myself at the Institute of Aquaculture, University of Stirling, Scotland and have not been submitted for any other degree of qualification. All information from other sources has been acknowledged.

Name of Candidate : Mochamad Syaifudin

Signature :

Principal Supervisor : Prof. Brendan McAndrew

Signature :

Co-Supervisor : Dr. David Penman

Signature :

Date : 21st December 2015

ABSTRACT

The tilapias are a group of African and Middle Eastern cichlid fish that are widely cultured in developed and developing countries. With many different species and sub-species, and extensive use of interspecies hybrids, identification of tilapia species is of importance in aquaculture and in wild populations where introductions occur.

This research set out to distinguish between tilapia species and sub-species by retrieving species-specific nuclear DNA markers (SNPs) using two approaches: (i) sequencing of the coding regions of the *ADA* gene; and (ii) next-generation sequencing, both standard RADseq and double-digest RADseq (ddRADseq). The mitochondrial DNA (mtDNA) marker cytochrome c oxidase subunit I (COI) was used to verify tilapia species status.

ADA gene sequence analysis was partially successful, generating SNP markers that distinguished some species pairs. Most species could also be discriminated using the COI sequence. Reference based analysis (RBA: using only markers found in the *O. niloticus* genome sequence) of standard RADseq data identified 1,613 SNPs in 1,002 shared RAD loci among seven species. *De novo* based analysis (DBA: based on the entire data set) identified 1,358 SNPs in 825 loci and RBA detected 938 SNPs in 571 shared RAD loci from ddRADseq among 10 species. Phylogenetic trees based on shared SNP markers indicated similar patterns to most prior phylogenies based on other characteristics. The standard RADseq detected 677 species-specific SNP markers from the entire data set (seven species), while the ddRADseq retrieved 38 (among ten species). Furthermore, 37 such SNP markers were identified from ddRADseq data from a subset of four economically important species which are often involved in hybridization in aquaculture, and larger numbers of SNP markers distinguished between species pairs in this group. In summary, these SNPs are a valuable resource in further investigating hybridization and introgression in a range of captive and wild stocks of tilapias.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express my sincere appreciation and gratitude to my supervisors Professor Brendan McAndrew and Dr. David Penman for their invaluable guidance, encouragement and assistance throughout my study.

I am very grateful for the technical assistance rendered during my experiments by Dr. John B. Taggart, Jacquie Ireland, Dr. Kerry Bertie and also Dr. Sarah-Louise Counter. I would like to give my special thanks to Dr. Michael Bekaert, the bioinformatician at the Institute of Aquaculture, for his profound pearl scripts for data mining, which helped me a lot to extract useful information from huge datasets.

My profound gratitude extends to Dr. Christos Palaiokostas and Dr. Jan Heuman who really helped me with my experiments and analyses during my study. Moreover, I would also like to acknowledge Juliet Natabbi and Dr. Monica Betancor for all the fruitful discussion and Joanna Wilson for the proof reading.

On this occasion, I woud like to pay special thanks to all my friends and colleagues at the institute, whose company and kind support were always there, especially Khalfan M.A Al-Rashdi, Dr. Stephen N. Carmichael, Dr. Greta Carmona-Antonanzas, Dr. Beatrix Berdal, Munevver Oral, Taslima Khanam, Sienna Gray and Bridie Grant at the Institute of Aquaculture. I also thank Prof. Slamet Budi Prayitno (University of Diponegoro Semarang) and Toni Kuswoyo, S.Pi, M.Si (Freshwater Research Center Unit, Klaten) Indonesia for their

contribution on sample's collection from Indonesia.

I wish to extend my sincere appreciation to my family (Parents, sisters and brothers), who were always the source of inspiration for me and I would be unable to complete this difficult without their kind support and prayers. I also want to acknowledge the support and company of the family found in Stirling (Winarti Sarmin, Unang Mulkhan, Amjad Ullah and Khalid Shahin). Moreover, I want to thank all the members of the Asian student community, Aquaculture Student Association, Islamic Society and Indonesian student community at the University of Stirling, which shaped my character in facing hardship and happy life occasions.

I wish to acknowledge the support by MASTS (Marine Alliance for Science and Technology for Scotland) for the financial funding. I also want to say special thank to the Director General of Higher Education, Ministry of Research, Technology and Higher Education (*Kemenristek Dikti*) Indonesia, for funding my PhD scholarship at the University of Stirling.

Finally, I am most grateful to my lovely wife Zubaidah Usman and son M. Irfan Bahits who provided their infinite love and encouragement throughout my adventurous journey. I also want to acknowledge my genuine respect and thanks goes to my extended family who inspired me to persevere to seek knowledge and obtain this degree.

GLOSSARY

Allele. Variant of a gene that can vary at the nucleotide level with or without affecting phenotypic expression.

Allozyme. One of several forms of an enzyme encoded by different alleles at a locus.

Chromosome. Threadlike structure that includes DNA and proteins (containning genes arranged in a linier sequence along the thread), which can be visualized when condensed during cell division.

Clade. Group of taxa diagnosed as monophyletic by the discovery of homologies (or synapomorphis), consisting of an ancestor and all its descendants. The ancestor may be an individual, a population or even a species (extinct or extant).

Heterochromatin. Regions of chromosomes that do not include coding DNA, generally make up the structure of chromosomes, and always remains condensed during a cell's life cycle.

Introgression. Movement of genes from one species or population into another by hybridization and backcrossing; carries the implication that some genes in a genome undergo such movement, but others do not.

Isozymes. Variants of an enzyme that differ in physical properties (e.g. stability, optimum pH, isoelectric point) but catalyze the same chemical reaction.

Lineage. A series of ancestral and descendant population through time; usually refers to a single evolving species, but may include several species descended from a common ancestor.

Linkage group. Equivalent to a chromosome. A cluster of markers which do not agree with independent assortment and linked with each other at a certain distance in recombination frequency and cM.

Locus. A precise location in the genome, whether a gene is found there or not; formerly this term was used interchangeably with gene, but the definition has become more specific in the era of molecular genetics.

Physical map. A diagram of a chromosome or DNA molecule with distance given in base pairs, kilobases or megabases.

Polymorphism. genetic variation at a locus. The terms 'allele', 'mutation', and 'polymorphism' have similar meaning but different connotations, with

polymorphism being the most inclusive term. Polymorphisms use naturally occurring variations in the DNA sequence and may not have a detectable phenotypic variation. 'Mutation' usually implies a genetic change that has been experimentally induced, that occurs rarely, or that substantially alters the phenotype. 'Allele' often implies a detectable phenotype difference, which may arise from natural or experimentally induced variation.

Repetitive DNA. Regions of DNA that include the same DNA sequence repeated up to several hundred or thousands times; regions with repeated segments that involve only 2-5 base pairs of DNA are called microsatellites.

Single-nucleotide polymorphism (SNP). genetic variation that arises from a change in a single nucleotide at a locus.

Subspecies. A named geographic race; a set of populations of a species that share one or more distinctive features and occupy a different geographic area from other subspecies.

Source of definitions:
Hartl & Jones (2008)
James, M (2001).
Smith et al. (2003)
The Encyclopedia of Molecular Biology (1994).

1. GENERAL INTRODUCTION

1.1 An Overview of Tilapia Culture

The tilapias are a group of African and Middle Eastern cichlid fish that are widely cultured in both developed and developing countries (major producers include China, Egypt, Indonesia, Philippines, Thailand and Brazil), with total world aquaculture production of 4,507,002 t and total value of 7,656,257,000 USD in 2012 (FAO, 2014). Of this, 3,791,913 t was *Oreochromis niloticus*, representing 84.13% of the total. Commercial culture of tilapia occurs in approximately 140 countries and in 2012, Nile tilapia (*O. niloticus*) alone was ranked fifth among the most cultured species of fish in the world (FAO, 2014).

There are more than 70 species and strains of tilapia, whereas these species are endemic to Africa, Jordan, and Israel (Popma & Lovshin, 1995). However, few are commercially important and even fewer are of aquacultural importance. *O. niloticus* is dominate on tropical freshwater, whereas in subtropical freshwater *O. aureus*, which has increased cold tolerance, is often substituted for *O. niloticus* or used to produce a hybrid with *O. niloticus*. In brackish or saline conditions a number of species can be used but the preference, particularly in the Caribbean or Latin America, has been for one of the hybrid red tilapia based on Taiwanese or Florida strains. These strains are red *O. mossambicus* that has been hybridized to pure or hybrid *O. niloticus*, *O. aureus* and/or *O. hornorum* strains. In higher salinity and temperature conditions such as in the Red Sea, species such as *O. spilurus spilurus* and *O. s. niger* have been assessed (Cruz et al., 1990 in

Beveridge & McAndrew, 2000). In most Asian countries tilapia farmers have changed from using *O. mossambicus* or *O. mossambicus*/*O. hornorum* hybrids to *O. niloticus* or *O. niloticus/ O. aureus*.

1.2 Morphology, Morphometric and Meristic

Tilapias have fairly conventional, laterally compressed, deep body shapes, where the body is covered with relatively large, cycloid scales, which are not easily dislodged. The dorsal and anal fins have hard spines and soft rays, meanwhile the pectoral and pelvic fins are large and more anterior in an advanced configuration, therefore they provide the fish with great control over swimming and manoeuvring. The fins are also used for locomotion, and this is why cichlid fishes have red muscles designed for relatively low-speed but continuous movements (Ross, 2000). Tilapia bodies are generally characterized by vertical bars, with relatively subdued colours, and response to stress, by controlling skin chromotophores. Tilapia have well-developed sense organs, represented by prominent nares and a clearly visible lateral line. Furthermore, the eyes are also relatively large, providing the fish with an excellent visual capability (El-Sayed, 2006).

Morphometrics is a field concerned with studying variation and change in the form (size and shape) of organisms. Morphometrics enables one to describe complex shapes in a rigorous fashion, and permits numerical comparison between different forms (Webster, 2006). 21 morphometrics have been utilized to identify tilapia at the species level (**Figure 1.1**) (Barriga-Sosa et al., 2004).

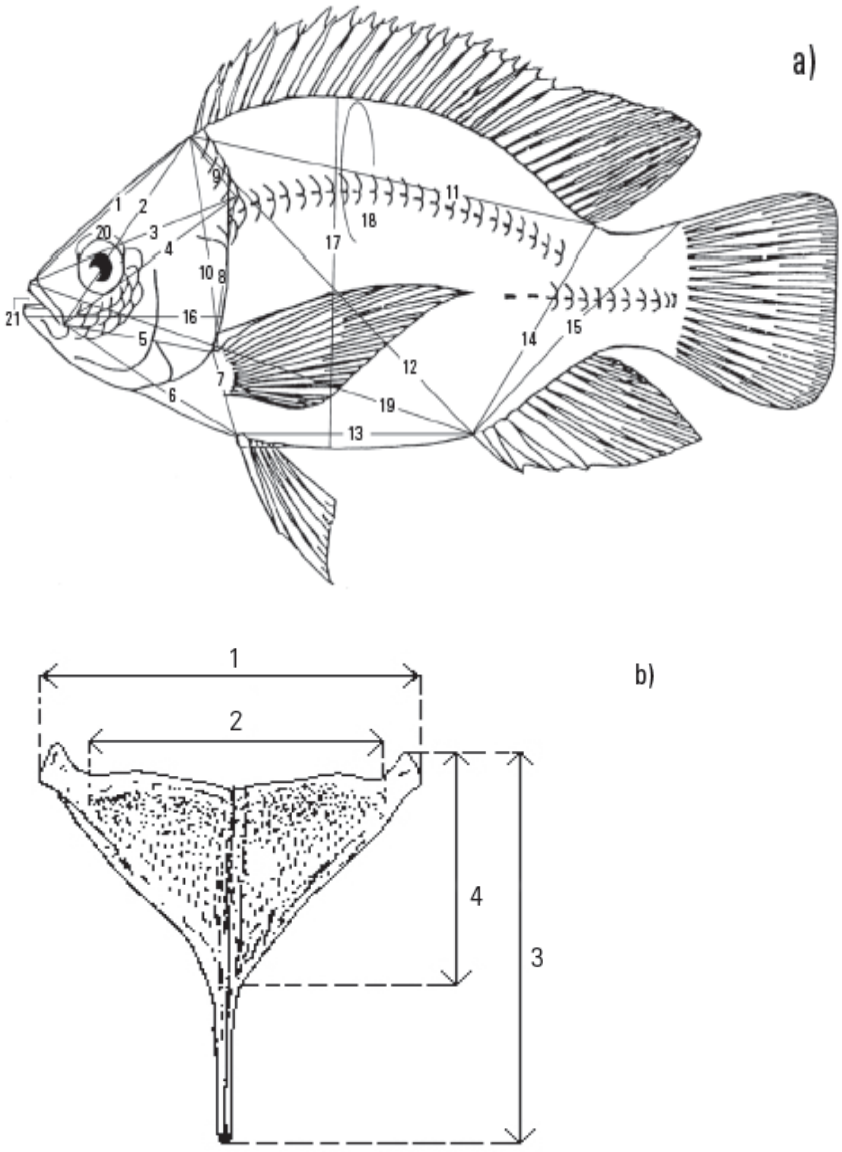


Figure 1.1 Morphometric and meristic measurements to identify tilapia at species level.

a) Twenty-one morphometric measurements for tilapia species identification (Barriga-Sosa et al., 2004), b) The tooth pharyngeal bone (TPB) variables: 1) total width; 2) width of the toothed area; 3) total length; and 4) length of the toothed area (modified from Vreven et al., 1998).

Meristic characteristic, for instance the numbers of fin rays and spines in dorsal, pectoral and anal fin are widely used for species distinction and identification. Nevertheless, the number of fin spines and/or rays of the same

species may vary from one aquatic environment to another, even one strain to another. In the tilapias low numbers of vertebrae are considered primitive (although sometimes secondarily reduced). The range in numbers of vertebrae in *Tilapia* is 26-30, in *Sarotherodon* 26-31, and in *Oreochromis* 27-34, where the mode of modes in *Tilapia* is 28, in *Sarotherodon* 29, and in *Oreochromis* 30. Meanwhile, the numbers of scales in the lateral series are generally one or two higher than the vertebrae (Trewavas, 1983). Stocks of *T. rendalli* and *O. aureus* in the Infiernillo Dam, and *O. mossambicus* at the Zicuiran reservoir can be identified into genera based on differences in the number of predorsal scales and the pharyngeal teeth morphology. *Oreochromis* was recognised into species using five meristic variables, which were: a). number of scales of the interior lateral series, b). number of predorsal scales, c). number of gill rakers (first gill raker), d). number of spines, and e). rays of dorsal fin (Espinosa-Lemus et al., 2009). Meristic differences between populations of fishes may be influenced by genetic or environmental factors, or both (Bailey & Goseline, 1955). Genes influence on morphology and physiology, shaping the behavior of an animal. Meanwhile, the environment can affect morphological and physiological development. Genes also create the scaffold for learning, memory, and cognition, remarkable mechanisms that allow animals to acquire and store information from their environment (Breed and Sanchez, 2010).

The genus *Oreochromis* has been divided by Trewavas (1983) into five subgenera: *Oreochromis*, *Alcolapia*, *Vallicolla*, *Nyasalapia*, and *Neotilapia*. The distinguishing characters of the subgenera include the size, shape, and number of tooth cuspids; shape of the preorbital bone; number of openings in the preorbital

bone; number of anal spines; relative size of belly scales; size and shape of pharyngeal teeth; presence or absence of microbranchiospines; number of lower gillrakers; enlargement of jaws in mature fishes; size of the male genital papilla; and length of the pectoral fin. Most of these characters are quantitative and display a considerable overlap in the various species of the five subgenera (reviewed by McAndrew, 2000). Morphology analysis of the paryngeal teeth have been used to differentiate among genera; *Oreochromis*-monocuspid and bicuspid, *Tilapia*-tricuspid and *Sarotherodon*-bicuspid and tricuspid. The genera *Sarotherodon* and *Oreochromis* are distinguished by the size of the belly scales, relative to the flank scales, the size of the male genital papilla, and the weight of ripe testes relative to body weight (Trewavas, 1983).

Based on anatomical characteristics of the *Oreochromis* genus, *O. mossambicus* can be readily distinguished from the other two principal cultured species, *O. niloticus* and *O. aureus,* by the presence of yellow pigmentation in the opercular region (most notable when comparing juveniles), an upturned, protruding snout and black colour in older males and lack of vertical banding on the tail. *O. niloticus* can be distinguished from *O. aureus* by the relatively strong, vertical banding in the caudal (tail) fin and by the grey-pink pigmentation of the opercular region (Popma & Lovshin, 1995). Sexual dimorphism has been noted in tilapiines including dorsal and anal fins pointed in mature males and rounded in females of *Sarotherodon galilaeus* (Linnaeus) and *Oreochromis aureus* (Steindachner) (Chervinski, 1965), pelvic fins reaching or passing the anus in males but not in females in *Tilapia zillii* (Gervais), *S. galilaeus* and *O. aureus* (Chervinski, 1983), males with one urogenital opening and females with two in *T.*

zillii, *S. galilaeus* and *O. aureus* (Chervinski, 1983), *O. mossambicus* (Peters) (Datta & Roy, 1984), a thicker and continuous dorsal fin in mature males and notched dorsal fin in females of *O. aureus* (Fishelson, 1996) and a thicker lip in upper jaw in mature males *O. mossambicus* (Seitz, 1949). In *O. mossambicus*, males have a strong and large mouths and high dorsal and anal fins, traits that are important in agonistic displays (jaw and fins), fighting and nest digging (jaw) (Oliveira & Almada, 1995).

1.3 Reproduction

In regard to reproductive behaviour and scientific names, commercially important tilapias are currently divided into three major taxonomic groups: substrate spawners (*Tilapia* spp), maternal mouthbrooders (*Oreochromis* spp), and paternal and biparental mouthbrooders (*Sarotherodon* spp). The majority of cultured tilapia belong to the maternal mouthbreeding *Oreochromis* genus. There is no or minimal sexual dimorphism or dichromatism in *Tilapia* and *Sarotherodon*, nevertheless in *Oreochromis*, males are generally bigger than females, have distinctive and conspicuous breeding colours, have enlarged jaws and unicuspid teeth when mature in some species; males in some species have a tassel-like appendage on the genital papilla (Trewavas, 1983). *Tilapia* in common with many other cichlids, lays eggs which adhere to the substrate and are guarded by both parents until the young are able to fend for themselves, meanwhile the eggs of *Sarotherodon* have only a vestigial adhesive layer and those of *Oreochromis* have none; in both these genera they are held in the mouth of one or both parents until the young are free-swimming and in *Oreochromis* after this at night and in case of

danger (Peters and Berns, 1982 cited In Nagl et al., 2001). The parental roles of the sexes in *Tilapia* are similar, however most of the fanning was done by the female, while most of the coping with intruders is by the male (Trewavas, 1983).

Sexual maturity in tilapia species is a function of age and size. In general, *O. mossambicus* reaches sexual maturity at a smaller size and younger age than *O. niloticus* and *O. aureus*. Tilapia populations in large lakes mature at a later age and larger size than the same species raised in culture ponds. Conversely, when growth is slow in farm ponds, sexual maturity will be delayed a month or two but fish may spawn at weights as low as 20 g. Under fast growing conditions in culture ponds *O. mossambicus* may reach sexual maturity in as little as three months of age, at which time they seldom exceed 60 to 100 g. In poorly fertilized ponds sexually mature fish may be as small as 15 g (Popma & Lovshin, 1995).

1.4 Food and Trophic adaptation

Many species of tilapia consume macrophytes, except populations of *T. kottae* and *T. mariae* in the extremely eutrophic Lake Barombi Kotto, which eat mainly phytoplankton. Conversely, many species of *Sarotherodon* and *Oreochromis* use epiphytic growths, some epilithic algae, and others feed almost entirely on phytoplankton. The long lower pharyngeal bone, the long array of gill-rakers and the broad skull are all measures of the large buccopharynx of *Sarotherodon* and *Oreochromis* which might be an advantage to a fish whose habits require the passage of large quantities of water through the mouth, as well as for buccal incubation. These parameters are characteristic of the most specialized plankton feeders (Trewavas, 1991). Furthermore, tilapias are more

tolerant than most commonly cultured fish to salinity, high water temperature, low dissolved oxygen and high ammonia concentrations. The tilapias most used in commercial culture are freshwater species, but all are tolerant to brackishwater. *T. zillii* is noted for this, having salinity tolerances upwards of 45 ppt (Chervinski, 1971), while conversely, *O. niloticus* is the least saline tolerant of the commercially important species, but grows well at salinities up to 15 ppt. Blackchin tilapia *Sarotherodon melanotheron* are broadly euryhaline, primarily inhabiting estuarine such as mangrove marshes, and travel freely between fresh and saltwater environments (Trewavas, 1983). Within the Brevard County portions of the Indian River Lagoon (IRL) watershed, *S. melanotheron* persists in freshwater as well as at salinities of up to 30 ppt (Dial and Wainright, 1983). *O. aureus* grows well in brackish water up to 20 ppt salinity, meanwhile *O. mossambicus* and *O. spilurus* grow and even reproduce at salinity levels near or at full-strength seawater (Popma & Lovshin, 1995).

The lethal lower temperature for most tilapia species is 10^0 or 11^0C, nevertheless *O. aureus*, the most cold tolerant, tolerates down to 8^0 or 9^0C. When hybridized with other *Oreochromis* species, cold tolerance appears to be inherited from the *O. aureus* parent (Popma and Lovshin, 1996). O. *mossambicus* are stenothermal; dying when temperatures fall below 5-10^0C (Wohlfarth & Hulata, 1983). Feeding generally ceases when water temperature falls below 16° or 17°C, meanwhile disease-induced mortalities after handling seriously constrain management below 17° or 18°C. Furthermore, reproduction is inhibited at water temperatures below 20°C and >30°C, slow in waters of 21 to 24°C and most frequent in water at 25-30°C (Popma & Lovshin, 1995).

1.5 Biogeography

The genus *Oreochromis* is widely distributed in the Rift Valley Lakes, rivers and the rivers that drain into Indian Ocean, but the number of species is low in Western Africa. On the other hand, all species of the genus *Sarotherodon*, except *S. galilaeus*, are restricted to West Africa. *O. niloticus* and *O. aureus* are distributed in the Nilo-Sudanian region. Moreover, *O. niloticus* is spreading eastwards into the Ethiopian Rift Valley and has moved southwards, colonizing all the Western Rift Lakes (Lake Albert, Lake George, Lake Edward, Lake Kivu, and Lake Tanganyika) and Lake Turkana in the eastern Rift Valley (El-Sayed, 2006). The Wami River Tilapia (*Oreochromis urolepsis hornorum*) originates from the Wami River, in eastern Tanzania. *O. urolepsis* has been found in four coastal locations in Tanzania, but the Wami River subspecies, *O. u. hornorum*, is known only from the Wami River and Zanzibar Island (Trewavas, 1983). This tilapia is famous in aquaculture because the male parental stock was crossed with female *O. mossambicus* to produce 'all male' hybrid progeny (Hickling, 1960).

The Mozambique tilapia is native to the eastward-flowing rivers of Southern Africa (down to South Africa). In the northern part of its range it is present below Kapachera Falls in the lower Shire River in Southern Malawi, the lower Zambezi, and in Mozambique in all coastal rivers down the South-eastern African coast to Algoa Bay, South Africa (Pullin, 1988). This species is more widely farmed in Asia and elsewhere than in its home of S.E. Africa, however the route of spread from Africa to Asia is unknown. *O. mossambicus* were "discovered" for the first time in Java, Indonesia in 1938, then sent to Honolulu in

1951 (Brock, 1960). All Asian *O. mossambicus* populations could be derived from Java, the origin of all of the feral population of this species established throughout the world (Pullin, 1988).

T. zillii is native to a large swath of north central sub-Saharan Africa from Senegal in West Africa through northern Zaire and the Sudan, and North into the Nile River basin and Asia Minor (Pullin, 1988). In Africa, its distribution extends from Morocco and Egypt in the North, Côte d'Ivoire and Nigeria in the West to Democratic Republic of Congo in Central Africa (El- Shazly, 1993). *T. zillii* were imported to Southern California due to their ability to feed on nuisance aquatic weeds and other macrophytes which were clogging irrigation canals; however for both biological control and aquaculture purposes this species is a poor choice because of its high fecundity and high spawning periodicity, and its slow overall growth rate to a small maximum size (Costa-Pierce, 2003).

With many different species and sub-species of tilapia, it is very difficult to differentiate hybrid or mixed populations. Wherever a mixture of tilapia species has been stocked, reproductively viable hybrids have generally resulted, and the use of external morphometric characterizations of the hybrid for species determination is fruitless (Wohlfarth & Hulata, 1983). The species intrageneric hybrids much occur more common than intergeneric. Therefore, the use of DNA fingerprint markers offer a method to accurately discern presence/absence of distinct tilapia species, the genetic composition of established, feral hybrids in new environments, and the composition of mixed species in culture (Costa-Pierce, 2003).

1.6 Genes Structure in the Genome

Across the range of cellular organisms there is an enormous diversity in gene structure. Genes are composed of a threadlike double-helical macromolecule called deoxyribonucleic acid, abbreviated to DNA, that is passed on from one generation to the next, and dictates the inherent properties of a species (Griffith et al., 1998). Bacterial genomes consist almost entirely of genes whereas in higher eukaryotes, genes can be small islands in a large sea of non-coding DNA (Primrose and Twyman, 2006). The coding sequences in some genes are interrupted by the presence of non-coding (untranslated) sequences called introns, while the parts of translated genes are known as exons. Split genes are rare in prokaryotes, but they are much commoner in eukaryotes (Martínez-Abarca & Toro, 2000). Exon length sizes are variable, but in most eukaryotic genes are less than about 300 nt long. However, introns are much more variable in length than exons and the distribution of their sizes varies greatly between different groups of organisms. The lengths of vertebrate introns are varied, about a third are less than 300 nt long, but at the other extreme, approximately 15% are over 2000 nt long (Martínez-Abarca & Toro, 2000).

The initial products of all genes are ribonucleic acids (RNAs), produced by a transcribing process of nucleotide sequence in DNA that is called transcription. RNA, a single-stranded nucleotide chain, has ribose sugar in its nucleotides, rather than deoxyribose, whereas the two sugars differ in the presence or absence of just one oxygen atom. RNA nucleotide carry the bases adenine, guanine, and cytosine, however the pyrimidine base uracil is found in the place of thymine, and it forms

hydrogen bonds with adenine just as thymine does (Griffith et al. 2008). Some RNAs are intermediates in the process of decoding genes into polypeptide chains, called “informational” RNAs, meanwhile in the others classes, RNA itself is the final product which is called “functional” RNAs. For the majority of genes, RNA is only an intermediate in the synthesis of the ultimate functional product, which is a protein: the majority of genes are messenger RNA (mRNA) (Griffith et al., 1999). The genes must be expressed to stimulate a function. Cells use the two-step process of transcription and translation to read each gene. For some genes, for instance those coding for tRNA and rRNA molecules, the transcript itself is the functionally important molecule, however for other genes the transcript is translated into a protein molecule (Brown, 2006).

1.6.1 Gene isolation by PCR

The PCR can be used as an alternative to cDNA cloning. The synthesis of DNA from an RNA template can be obtained using reverse transcription, followed by the polymerase chain reaction (PCR), producing complementary DNA (cDNA) (Dieffenbach & Dvesler 1995). cDNAs are especially useful because RNAs are inherently less stable than DNA, and techniques for routinely amplifying and purifying individual RNA molecules do not exist (Griffiths et al., 2002).

Target genes derived from PCR, which lead to direct DNA sequencing of targeted regions within the genome, can be used in SNP marker development. DNA sequencing is ideal for determining the evolutionary history of a group of organisms and for inferring evolutionary process and pattern such as the genetic basis of adaptive trait loci (for instance, genes involved in responses to fish

growth), the historical patterns of migration and expansion of animal species (e.g., from Pleistocene to current day), and the evolution of specific traits involved in taxonomic diversification (for example the origin of notochord leading to vertebrates). DNA sequencing has also enabled the development of another highly polymorphic, co-dominant marker type called single nucleotide polymorphisms (SNPs) (Allan & Max, 2010). If genomic DNA fragments are used in SNP marker development, the SNPs will mostly provide neutral markers, by contrast if cDNA is used, the SNP will be more likely to represent loci that are subject to natural selection due to the location within protein-coding regions of the genome (Freeland et al., 2011).

1.6.2 Sequence Variation Within and Between Species

Genetic variation can be described as having three main components: genetic diversity (the amount of genetic variation); genetic differentiation (the distribution of genetic variation among populations); and genetic distance (the amount of genetic variation between pairs of populations). Factors influencing diversity and differentiation mainly are population size, gene flow and reproductive system (Lowe et al., 2004). DNA sequence polymorphism can be implemented to examine variation among individuals and between species in their DNA sequences. Two types of studies can be conducted. Firstly, studying variation in the sites recognized by restriction enzymes provides a coarse view of base pair variation. Secondly, the variation can be observed base pair by base pair by DNA sequencing methods. Restriction digestion is a method that utilizes restriction

endonuclease enzymes. These enzymes will cleave double stranded DNA at a specific site (generally palindromic and 4-8 bp long) located throughout the sequence, each leaving distinctly different terminal ends (Griffith, 2008). When cleaving large pieces of DNA such a whole genome, the size of the recognition site for the restriction enzyme determines the relative number of expected digested fragment. For instance, assuming a sequence to be totally random (50%G+C), a four base recognition site occurs 4^4 or every 256 bases, while an eight-bases recognition sites would be recognized and cleaved on average every 65,536 bases (4^8) (Oveturf, 2009). The polymorphism can be detected if one of the particular bases at recognition sites is different, so there will be a restriction fragment length polymorphism (RFLP) in the population, because in one variant the enzyme will recognize and cut the DNA, whereas in the other variant it will not (Griffith, 2008).

Genetic diversity is commonly used expression to describe the heritable variation found within biological entities and can be measured at the individual, population and species level. At any particular locus diversity may be present within an individual, for example an individual may be heterozygous. It may also be present within a population, when the alleles present at the variable locus are found in different individuals (Lowe et al., 2004). Sequence variants within a species that give rise to amino acid changes are often deleterious mutations that will be eliminated by selection. Meanwhile, variations between species are as a result of fixation of mutations in the population of one species or the other (Griffiths et al. 2008). A mutation that does not impacts on amino acid changing is called a synonymous substitution, while one that does lead to an amino acid

change is called a non-synonymous substitution. Non-synonymous changes are sometimes classified as either missense mutations, where an amino acid is replaced by another amino acid, or nonsense mutations, where stop codon is introduced into the middle of the sequence (e.g., the tryptophan codon TGG could changes to the stop codon TAG) (Higgs & Attwood, 2005). The ratio of synonymous polymorphisms (dS) and non-synonymous polymorphisms (dN) indicates the evidence of selection, where for most genes and regions of the genome, dN/dS will be much less than 1 since most non-synonymous change produces a less favourable allele (Meneely, 2014).

1.7 DNA Marker Technologies

With many different species and sub-species of tilapia, extensive introductions (into approximately 140 countries) and use of interspecies hybrids in aquaculture, it is often difficult to differentiate these, or ascertain contribution to hybridized/introgressed stocks. The published descriptions of the species based on meristic and morphometric characters show considerable variation and broad interspecific overlaps (B-Rao & Majumdar, 1998; Trewavas, 1983; Wohlfarth & Hulata, 1983).

DNA marker technologies have become essential tools for aquaculture genetics research and the genetic improvement of aquaculture species. The task of DNA marker technologies is to provide the means to reveal DNA-level differences of genomes among individuals, populations and various related taxa (Liu, 2011). There have been rapid advances in molecular technologies that will also assist in the management and genetic improvement of farmed tilapia strains, for instance

by integration into selective improvement program to help geneticists to maximize genetic gains (Penman & McAndrew, 2000). Basically, the markers have been classified into two categories: type I are markers associated with genes of known function, while type II markers are associated with anonymous genomic segments (O'Brien, 1991). A range of different types of genetic markers have been used to assess genetic variation and apply this to further understanding and management of wild and cultured species and populations. These include allozymes (protein enzymes), mitochondrial DNA (mtDNA), Randomly Amplified Polymorphic DNA (RAPD), Amplified Fragment Length Polymorphic (AFLP), microsatellites and Single Nucleotide Polymorphisms (SNPs). Different type of markers, characteristics and their potential application are listed in **Table 1.1**

1.7.1 Allozyme

Allozymes are forms of an enzyme that differ in electrophoretic mobility as a result of allelic differences in a single gene. Therefore, allozyme variation in a population is an indication of simple Mendelian genetic variation (Hartl & Clark, 1997). Genetic variations detected in allozymes might be the result of point mutations, insertions, or deletions (indels). This has been widely utilized in fisheries and aquaculture including tilapia species identification (Deines et al., 2014; De Silva & Ranasinghe, 1989; Penman & McAndrew, 2000; Costa-Pierce, 2003; Sodsuk & McAndrew, 1991). Previous research reported on the alleles at twenty-two variable loci in protein from muscle, liver, and eye tissues from nine tilapiine species; *T. zillii*, *S. galilaeus*, and seven species of *Oreochromis*. Of the 22 enzymes tested adenosine deaminase (*ADA*) had unique alleles in most species,

making it useful for species identification (McAndrew & Majumdar, 1983).

Allozyme markers were used to identify fish species before the discovery of DNA markers, however the limited number of variable loci prohibits genome-wide coverage for the analysis of complex traits. Mutation detection at the DNA level that does not result in a mobility change, but replacement by a similarly charged amino acid, may not be detected by allozyme electrophoresis (Kucuktas & Liu, 2007). It is difficult to extrapolate from electrophoretic surveys of enzymes to the entire genome because the enzymes may not be representative. Furthermore, these markers present considerable difficulties for collection and storage because fish must be killed and tissues such as muscle, liver, eye, and heart need to be kept frozen until analysed (Toniato et al., 2010).

1.7.2 Mitochondrial DNA

The rate of evolution of the mitochondrial genome appears to exceed that of the protein-coding region of the nuclear genome by a factor of about 10 due to an elevated rate of mutation in mitochondrial DNA, however this region also lacks any recombination (Boore, 1999). Because of the high rate of evolution, mitochondrial DNA is likely to be an extremely useful molecule to employ for high-resolution analysis of the evolutionary process (Brown et al., 1979). The polymorphism is especially high in the control region (D-loop region), which makes this region highly useful in population genetics and as a markers in stock management for aquaculture (Liu, 2011).

Table 1.1 Types of molecular markers, their characteristics, and potential applications (modified from Liu and Cordes, 2004).

Marker	Mode of inheritance	Type	Allele numbers	Power	Application	References
Isozyme	Mendelian, codominant	Type I	2-6	Low	Species Identification	McAndrew and Majumdar, 1983; Sodsuk & McAndrew, 1991
					Population studies	Agnese et al., 1997; Adépo-Gourène et al., 2006
mtDNA	Maternal inheritance	-	Multiple haplotype	-	Species identification	Nagl et al., 2011; He et al., 2011
					Population studies	Rognon and Guyomard, 2003; D'Amato et al., 2007; Espinosa et al., 2009
RAPD	Mendelian, dominant	Type II	2	Intermediate	Species identification, ,	Bardacki and Skibinski, 1994; Dinesh et al., 1996
					Population studies	Hassanien et al., 2004
					Hybrid identification	Appleyard, 2000
AFLP	Mendelian, dominant	Type II	2	Low	Linkage mapping	Agresti et al., 2000
Microsatellite	Mendelian, codominant	Mostly Type II	Multiple	High	Linkage mapping	Lee & Kocher, 1996; Cnaani et al., 2003; Lee et al., 2005
					Species identification	Hong-Mei et al., 2009; Hassanien and Abdallah, 2005

					Population studies	Boris Briñez et al., 2011; Bhassu et al., 2004
SNP	Mendelian, codominant	Type I or II	2-4	High	Population studies	Baird et al., 2008; Messmer et al., 2011; Seeb et al., 2011; Willing et al., 2011; Scaglione et al., 2012
					Linkage mapping	Baird et al., 2008; Sarropoulou et al., 2008
					Hybrid identification	Hohenlohe et al., 2011

Sequencing of specific regions of mitochondrial DNA (mtDNA) can be used to discriminate between tilapia species (Nagl et al., 2001) and population studies (Rognon & Guyomard, 1997; D'Amato et al., 2007). Mitochondrial DNA also has been used to identify tilapia species that exist in Hawaii (Wu & Yang, 2012). One of the mtDNA genes used to distinguishing species is the conserved sequence of the 5' region of the mitochondrial gene cytochrome oxidase subunit I (COI or Cox1), for instance in Australian fish (Ward et al., 2005), marine fishes in the Northwest Atlantic Ocean, Canada (McCusker et al., 2013).

Nine different mtDNA haplotypes of seventeen natural populations of the Nile tilapia *O. nilolicus* were found in the RFLP analysis of a 1 kb portion of the DLoop region (Agnése et al., 1997). Mitochondrial DNA DNA-RFLP markers (*r16S* and *cytochrome b)* and 14 allozyme loci also have been used to evaluate the status of tilapia introduction to the Infiernillo Lake in Mexico (Espinosa et al., 2009). There were discrepancies between allozyme and mtDNA results and Trewavas' classification related to species classification. Trewavas (1983) stated that the West African and Nile River populations of Nile Tilapia belong to the same subspecies, *O. niloticus niloticus*, however an allozyme study (Rognon et al., 1996) did not show congruent results because the population from the Nile clustered to the Lake Turkana one which was described as a distinct morphological subspecies, *O. n. vulcani*. Restriction fragment length polymorphism (RFLP) of mitochondrial DNA (mtDNA) showed that all West African *O. niloticus* populations and *O. aureus* are clustered, whereas Nile *O. niloticus* populations show affinities both with West African populations and

with specimens from Lakes Tana and Turkana (Agnése et al., 1997). Furthermore, mitochondrial DNA study also stated evidence that the morphological similarity between Nile River and West African populations reflects convergence, common ancestral morphology or non-genetic environmental effects rather than derived phylogenetic relatedness (Rognon & Guyomard, 1997).

1.7.3 Randomly Amplified Polymorphic DNA (RAPD) and Restriction Fragment Length Polymorphism (RFLP)

Randomly Amplified Polymorphic DNA, first developed in 1990 (Welsh & McClelland, 1991), is a polymerase chain reaction (PCR)-based multilocus DNA fingerprinting technique. Segments of nuclear DNA are amplified using PCR with a single short PCR primer (8-10 bp). RAPDs have been used in a variety of aspects of aquaculture genetics, such as species identification, detection of interspecific hybridization, analysis of population structure, estimation of heterosis in strain crosses and analysis of genetic diversity (Liu & Cordes, 2004).

RAPD analysis was applied to three species of the tilapia genus *Oreochromis* and four subspecies of *O. niloticus*, where different RAPD fragment patterns were observed for those species, although not always for different subspecies (Bardakci & Skibinski, 1994), inheritance patterns of feral Australian *Oreochromis mossambicus* (Peters) (Pisces: Cichlidae) and an interspecific hybrid population (Appleyard & Mather, 2000). Genetic differentiation among seventeen natural populations of the Nile tilapia from River Senegal to Lake Tana and from Lake Manzalla to Lake Baringo using allozymes and RFLP have been conducted,

where sixteen variable nuclear loci showed that these populations can be clustered in three groups: 1. West African populations (Senegal, Niger, Volta and Chad drainages); 2. Ethiopian Rift Valley populations (Lake Awasa, Ziway, Koka and Awash River); 3. Nile drainage (Manzalla, Cairo, Lake Edward) and Kenyan Rift Valley populations (Lake Turkana, Baringo and River Sugata) (Agnése et al., 1997). RAPD markers were also used to distinguish three species of tilapia i.e. Mozambique/Nile tilapia, Blue/Nile tilapia and Blue/Mozambique tilapia (Dinesh et al., 1996) and population genetic studies (Hassanien et al., 2004).

RAPD tends to exhibit low levels of polymorphism among individuals of the same population, and thus are not ideal markers for parentage analysis. The dominant inheritance pattern displayed by this marker makes it difficult to distinguish between dominant homozygotes and heterozygotes (Liu, 2011). Furthermore, the presence of paralogous PCR products, amplified from different DNA regions that have the same lengths and thus appearing to be a single locus, limit the utility of this marker type (Wirgin & Waldman, 1994).

1.7.4 Amplified Fragment Length Polymorphic (AFLP) Markers.

AFLP is based on the selective amplification of a subset of genomic restriction fragments using PCR to solve the major problems of low reproducibility in RAPD (Liu, 2011). The molecular bases of AFLP polymorphism are base substitutions at the restriction sites, insertion or deletion between the two restriction sites, base substitution at the pre-selection and selection bases, and chromosomal rearrangement. This type of marker is highly reliable because it combines the advantage of RFLP and RAPD, but it is much quicker and has

higher levels of polymorphism compared to RFLP and greater reproducibility than RAPDs (Dunham, 2011).

AFLP have been widely used in tilapia for strain identification, hybrid analysis, population structure and sex-linked markers. AFLP has been used to determine the status of three commercial strains of tilapia New Gift tilapia, GenoMar tilapia and the hybrid tilapia (*O. niloticus*♀ × *O. aureus*♂) (Yun et al., 2008). Combining microsatellite and AFLP markers, a genomic map for each of the parents in an *O. mossambicus x (O. aureus x* red and wild *O. niloticus)* was constructed (Agresti et al., 2000). In spite of its popularity, AFLP has two fundamental flaws that prohibit its wider applications - dominant inheritance and lack of information to link to it to genome sequence information (Liu, 2011). Beside, these markers are more technically specialized, and required expensive equipment such as DNA sequencers (Dunham, 2011).

1.7.5 Microsatellites

The simple sequence repeats (SSRs) or microsatellite, representing a unique type of tandemly repeated genomic sequences, were discovered at the end of the 1980s. They became the most preferred marker type because of their characteristics – biparental, codominant inheritance, high level of polymorphism, abundance, roughly even genome distribution, and small locus size that facilitate PCR-based genotyping (Tautz, 1989). These markers are present in both coding and non-coding regions but more likely to be type II markers (e.g. associated with anonymous genomic segments) (Zane et al., 2002).

There are several forms of microsatellites; dinucleotide, trinucleotide, and

tetranucleotide repeats, but dinucleotide repeats are the most abundant forms. Of the dinucleotide repeat types, (CA)n is the most common, followed by (AT)n, and then (CT)n (Xu et al., 2006), while (CG)n repeat type is relatively rare in the vertebrate genomes. A/T-rich are also generally more abundant than G/C-rich type among trinucleotide and tetranucleotide repeats (Tóth et al., 2000). In tilapia, (CA)n microsatellites were found abundantly (Carleton et al., 2002; Lee et al., 2005).

In Nile tilapia, these markers have been widely used to assess genetic diversity (Brinez et al., 2011), population structure and gene flow (Bhassu, Yusoff, & Panandam, 2004; Hassanien et al., 2004) and quantitative trait loci (Cnaani et al., 2003; Lee et al., 2005). Micosatellite and allozyme studies have been used to confirm that the *O. esculentus* population from Lake Kanyaboli has not hybridized with *O. niloticus* (Agnese et al. 1999). One study demonstrated that a polymorphism of 17 di-nucleotides between the microsatellite alleles in the tilapia PRL I promoter is associated with growth in salt water and with differential expression of the PRL I gene (Streelman and Kocher, 2002). Microsatellite markers have also been used for development of linkage maps (Lee et al., 2005) and distinguishing three species of tilapia, where *O. aureus* was more closely related to *O. mossambicus* than to *O. niloticus* (Hong-Mei et al., 2009).

Microsatellite markers are taxon-specific (species or closely related groups of species) and isolation was fairly laborious (preparation of DNA libraries, enrichment and probing with repeat motifs, sequencing of positive clones). This type of marker is often highly polymorphic, shows codominant inheritance and small locus size, nevertheless it is practically impossible to develop hundreds of

thousands of microsatellite markers (Liu, 2011). Amplified microsatellite DNA of the same size can have different sequences, so in fact being different alleles, which may limit their usefulness in clearly discriminating species and in studying hybridization (Toniato et al., 2010).

1.7.6 SNP (Single Nucleotide Polymorphisms)

Single Nucleotide Polymorphisms are single base-pair positions in genomic DNA at which different sequence alternatives (alleles) exist in the population. In highly outbred populations, such as humans, polymorphisms are considered to be SNP only if the least abundant allele has a frequency of 1% or more. This is to distinguish SNPs from very rare mutations (Primrose & Twyman, 2006). A SNP within a locus can produce as many as four alleles, each containing one of four bases at the SNP site: A, T, C, and G. However most SNPs are usually restricted to one of two alleles (quite often either the two pyrimidines C/T or the two purines A/G) and have been regarded as bi-allelic (Liu, 2011). SNP sites are abundant throughout the entire genome (3×10^7 different sites in humans) and show co-dominant inheritance (Dunham et al., 2004). In various organisms, SNPs are found anywhere from every 76 to every 2000 bp and are found in both non-coding and coding regions (Liu, 2007). SNP replacement polymorphisms change the amino acid, and SNP synonymous polymorphisms change the codon but not the amino acid. SNP regulatory polymorphisms can occur which alter gene regulation (Durham, 2011).

A large number of polymorphic SNP markers can be identified both from gene and alter genomic (non-coding) regions. In contrast to microsatellites which

are usually type II markers, SNPs, the most abundant polymorphic marker in organism are more commonly type I markers. SNPs can be used to analyze QTL (Quantitative Trait Loci) regions associated with important production traits such as growth, disease resistance and cold tolerance (Liu & Cordes, 2004). Next Generation Sequencing (NGS) offers new opportunities to rapidly and cheaply isolate very large numbers of genetic markers, primarily SNPs and microsatellites, and to do so from structured samples (families, populations or species) in a way that will identify markers associated with specific traits or differences between populations or species (Baird et al., 2008). Previously, DNA sequencing was performed by the Sanger method, which has an excellent accuracy and reasonable read length but very low throughput. It was used to obtain the first consensus sequence of the human genome project in 2001 (Venter et al., 2001 *In* Zhang et al., 2011). Since then, several genomes in aquatic organisms have been sequenced with NGS with varying degrees of coverage (Baird et al., 2008).

A common strategy for NGS is to use DNA synthesis or ligation process to read through many different DNA templates in parallel (Fuller et al., 2009). Therefore, NGS reads DNA templates in a highly parallel manner to generate massive amount of sequencing data, however, the read length for each DNA template is relatively short (350-500 bp) compared to traditional Sanger sequencing (1000-1200 bp). Traditionally, a standard DNA sequencing workflow involved three key steps: library preparation, sequencing, and data analysis. Library preparation is accomplished by random fragmentation of DNA, followed by *in vitro* ligation of common adaptor sequences. There are five NGS platforms available commercially, including the Roche GS-FLX 454 Genome Sequencer

(originally 454 sequencing), the Illumina Genome Analyzer (originally Solexa), the ABI SOLiD Analyzer, Polonator G.007 and the Helicos HeliScope. Nevertheless, the last two platforms are not widely used (Liu, 2011).

The ability to produce gigabases of DNA sequence in a short time and at minimal cost using platforms such as Illumina, Roche 454 and AB SOLID means that genomes can now be sequenced from scratch within the limit of a normal research grant. When selecting an NGS platform, laboratories working with non-model species must consider the cost, research question and availability of resources for sequence assembly. Both Roche 4564 and ABI SOLiD use emulsion PCR as template amplification, while the Illumina platform uses bridge PCR. Roche 454 has the longest read length reaching 400 bp compared with 100 bp in Illumina and 50 bp in ABI SOLiD, however it has the lowest capacity, which is only 0.4-0.6 Gb per run. The ABI SOLiD platform has the biggest capacity and the longest run time, which is 25-30 Gb for 6-7 days in single-end library and 50-60 Gb for 12-14 days per run in paired-end library among those platforms. Using the SOLiD system can sequence more individuals at very high coverage, where several million reads are obtained from each library (Everett et al., 2011). Conversely, the Illumina platform has a capacity just 0-10 Gb per run below the ABI SOLiD, nevertheless it has almost half the time of the ABI SOLiD (Liu, 2011).

The read length (the actual number of continuous sequenced bases) for NGS is much shorter than that attained by Sanger sequencing which reach 100-1,200 bp, however at present, NGS can provide 50-500 continuous base-pair reads. The Illumina/Solexa Genome Analyzer is widely recognized as the most

adaptable and easiest to use sequencing platform, because it has superior data quality and read lengths. To date, the new Illumina HiSeq 2000 Genome Analyzer is capable of producing single reads of 2 x 100 basepairs (pair-end reads), and generates approximately 200 giga basepair (Gbp) per run (Zhang et al., 2011).

NGS technologies can be applied in a variety of fields, such as de novo genome sequencing, whole genome resequencing, or more targeted discovery of mutations or polymorphisms, transcriptome analysis, large-scale analysis of DNA methylation and genome-wide mapping of DNA-protein interactions. It also has advantages relative to Sanger sequencing, because of efficiency in making *in-vitro* sequence libraries, enabling a much higher degree of parallelism of conventional sequencing in array-based sequencing and immobilizing to a planar surface of array features, so they can be enzymatically manipulated in a single reagent volume (Liu, 2011). The initial generation of the primary genetic sequence of a particular organism is called *de novo* sequencing, thus it might be possible to determine a detailed genetic analysis of any organism, for instance assembly of the rice pathogen *Pseudomonas syringae* yields >75% of the predicted genome covered by scaffolds over 100,000 bp (Reinhardt et al., 2009), and 10% assembly of a guppy genome (Willing et al., 2011). Based on genome sequence, it also can identify single-nucleotide polymorphisms (SNPs), indels, copy number and haplotype structural variations in populations (Etter et al., 2011).

Restriction-site Associated DNA sequencing (RADseq), a method that samples at reduced complexity across target genomes, promises to deliver high resolution population genomic data-thousands of sequenced markers across many individuals-for any organism at reasonable costs. It was developed to speed

discovery of SNPs and is particularly attractive in systems lacking a reference genome (Etter et al., 2011). It can be used to detect restriction site presence-absence polymorphisms, by identifying a marker that is present in one set of individuals but absent in another, indicating a variation in the restriction site (Davey & Blaxter, 2010). RADseq was first applied to investigation of the genetics of an important ecological trait (lateral plate phenotype) in the three-spine stickleback (*Gasterosteus aculeatus*) (Davey and Blaxter, 2011). Sequence RAD tags have several attractive features for genetic mapping. First, RAD tags create a reduced representation of the genome, allowing over-sequencing of the nucleotides next to restriction sites and detection of SNPs. Second, a suitable number of markers for an application can be selected and increased almost indefinitely by choice of restriction and additional enzyme, respectively. Third, the approach is amenable to genotyping populations by bulk segregant analysis and also multiplexed genotyping of individuals for fine-scale mapping. RADseq identified more than 13,000 polymorphic markers in the freshwater low lateral plate phenotype (Bear Paw, BP) and the saltwater complete plate ancestral (Rabbit Slough, RS) populations of threespine stickleback (Baird et al., 2008). SNPs have been used to identify genetic population structure of sea lice among farmed and wild host salmon from 12 Pacific Ocean samples ranging from the Bering Sea to southern Vancouver Island (Messmer et al., 2011). One study using *Sbf*I RAD-seq data demonstrated that orthologous *Sbf*I RAD loci were identified across closely and distantly related species amongst ten teleost fish. This also suggests that similar meta-datasets could be utilized in the prediction of evolutionary relationships across populations and species (Gonen et al., 2015).

1.8 Restriction-site Associated DNA (RAD) Sequencing Tags using High-throughput Illumina Platform

RADSeq based on the Illumina platform combines two standard molecular biology techniques, restriction enzymes and molecular identifiers (MID). Restriction enzymes cut DNA into fragments then added MID associate sequence identifiers to particular individuals or pooled groups. The illumina platform currently permits sequencing out to 150 bases, and thus approximately 300 bases flanking each restriction site can be screened for polymorphisms (Davey & Blaxter, 2010). Despite the power of massively parallel sequencing platforms, a drawback is the short length of the sequence reads produced if it is to be used to put together a complete genome. Short reads can be locally assembled into longer contigs using paired-end sequencing of restriction-site associated DNA (RAD-PE) fragments. One difficulty in assembling a genome from short reads is bridging repetitive sequences, which may exist in thousands to millions of locations in a genome, and are nearly indistinguishable in the context of a short sequence read. To overcome this problem, the genome is physically breaked into smaller fragments, cloning and sequencing each fragment independently (Etter et al., 2011).

RAD-Tag Sequencing is a reliable tool to determine the presence/absence of SNPs in species, sub-species or populations. Using PE RAD-seq in order to produce extended contigs flanking a restriction site, it was possible to reconstruct one tenth of the guppy genome represented by 200-500 bp contigs associated to *Eco*RI recognition sites. It was also possible to produce 283,842 RAD tags of

which ~50% overlapped. Albeit, this ratio could be significantly increased either by reducing the insert size of the library or by sequencing with longer read length (Willing et al., 2011). RAD paired-end contigs provide a low-cost method for SNP discovery in a format suitable for high-throughput genotyping platforms that require flanking sequence for primer design (Etter et al., 2011). If a reference genome is available, raw sequence reads can be aligned to the reference sequence, and SNPs and indels identified using existing next-generation sequencing bioinformatics tools, such as Bowtie (BWA) and SAMtools. Nonetheless, if a reference genome is not available, RAD tags can be analysed *de novo*, where identical reads are aggregated into unique sequences and treated as candidate alleles at the same locus (Davey and Blaxter, 2011).

The reference sequence can be a whole genome sequence assembly or an existing normalized sequence collection such as BAC-end or expressed sequence tag (EST) sequences. Alternately, individual reads from the SNP discovery sequencing project can be assembled into contigs, and the contig consensus sequences can serve as a pseudoreference (Liu, 2011). ESTs (Expressed Sequence Taq) are transcribed sequences, so that EST-derived SNPs are associated with actual genes, allowing use of gene-associated SNPs for mapping and subsequent use in comparative genome studies (Sarropoulou et al., 2008). However, SNPs derived from ESTs can only be identified where contigs contain a minimum of two sequences and the frequency of sequencing of ESTs is not random. Large-scale sequencing is required to identify SNPs from rarely expressed genes. Besides, SNP discovery rates could be lower in coding regions because of evolutionary restraints of selection pressure (Liu, 2011).

RADseq has been used to retrieve 2923 (33.7 % of total) candidate species-specific SNPs for distinguishing two species, rainbow trout (*Oncorhynchus mykiss*) and native cutthroat trout (*Oncorhynchus clarkia lewisi*). Further analysis also provide flanking sequence for design of qPCR-based TaqMan® using a subset of 50-100 loci which potential to be applied in distinguishing hybrids and quantifying genome-wide average levels of introgression between two species which impact on the genetic purity of these species (Hohenlohe et al., 2011).

1.9 Need to manage tilapia broodstock

In aquaculture, there are major breeding programmes for genetic improvement of tilapia (e.g. GIFT – Ponzoni et al., 2011), however also many examples of declining performance in captive stocks, which in some cases have been linked to reduced genetic variation through poor management (McKinna et al., 2010). In wild populations, introgression resulting from species movements is perceived as a major threat to biodiversity (D'Amato et al., 2007). It is the process of transferring a portion of genetic material from one species into the genetic background of another and sometimes involving significant genomic quantities (Anderson, 1949). Genetic transfer occurs via a fully or partially fertile interspecific hybrid, which hybridizes with one or both of the parental species (backcrossing) (Lowe et al., 2004). The generally high propensity of fishes to hybridize, combined to (1) the wide use of hybrids and synthetic strains in aquaculture (Bartley et al., 2001) and/or (2) the difficulty of accurate identification of closely related species in some groups, is likely to have led to introgressions in

some aquaculture or laboratory strains (Bezault et al., 2007). Introgression of unwanted genes into cultured stocks can lead to major declines in productivity, for instance alleles from a congeneric species, *Oreochromis macrochir* (Boulenger) introgressed into *O. niloticus* culture stock caused 20% reduction in growth performance (Micha et al., 1996). Based on mtDNA control region sequences, there was no introgression detected between the native, *Oreochromis esculentus* and the invasive species, *O. niloticus* in Lakes Kanyaboli and Namboyo, however based on eight nuclear microsatellite loci, there was a low level of nuclear admixture, primarily from *O. niloticus* to *O. esculentus* (Angienda et al., 2011). One study indicated there was an introgression by mtDNA of *O. leucosticus* to *O. niloticus* derived from three hot spring populations in Kenya, while microsatellite analysis suggested that some nuclear genes might also have crossed the species barrier (Ndiwa et al., 2014).

The genetic diversity in tilapias has been researched over many years; however a reliable method for accurate assessment of hybridisation and introgression still remains problematic. With many different species and many captive aquaculture stocks of tilapia around the world, there is a strong need for better tools for such analyses. In particular, larger numbers of species-specific markers would be very useful. High-throughput sequencing offers new opportunities to rapidly isolate and genotype very large numbers of genetic markers, primarily SNPs, and to do so from structured samples (families, populations or species) in a way that will identify markers associated with differences between populations or species (Baird et al., 2008). This research will set out to find out DNA sequence differences that distinguish between tilapia

species and sub-species using genome-wide (RAD sequencing) and a candidate gene (adenosine deaminase *ADA*) approaches and to apply these to important questions, such as phylogenetic relationships and potential for application in the studies on hybridization and introgression between species in aquaculture and in the wild. Another study was also conducted to confirm the species status of the RADseq and ddRADseq study samples using mtDNA barcoding DNA for tilapia species. Mitochondrial DNA sequences based on Cytochrome c oxidase subunit I (COI) were used to distinguishing tilapia species and generating a parallel phylogenetic tree among these species.

1.10 Aims and Objectives

The main objectives of this research were as follows:

1. To look for species-specific markers that distinguish between tilapia species using candidate gene adenosine deaminase (*ADA*) and genome wide (both standard and double digest RAD-sequencing) approaches.
2. To construct phylogenetic trees both from *ADA* gene and RAD-sequencing.
3. To build a physical map of these markers derived from standard and double digest RAD-seq.
4. To develop such sequence differences derived from *ADA* into allelic discrimination marker assays.
5. To confirm species authentication in tilapia using cytochrome c oxidase subunit I (COI).

1.11 Thesis Structure

This research is organized as follows: beginning with a General Introduction (Chapter I) and General Materials and Methods (Chapter 2), followed by three experimental chapters (3-5) and a final discussion (Chapter 6). The first experiment (Chapter 3) describes SNP marker developement from the Adenosine deaminase (*ADA*) gene, consisting of mRNA isolation, characterization, a gene tree and allelic discrimination between tilapia species. Chapter 4 describes the second experiment entitled "Species-specific SNP markers and their genomic distribution in tilapia based on standard RADSeq" using Restriction Enzyme from *Sbf*I involving seven tilapia species and *Pelvicachromis pulcher*. This chapter consists of DNA library preparation for standard RADseq, sequencing, analysis, phylogenetic tree reconstruction, species –specific diagnostic SNP marker and physical mapping. Chapter 5 describes SNP markers development using double digest RADseq involving a much broader range of tilapia species, sub-species and populations. This chapter consisted of DNA library preparation for ddRADseq, sequencing, analysis, phylogenetic tree reconstruction, species –specific diagnostic SNP marker, and physical mapping. Species authentication using DNA barcode – Cytochrome c oxidase subunit I (COI) was also conducted in parallel with SNP marker development both from standard RADseq and ddRADseq. The last, Chapter 6 is a general discussion about the SNP marker development from *ADA* gene and genome approaches, allelic discrimination assay using KASP PCR, the strength and weakness of such technology in revealing the genetic evidence across tilapia species and populations.

2. GENERAL MATERIALS AND METHODS

In general, the materials and methods used during the experiments are listed in this chapter, nevertheless some of them which were specific to certain experiments and are described in the relevant chapter.

2.1. Materials

Biological materials

Fin samples were collected from 10 different tilapia species and one individual of *Pelvicachromis pulcher* (Boulenger) as an out-group (**Table 2.1**). Three of the *Oreochromis* species (*O. niloticus*, *O. mossambicus* and *O. aureus*) and *T. zillii* consisted of at least two populations, while the remaining species were only represented by samples from one population. All of them are considered as being pure species in the wild and culture. The origin and description of samples are also described in the related chapter. Samples were stored in 99% ethanol at -20°C until required.

Table 2.1 Origin of samples used in SNP markers development in tilapia.

No	Species/sub species	Strain/ Population	Origin
1.	*O. niloticus*		
	a. niloticus	a. Stirling	L. Manzala, Egypt
		b. Kpandu	Ghana
		c. Nyinuto	Ghana
	b. *cancellatus*	a. Hora	Ethiopia
		b. Koka	Ethiopia
		c. Metahara	Ethiopia
2.	*O. mossambicus*	a. Stirling	Zimbabwe
		b. Natal	South Africa
3.	*O. aureus*	a. Stirling	L. Manzala, Egypt
		b. Ain Faskha	Israel
4.	*O. karongae*	Stirling	L.Malawi, Tanzania
5.	*O. u. hornorum*	Israel	Israel
6.	*T. zillii*	a. Stirling	L. Manzala, Egypt
		b. Ghana	Ghana
7.	*S. galilaeus*	Israel	Israel
8.	*O. andersonii*	Itezhi-tezhi	Zambia
9.	*O. macrochir*	Itezhi-tezhi	Zambia
10.	*S. melanotheron*	Ghana	Ghana
11.	*Pelvicachromis pulcher*	NA	Stirling

The key equipment used in the experiment is listed in Table 2.2, while consumables are listed in Appendix II-1.

Table 2.2 Key equipment used for SNP markers development in tilapia.

No	Equipment	Specification	Purpose
1.	Large agarose gel apparatus & multi-channel compatible combs		DNA quality checks
2.	Small gel apparatus	Tray c. 10 x 12 cm; preferably UV transparent	Band excision and library quality checks.
3.	8 (or 12) Channel multi pippetes	Low volume (1-10 µL)	Accurate dispensing of 3 µL aliquots
4.	Accurate adjustable pipettes with filter tips	Volume: 2, 10, 10, 200 and 1000 µL	
5.	Plate vortex		
6.	Microtitre plate compatible centrifuge		
7.	Gel documentation system		
8.	Accessible UV/blue light box		Gel band cutting
9.	96 well plate thermocycler		
10.	96 well plate qPCR cycler or UV microtiter plate reader		Fluorescent quantification of DNA
11.	Microtube variable speed centrifuge	Capable at least 12 K g	
12.	Tube compatible heat block	Volume 1.5 mL	incubation
13.	96 well tube racks	0.2 mL PCR microtubes	

2.2. Methods

2.2.1 Genomic DNA extraction

Total genomic DNA was extracted using the Realpure Genomic DNA Extraction Kit (Durviz S.L). In general, DNA extraction consists of five steps: cell lysis, RNAse treatment, protein precipitation, DNA precipitation and DNA dissolution. Initially, each fin sample (three punches of ⌀ 2 mm) was placed in 300 µL lysis solution, then 3 µL Proteinase K (10 mg/ml) was added, mixed and incubated overnight at 55ºC or until total lysis occurred (observed solution inside the tube), then 3 µL RNase was added to each sample, mixed by vortexing and incubated at 37ºC for 15-60 min. In the protein precipitation step, samples were allowed to cool to room temperature, then 180 µL of protein precipitation solution was added and mixed by vortexing at high speed for 30 s. Next, the samples were centrifuged at 14,000 *g* for 10 min. The precipitated protein formed a pellet at the bottom of the tube. To precipitate the DNA, the supernatant containing the DNA was poured into a new microfuge tube containing 150 µL isopropanol, mixed by inversion (4-6 times) and centrifuged at 14,000 *g* for 3 min. The supernatant was removed by decanting and/or pipetting. The pellet was washed in 1 mL of 70% ethanol, inverted several times and centrifuged at 14,000 *g* for 2 min. Most of the ethanol was removed with a pipette (1 mL), then samples were briefly spin again (2 s pulse). All remaining ethanol was removed with a small volume pipette (20 µL) without touching the pellet. Next, the tubes were inverted and air-dried on absorbent paper for 30 min to 1 hour. The final step was DNA dissolution. To each

sample, 40-50 µL 5 mmol/L Tris solution was added and the pellet dissolved, then incubated at 65ºC for 1 hour or overnight at room temperature (RT), approximately 20-22ºC. DNA samples were stored at -20ºC for long-term storage.

2.2.2 DNA Quantification and Visualization

Extracted DNA was quantified using a Nanodrop (ND 1000 Spectrophotometer, NanoDrop Technologies Inc., Montchanin, DE) and diluted to 50 ng/µL in 5 mmol/L Tris. The purity of DNA was assessed by the ratios of absorbance A 260/A280 and A260/A230. The value of wavelength reading 260 nm allows calculation of DNA concentration in the templates, meanwhile the value of 280 nm showed the amount of protein in the samples. Absorbance at 230 nm is accepted as being the result of other contamination. Good DNA samples give a value for A260/A280 of approximately 1.8-2.0, while expected A260/A230 values are commonly in the range of 2.0-2.2.

To prepare a 1% agarose gel, 25 mg agarose was heated up and dissolved in 25 mL TE buffer using a microvave. The fluorescent visualization dye EtBr (0.5 µL of 5 mg/mL stock) was added when the liquid had slightly cooled. The molten agarose was poured into a tray, and a comb/well former was immediately inserted into the gel, which was allowed to set for at least 50 mins at room temperature (it could be put in fridge to speed up setting). The comb was removed carefully, and the gel was placed in a buffer tank, covered with 0.5x buffer and left for at least 5-10 mins before adding samples. The sample and loading buffer (15% Ficoll equal to 6x loading buffer) with ratio 1:5 (1x loading buffer) were first added to the well so it was ¾ full and loading dye/buffer was used to fill the well. Fragments sizes

and approximate amount of DNA were estimated by calibrating DNA concentration against known size standards, e.g λDNA/HindIII digested DNA (Appendix II-2). Genomic DNA was expected to have a high molecular weight with a single band present above 23 Kb without any significant DNA degradation.

2.2.3 Extraction of DNA from Agarose Gel

The MinElute Gel Extraction Kit provides spin columns, buffers, and collection tubes for silica-membrane-based purification of DNA fragments of 70 bp – 4 kb from up to 400 mg gel slices. The spin columns are designed to allow elution in very small volumes (as little as 10 µl), delivering high yields of highly concentrated DNA. An integrated pH indicator allows easy determination of the optimal pH for DNA binding to the spin column. DNA fragments purified with the MinElute system are ready for direct use in PCR and restriction digestion in RAD library preparation.

The DNA fragments from the agarose gel were excised with a clean, sharp scalpel. The gel slices were weighed in a colourless tube. Then, 3 volumes of buffer QG were added to 1 volume of gel (100 mg~100 µl) and incubated with agitation on a rotator for 30 min (until the gel slice has completely dissolved). With 0.5x TAE gel buffer there is usually no need to adjust pH at 7.8-8 (buffer is yellow). The liquid was spun briefly after dissolving. To increase the yield of DNA fragments, 1 gel volume of isopropanol was added to the sample, mixed and spun briefly. Next, a QIAquick spin column was placed in the provided 2 ml collection tube. Sequential aliquots of QG buffer/gel mix were loaded in single column and spun for 10 sec (between aliquots). The column was placed in the

collection tube each time.

After the final sample was applied to the QIAquick column, the column was allowed to stand for 5 minutes to bind DNA and centrifuged at 17,800 g for 1 minute. The remaining solution in the collection tube was discarded and the flow-through. The QIAquick column was placed in clean collection tube. QG buffer 0.5 mL was added to the column and spin for 1 min, then placed in a clean collection tube. Wash buffer (plus ethanol) 0.75 ml was added to the QIAquick column, the column was allowed to stand for 5 minutes to wash, then centrifuged for 1 minute. The column was placed in a clean collection tube with the lid open and spun again for 1 minute. Next, the QIAquick column was placed in a final nuclease free 1.5 micro centrifuge tube and incubated at 60^0C for 5 minutes. To elute DNA, 20-30 µl buffer EB (10 mM Tris-Cl, pH 8.5) or water (pH 7.0-8.5) was added to the center of the QIAquick membrane and centrifuged for 1 minute. Alternatively, for increased DNA concentration, buffer EB or water was added two times (approximately between 10-15 µl), the column let to stand for five minutes, then centrifuged. To be analysed on a gel, one volume of loading dye was added to five volumes of purified DNA then the solution was mixed, by pipetting up and down before loading onto gel.

2.2.4 Purification of DNA from PCR

First of all, 3-5 volumes (3 volumes in RAD sequencing and 5 volumes in *ADA* and COI gene) of buffer PB was added to 1 volume of PCR sample, mixed properly, and briefly spun. The colour of the mixture should be yellow, if pH indicator I had been added to buffer PB; however if the colour is orange or violet,

1-3 µl of 3 M sodium acetate (pH 5.0) was added and mixed briefly. QIAquick spin columns were placed in the provided 2 ml collection tubes. The samples were applied to the QIAquick columns, allowed to stand for 5 minutes to bind DNA, and centrifuged at 17,800 g for 1 minute. The remaining solution in the collection tube was discarded, then the QIAquick column was placed back into the same tube. Buffer PE (0.75 ml) was added to the QIAquick column, allowed to stand for five minutes to wash, then centrifuged for one minute. This step was repeated two times. Next, QIAquick columns were placed in a clean 1.5 ml microcentrifuge tube and heated up to 60^0C for 5 minutes. To elute DNA, 20-30 µl buffer EB (10 mM Tris-Cl, pH 8.5) or water (pH 7.0-8.5) was added to the center of the QIAquick membrane and the column centrifuged for 1 minute. Alternatively, for increased DNA concentration, buffer EB or water was added two times (approximately between 10-15 µl), the column let stand for five minutes, then centrifuged. To be analysed on a gel, one volume of loading dye was added to five volumes of purified DNA and the solution mixed by pipetting up and down before loading the gel.

2.2.5 Magnetic Bead Clean-up of DNA

A paramagnetic bead approach gave a purer product with more consistent removal of unwanted smaller fractions (PCR primers and primer dimer product). This protocol was used to purify 50 µl ddRAD library. Initially, fresh 73% ethanol (730 µl 100% EtOH + 270 µl nuclease free water) was prepared. The Ampure beads were removed from the fridge and equilibrated to room temperature. Meanwhile, the heat block was set up to 60^{o}C to warm up

50 µl of Qiagen EB buffer. The beads were mixed well, then an equal volume of beads was carefully added to the DNA solution. They were mixed gently by pipetting, but without vortexing or flicking the mix, so the solution remained in the bottom of the tube. The mixture was left at room temperature for 5 minutes. The tube cap was opened and the microfuge tube was placed in a magnetic stand (the tube remains undisturbed in the magnetic stand until the beads drying step at 60°C in a heat block**).** It was left for 3-4 min until the beads had fully migrated to the side of tube, and the supernatant was carefully discarded. 190 µL 73% EtOH wash was added to the tube, and left for 30-60 s. The washing was repeated for a second time, ensuring that all wash solution was removed. The tube was removed from the rack and placed in a 60°C heat block for 2-3 mins to completely dry the beads. The beads were gently resuspended in EB buffer (usually c. 20 µL) by gentle pipetting. The tube was incubated in a heat block (60°C) for 2-3 m. The tube cap was opened and the microfuge tube was placed in magnetic stand. It was left for 3-4 m until the beads had fully migrated to the side of the tube. All of the supernatant was carefully pipetted into a new tube.

2.2.6 Standard Restriction-site Associated DNA sequencing (RAD-seq) library preparation and sequencing

Initially four RAD libraries were constructed, each comprising pooled DNA from 11-12 individually barcoded fish. Later a fifth library comprising DNA from a further 7 individuals was made. RAD libraries are prepared according to Etter et al. (2011). The first four libraries consisted of 6 tilapia species (6-8 each), the fifth contained five individuals of *S. gallilaeus* and one specimen of

Pelvicachromis pulcher. High quality genomic DNA was digested for 45 min at 37^0C in a 11 µl reaction volume containing 1.10 µl NEB 4 Buffer x10, 5.5 µl DNA template (0.045 - 0.05 µg/µl), 0.18 µl of 20 units (U)/µl SbfI (New England Biolabs [NEB]) and 4.22 µl H_2O. Samples were heat-inactivated for 23 min at 65^0C. For ligation, 0.6303 µl of barcoded SbfI-P1 adapter (100 nM) was added to each sample along with 0.132 µl rATP (100 mM), 0.22 µl NEB2 Buffer, 0.11 µl T4 DNA Ligase (2K U/µl), 0.4477 µl H_2O and incubated at room temperature (RT) for 60 min. A master mix of ligation reaction was prepared for each library. Samples were again heat-inactivated for 20 min at 65^0C, combined into 4 libraries (n=11-12 samples each), and randomly sheared using a Covaris sonicator to an average size of 450 bp. A 123 µl aliquot of each library was taken and sheared 10 times for 30 sec on the high setting, following the manufacturer's instructions.

The sheared sample was purified using a QIAquick Spin column (Qiagen) and run on a 1.1% agarose (Sigma), 0.5x TAE gel. A smear of DNA approximately 350–650 bp was isolated with a clean razor blade and purified using the MinElute Gel Extraction Kit (Qiagen). The Quick Blunting Kit (NEB) was used to polish the ends of 20 µl of eluted DNA in a 26 µl reaction volume containing 2.5 µl Blunting Buffer, 2.5 µl dNTP Mix (1m*M*) and 1.0 µl Blunt Enzyme Mix. The sample was incubated at 37^0C, then purified with a QIAquick column, and eluted in 45 µl EB. This prepares the DNA fragments for ligation to the P2 adapter, which possesses a single 'T' base overhang at the 3' end of its bottom strand.

To the eluate from the previous step, 1 µl dATP (10m*M*), 3.0 µl Klenow (3'-5' exo^- to add 3'adenine overhangs to the DNA) (0.2 U/µl) were added and incubated at 37^0C. After another purification and elution in 45 µl EB, the P2

adapter (a divergent modified Illumina adapter that contains a 3' dT overhang) was ligated onto the end of blunt DNA fragments with 3' dA overhangs. 1.0 µl P2 adapter (10 µM), 0.5 µl rATP (100m*M*), and 0.5 µl concentrated T4 DNA ligase (2K U/µl) was added to the reactions, and incubated at 37^0C. The sample was purified and eluted in 52 µl.

High fidelity PCR amplification of P1 and P2 adapter-ligated DNA fragments for RAD tags enrichment was carried out, prior to being hybridized to an Illumina Genome Analyzer flow cell. In a thin-walled PCR tube, 10.8 µl dH2O, 12.5 µl Phusion High-Fidelity Master Mix, 0.7 µl Solexa primer mix (10 µ*M*), and 0.5 µl RAD library template (eluate from last step) were combined. 18 cycles of amplification in thermal cycler were carried out: 30 sec at 98^0C, 18X [10 sec at 98^0C, 30 sec at 65^0C, 30 sec at 72^0C], 5 min at 72^0C, hold at 4^0C. 5.0 µl PCR product in 1X Orange Loading Dye was run out on 1.5% agarose gel next to 1.0 µl RAD library template and 2µl GeneRuler 100 bp DNA Ladder. If the amplified product is at least twice as bright as the template, a larger volume amplification (typically 100-250 µl) was performed to retrieve a large amount of the RAD tag library from the final gel extraction in the protocol. If amplification looked poor, more library template was used in a second test PCR reaction. The template should be dim, yet visible on the gel.

Large volumes of reaction mixture were purified with a MinElute column, and eluted in 23 µl EB. This purification step was performed to eliminate any contaminant bands that may appear due to an improper ratio of P1 adapter to restriction-site compatible ends. The entire sample, in 1X Orange Loading Dye, was loaded on a 1.1% agarose, 0.5X TAE gel and run for 5 min at 40 V, then 5

min at 60 V and 50 min at 100 V, next to 2.0µl GeneRuler 100 bp DNA Ladder Plus for size reference. PCR amplification of a wide range of fragment sizes often results in biased representation of amplified products with an increased number of short fragments. The library was cleaned through a column and gel purified. Being careful to exclude any free adapters or P1 dimer contaminants running at ~130 bp and below, a fresh razor blade was used to cut a slice of the gel spanning 350-650 bp. DNA was extracted using MinElute Gel Purification Kit following the manufacturer's instructions. Agarose gel slices were melted in the supplied buffer at room temperature and eluted in 20 µl EB.

All libraries were sequenced on the Paired-end module of the Illumina Genome Analyzer II following the manufacturer's instructions. Equimolar amounts of libraries 1 – 4 were combined and sequenced on a single lane of the Illumina HiSeq 2000 platform (100 bases, paired end reads) at the Genepool Genomic Facility, University of Edinburgh. Library 5 was sequenced in house at University of Stirling using two runs of Illumina Miseq (v3 chemistry, 100 forward/75 reverse reads).

2.2.7 Double Digested (dd) RADseq Library Preparation and Sequencing

Double digested RAD library preparation

Double digested RAD libraries were constructed from 10 tilapia species (4-13 individuals of each species/subspecies/population). RAD libraries were prepared according to Palaiokostas et al. (2015), modified from Peterson et al. (2012). High quality genomic DNA with a concentration of approximately 7 ng/µl

based on fluorometry was digested using enzymes restriction *Sbf*I (recognizing the CCTGCA⏐GG) and *Sph*I (recognizing the GCATG⏐C motif). In a 96 well plate format a 6 µl reaction volume containing 3 µl (21ng) DNA, 0.6 µl 10x CutSmart Buffer, 0.010 µl 10 units (U)/µg *Sbf*I, 0.010 µl 10 units (U)/µg *Sph*I and 2.380 µl double distilled water (ddH_2O) was mixed well, incubated for 40 min at 37 ^{0}C, then cooled to room temperature. A 3µl aliquot of barcode mix (*Sbf*I:*Sph*I 1:10), an individual specific combination of P1 (25 nM) and P2 adapter (100 nM), each with unique 5 or 7 bp barcode, was ligated to the digested DNA by adding 0.3µl 1x CutSmart Buffer, 0.120 µl rATP (1mM), 0.020 T4 Ligase 2 K ceU/ug and 2.560 ddH_2O, mixing well and incubated in thermocycler (heated lid off) for 2 hours 30 min at 22^{0}C. The ligation reactions were heat inactivated by incubating at 65^{0}C for 20 minutes in thermocycler (heated lid set to 70^{0}C and briefly spin down the plates).

Sample Pooling, Cleaning and Fragment Size selection

Recombined all 12 µl from each samples into 5.1 ml Qiagen PB buffer (approximately 3x volume of samples). 3 µl of 3 M sodium acetate (NaAc) (pH 5.2) was added to the template in the column and centrifuged. The template was eluted in 2x of 65 µl of EB buffer, to obtain 125µl total. The DNA concentration resulting from this purification was 17.72 ng/µl (Qubit 2.0 analysis). The DNA was extracted from the gel to obtain the appropriate size range of the fragments. A single well 2.5 cm wide was made to hold >200 µl of template. A 1.1% agarose gel without EtBr (0.42g agarose in 38 ml 0.5x TAE) was prepared in a small Biorad gel tray (c. 10 x 12 cm) with small tooted comb (wells 2-3 mm wide),

stored overnight in 0.5x TAE buffer in the fridge (4^0C). The apparatus was on ice, then it was frozen overnight.

To specify the target range for selection, 2x marker reactions were prepared containing 2 µl marker 4 (590bp), 2 µl marker 5 (320bp), 1.8 µl 6x LD and 6.2 µl of ddH_2O, then loaded next to the template, one on the left and another one on the right side of the template. 25 µl of 6x LD was added to the 125µl eluate, then loaded (approximately 175 µl in total). The 1.1% agarose gel was run in 3x TAE buffer at 40V for 5 minutes, then 5 minutes each at 60 V, 80 V and 100 V until the base pairs band (BPB) distance was about 3.3 cm from the origin (Figure 2.1). A smear of DNA approximately 300–600 bp was cutted with a clean razor blade, then it was extracted using the MinElute Gel Extraction Kit (Qiagen). The band was cut out the central part of gel, deliberately leaving the edges (in case of edge drag). The rest of the gel was stained, while the fragment was marked approximately 7 mm wide. Initially, 2x 0.375g of agarose gel slices were placed in two differents Eppendorf tubes, and processed with 1 column. A volume of 1.2 ml QG buffer was added to each tube, then 0.38 ml isopropanol was added. It was eluted in 2x 35 µl of EB buffer.

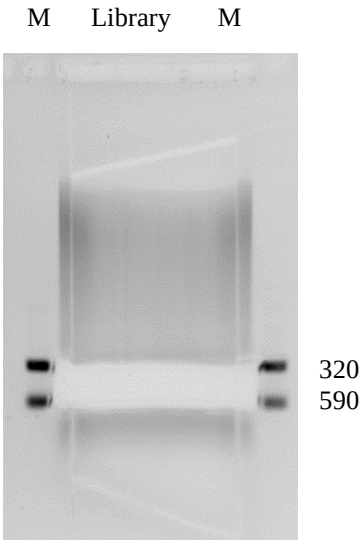


Figure 2.1 Size selection of PCR product from agarose gel.
M : marker

PCR amplification

Two differents PCRs were conducted at the same time: a standard 1µl template per 25 µl plus a non-template control (NTC) with 16 cycles and double template (2 µl per 25 µl) with 13 cycles were performed in a half reaction (12.5 µl) for the first test. In a thin-walled PCR tube, 5.05 µl ddH2O was combined with 6.25 µl master mix of NEB Q5, 0.2 µl Solexa primer mix (10 µ*M*), 0.5 µl ddRAD library template (for 16 cycles), and 1µl ddRAD library template (for 13 cycles). The amplification was performed as follows: 30 sec at 98^0C, 16X [10 sec at 98^0C, 30 sec at 65^0C, 30 sec at 72^0C], 5 min at 72^0C, hold at 4^0C. Then 5.0 µl of each PCR product plus 1X Orange Loading Dye was loaded onto a 1.5% agarose gel, next to 1.0 µl RAD library template and 2µl GeneRuler 100 bp DNA Ladder and run in 0.5x TAE buffer at 75 V. The gel was captured in the Syngene gel documentation with 40 ms exposure (Figure 2.2).

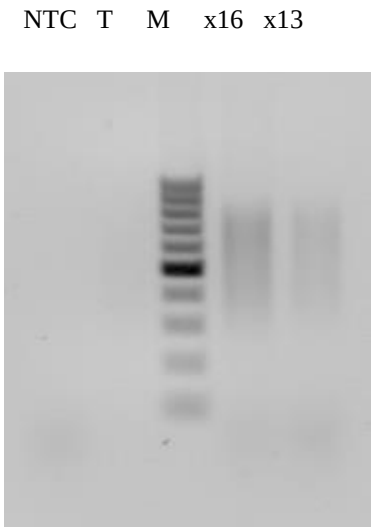


Figure 2.2 The first amplification of library template with two different cycles. NTC: non template control, T: template, M : 100 bp marker

As shown in Figure 2.2, the amplification of library template was more than four times as bright as the template, so it was decided to conduct bulk prep (400 µl) with 2 µl library template and 13 cycles of amplification. A large volume of amplification was conducted in 32 PCR tubes with 12.5µl reactions volume containing 5.05 µl ddH2O, 6.25 µl master mix of NEB Q5, 0.2 µl Solexa primer mix (10 µ*M*), and 1µl ddRAD library template. In the post PCR, all 32 aliquots were recombined and stored overnight at 4°C. 5 µl of library template was checked on a 1.5% agarose gel to make sure that the product was consistently dim, yet visible on the gel (Figure 2.3).

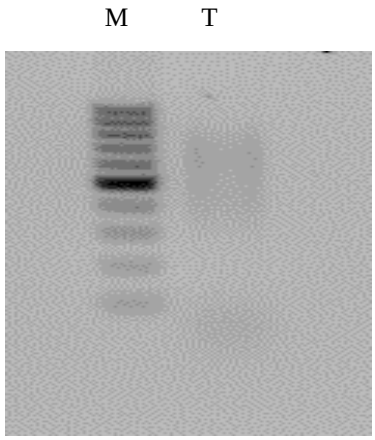
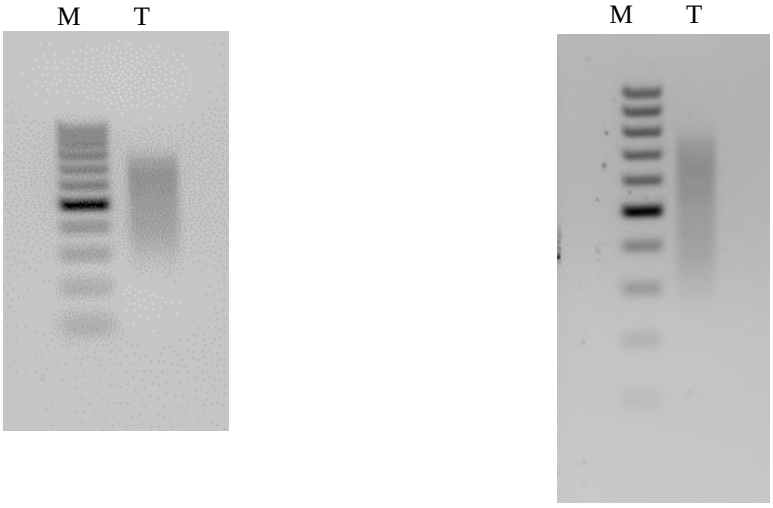


Figure 2.3 Large volume of amplification with 2 µl library template and 13 cycles, M: 100 bp marker, T:Template

A purification step was performed to eliminate any contaminant bands that may appear due to an improper ratio of P1 adapter to restriction-site compatible ends. The procedure was as previously described with a standard MinElute clean up followed by a paramagnetic bead clean. 1,350 µl (3x) of PB buffer was added to 450 µl of bulk prep library template. The template was eluted in 30 µl plus 25 µl of EB buffer, to obtain 52 µl returned. Final purification was conducted using AMPURE clean. An equal amount of AMPURE beads (52 µl) was added to library template. 2x volume of 75% EtOH was added for washing, and final elution was done in 23 µl EB buffer which resulted in 19 µl of product. For final library quality control, 1 µl template was loaded and run in the gel electrophoresis on fresh 1.5% gel, 0.5x TAE buffer. The image was captured two times, a short run with 40 ms and a long run with 80 ms (Figure 2.4).



a. Short run 0 ms b. Long run 80 ms

Figure 2.4 Final library quality control on fresh 1.5% agarose gel. Minimum size of the band 300 bp, maximum size 760 bp, mean 530 bp and median 550 bp.

The quantity of DNA concentration from purified product of library templates were measured in QUBIT® 2.0 Fluorometer:

Two readings were produced: the concentration of library template from two aliquot were 90 ng/ml and 87.2 ng/ml (mean = 88.6 ng/ml), while the concentration of two aliquot PCR samples were 4.78 ng/ml and 4.25 ng/ml (mean 4.52 ng/ml). Each concentration was multiplied 200x in actual condition, so on average, it gave 17.72 ng/µl for the library and 0.903 ng/µl for PCR template (Table 2.3)

Table 2.3 DNA content of purified product of library templates.

No	Type of measurement	Initial template	Overall library
1	Mol Wt of one bp	660	660
2	Median Size	525 bp	550 bp
3	Concentration	0.903 ng/uL	17.72 ng/uL
4	Volume	33 uL	22 uL
a.	Yield (ng)	29.799	389.84
b.	nmole DNA	0.00009	0.00107
c.	Molarity (nM)	2.60606	48.8154
d.	Total nmoles	660	0.00107
e.	Template nmoles		0.00009
f.	PCR nmoles		0.00099
g.	% template		0.08008
h.	% amplicon nmoles		0.91992

Note :

a. The yield (ng) = concentration (ng/µl) x volume (µl)
b. Nmole DNA = yield (ng)/(Mol weight 1 bp x Median size)
c. Molarity (nM) = (nmole DNA x 1000000)/volume (µl)

91.99% of amplicon was available for MiSeq. Therefore, 10 nM library needed a final concentration 10 nM/91.99% = 10.87 nM. So, for 10.87 nM solution that was equivalent to 10 nM available, the following volumes were added:

No	Reaction	Volume (µl)
1	Library	19.00
2	EB buffer	57.79
3	Tween 10, 1%	8.53
	Total volume	85.32

Note :

Volume EB buffer = (Molarity (nM)/Final concentration x vol. library) – vol. library – 1% tween 20.

1% Tween 20 = (Library molarity (nM)/Final concentration) x volume library/10

The final library solution was frozen at -20^{o}C and was ready for sequencing preparation.

Sequencing:

The following chemicals were required during sequencing preparation: HT1 (Hybridization Buffer), thawed and pre-chilled (Illumina-supplied provided in the MiSeq Reagent Kit), Illumina PhiX Control, stock 1.0 N NaOH (molecular biology-grade) and Tris-Cl 10 mM, pH 8.5 with 0.1% Tween 20.

Library preparation for Miseq sequencing consisted of three main steps:

1.Hybridization Buffer (HT1) and Fresh Dilution of NaOH preparation.

The tube of HT1 is used to dilute denatured libraries before loading libraries onto the reagent cartridge for sequencing. The tube of HT1 was removed from -20°C storage and set aside at room temperature to thaw, then stored at 2°C to 8°C until ready to dilute denatured libraries. Freshly diluted NaOH for denaturing libraries

for cluster generation is essential to the denaturation process. To avoid small pipetting errors from affecting the final NaOH concentration, at least 1 ml of freshly diluted 0.2 N NaOH was prepared by combining 800 µl of ddH_2O and 200 µl of 1.0 N NaOH, then the tube was inverted several times to mix.

2. DNA Denaturation and Dilution

It is important that the concentration of NaOH is equal to 0.2 N in the denaturation solution and not more than 1 mM in the final solution after diluting with HT1. As recommended in the protocol, either v3 reagents or v2 reagents were used in 4 nM library denaturation. This denaturation requires a 4 nM library, supports high library concentrations (10–20 pM), and results in a 20 pM DNA solution in 1 mM NaOH. The calculation was showed in the following table:

No	Reactions	Library	phiX
1	'10' nM Library	2	1
2	Water	3	1.5
	Mix = 4 nM		
3	0.2M NaOH (fresh)	5	2.5
	Incubate 5 min RT		
4	HT1 buffer chilled	990	495
	The reactions were mixed for 20 pM stocks		

Note:

a. For tilapia library = Final library concentration/20 x 600
b. For phiX library =Total volume of phiX library x Final library concentration/20.

To obtain 4 nM DNA library, 3 µl of ddH_2O was added to 2 µl of DNA library. Then, it was combined with 5 µl of freshly diluted 0.2 N NaOH in a microcentrifuge tube. The remaining dilution of 0.2 N NaOH was discarded or set aside to prepare a Phix control (within the next 12 hours). A phix control was

prepared as a control for illumina sequencing runs. The sample was mixed properly by vortex, and centrifuged at 280 xg for 1 minute. The microcentrifuge tube was incubated for 5 minutes to denature the DNA into single strands, then 990 µl of Pre-chilled HT1 was added to make a 20 pM denatured library in 1 mM NaOH. The denatured DNA was placed on ice until it was ready to proceed to final dilution.

3. Dilution of Denatured DNA for 4 nM Library

In the final library, 9.5 pM of denatured DNA was required, therefore the 20 pM denatured DNA was diluted to the desired concentration using the following guidance:

Final concentration	6 pM	8 pM	10 pM	12 pM	15 pM	20 pM
20 pM denatured DNA	180 µl	240 µl	300 µl	360 µl	450 µl	600 µl
Pre-chilled HT1	420 µl	360 µl	300 µl	240 µl	150 µl	0 µl

In order to make 9.5 pM stocks of each, the following reactions were prepared:

	Lib	HT1	Total
For tilapia library	285	315	600
For phiX library	14.25	15.75	30

To obtain the concentration required in the tilapia library, 285 µl of denatured DNA was added to 315 µl of HT1, then it was inverted several times to mix, then pulse centrifuged. Next, in the phiX library, 14.25 µl library was added to 15.75 HT1. The denatured and diluted DNA was placed on ice until it was ready to be loaded onto the MiSeq reagent cartridge.

4. PhiX Control Preparation

The 10 nM PhiX library was diluted to 20 pM using the v2 kit, then further diluted to 12.5 pM. 1 µl of 10 nM PhiX library was added to 1.5 ddH_2O then combined to 2.5 µl of 10 mM Tris-Cl, pH 8.5 with 0.1% Tween 20. PhiX control was denatured with combining 2.5 µl of 4 nM PhiX library and 2.5 µl of 0.2 N NaOH, then mixed with briefly vortex the 2 nM PhiX library solution. The template solution was centrifuged to 280 x g for 1 minute, then it was incubated for 5 minutes at room temperature to denature the PhiX library into single strands.

To obtain 20 pM PhiX library, 10 µl of denatured PhiX library was added to 990 µl pre-chilled HT1. The denatured 20 pM PhiX library could be stored for up to 3 weeks at -15^{o} - 25^{o}C.

Sequencing

The libraries were sequenced in three different lanes in house at University of Stirling using two runs of Illumina Miseq (v3 chemistry, 100 forward / 75 reverse reads).

3. DEVELOPMENT OF SPECIES-SPECIFIC SNP MARKERS FROM THE ADENOSINE DEAMINASE (*ADA*) GENE

3.1 Abstract

Identification of tilapia species, hybrids and introgressed populations is of importance in aquaculture and in wild populations where introductions have occurred. Although Tilapia species are morphologically similar, they are distinguishable, however hybrids are problematic. Molecular genetic variation for one enzymes, Adenosine deaminase (*ADA*), is a potential candidate marker for distinguishing tilapia species, since this locus is highly polymorphic among species at the protein (allozyme) level, and has been used for species identification. Therefore, we set out to identify species-specific nuclear DNA markers (single nucleotide polymorphisms, SNP) using sequencing of the coding regions of the *ADA* gene. The mRNA of liver tissue from 5 different tilapia species (2-5 individuals per species) were extracted, reverse transcribed, and sequenced to obtain coding sequences. The results indicated that *ADA* sequences were polymorphic in the species *O. niloticus*, *O. aureus*, *O. mossambicus*, *O. karongae* and *T. zillii*. *Tilapia zillii* was the most genetically distant from the other species, while *O. niloticus* showed the highest polymorphism within species. Primers for ten identified SNPs were then designed into SNP assays using the KBioscience Competitive Allele-Specific PCR (KASP-PCR) genotyping system (KBioscience Ltd, UK). SNP development using *ADA* was partially successful, where four out of ten SNP markers derived from *ADA* sequences, for *T. zillii* (Tzil_3_M170), *O. aureus* (Oaur_3_R122, Oaur_7_R626) and *O. mossambicus* (Omoss_10_Y879), can be applied in identifying and discriminating among tilapia species.

Key words: SNP marker, Tilapia, *ADA*, KASP assay

3.2 Introduction

A range of different types of genetic markers, both type I and type II have been used to assess genetic variation and apply this to further understanding and management of wild and cultured species and populations. One category of type I marker is allozyme. This have been used for species identification in brown trout (*Salmo trutta* L.) (Allendorf et al., 1977), allis shad (*Alosa alosa* L.) and twaite shad (*Alosa fallax* L.) (Alexandrino et al., 1993). Allozymes were also used in analyzing the geographic range of the tropical shad, hilsa *Tenualosa ilisha* indicated substantial gene flow between groups of hilsa within The Bay of Bengal (Salini et al., 2004), distinguishing hatchery stocks and native population of indigenous brown trout (*Salmo trutta* L.) population in Spain (Cagigas et al., 1999) and evaluate the genetic variability of Apennine stream populations (Northern and Central Italy) in *Salmo* (*trutta*) *macrostigma* (Marzano et al., 2003). In the tilapia, allozyme has been used for species identification (McAndrew & Majumdar, 1983; Sodsuk et al., 1995). One study indicated that 26 loci were polymorphic and 12 were diagnostic between *T. zillii* and *O. niloticus* (Rognon et al., 1996). Furthermore, two loci *LDH*-I and *PGI*-2 were also diagnostic between *T. zillii* and *T. guineensis* and can be used to identify hybrids in the River Bia Basin (Adépo-Gourène et al., 2006).

3.2.1 Recent study in adenosine deaminase

The nucleoside adenosine is a molecule that plays several roles in different tissues. In the central nervous system (CNS), adenosine acts as a neuromodulator,

controlling the excitatory and inhibitory synapses (Fredholm et al., 2005). It contributes nearly 50% to insulin-stimulated muscle glucose transport by activating the A1 Adenosine Receptor (Thong et al., 2007). Adenosine deaminase (ADA, EC 3.5.4.4) is an important enzyme that promotes the irreversible hydrolytic deamination of adenosine and 2'-deoxyadenosine to inosine and 2'-deoxynosine, respectively (Rosemberg et al., 2007). It is an indispensable enzyme in purine metabolism that affects the methylation process, cell growth and differentiation, apoptosis, DNA replication and immune system (Dong et al., 1996). It is required for B and T-cell development and plays a central role in the maintenance of a competent immune system (Cristalli et al., 2001).

ADA has been found in a wide variety of microorganisms, plants, invertebrates and mammals. Genetic deficiency of ADA human results a disease known as severe combined immunodeficiency disease (SCID), which is characterized by a lack of T- and B-lymphocytes (Kaneijam et al., 1993). This disease is caused by a mutation in the gene coding for the blood enzyme adenosine deaminase (ADA), as a result the precursor cells that give rise to one of the cell types of the immune system are missing (Griffiths et al., 2002). The polymorphism of the *ADA* gene (20q13.11) in the human resulting from the substitution of G by A at nucleotide 22 of exon 1 replaces the Asp amino acid (*ADA*1* allele) with Asn (*ADA*2* allele) amino acid in position 8 of the enzyme. Consequently, individuals with the *ADA*2* allele express low levels of ADA compared activity to homozygous *ADA*1/*1* individuals (Hirschhorn et al., 1994).

Recent research on rats infected with *Trypanosoma evansi* showed that the enzyme ADA plays an important role in the production and differentiation of

blood cells, erythrocytes and lymphocytes (Da Silva et al., 2011). There are also a group of proteins having similarity to ADA, the adenosine deaminase related growth factors (ADGF; known as CECR1 in vertebrates), and a novel paralogue, ADA-like (ADA-L) was also discovered which having significant amino acid similarity to ADA. These two domains are located just upstream of two ADA catalytic residues, of which all eight are conserved among the ADGF and ADA-L proteins that indicated both of them may share the same catalytic function as ADA (Maier et al., 2005).

3.2.2 Application of ADA enzyme for species discrimination in fish

Previously, isozymes have been used for species determination in rainbow trout (*Salmo trutta* L), where 37 enzymes were used to investigate genetic variation. Of a total of 69 loci detected, 54 loci were considered usable in population genetics screenings. Many enzymes, such as glutamate pyruvate transaminase (GPT), nucleoside phosporhylase (NP), pyruvate kinase (PK), phosphoglycerate kinase (PGK) including ADA enzyme could not resolve a clear genetic result (Allendorf et al., 1977). In Allis shad (Alosa alosa L.) and twaite shad (Alosa fallax L.), screening of allelic variation across eight allozyme loci (including ADA) and sequencing 448 bp of the mtDNA cytochrome b gene in 14 rivers throughout the range of the species supported that the two taxa were independent lineages (13% net nucleotide divergence) (Alexandrino et al., 1993). Adenosine deaminase (*ADA*), is a potential candidate marker for distinguishing tilapia species where this locus is highly polymorphic between 9 tilapia species with majority being unique and at least two alleles in every species (McAndrew &

Majumdar, 1983). Another study indicated that genetic variation in the ADA locus based on biochemical polymorphism using fin and muscle tissues can be designated to differentiate between genus *Sarotherodon* and *Oreochromis*, distinguishing between *O. andersonii*, *O. mortimeri* and *O. macrochir*, but not between chambo species *O. (Nyasalapia) karongae*, *O. lidole* and *O. (N.) squamipinnis* (Sodsuk et al., 1995).

The advantage of using ADA allozyme is due to its high variability among tilapia species, however fish must be killed and tissues such as fin, muscle, liver, eye, and heart need to be kept frozen until analysed (Toniato et al., 2010). Furthermore, mutation detection at the DNA level where replacement by a similarly charged amino acid, may not be detected by allozyme electrophoresis (Kucuktas & Liu, 2007).

3.2.3 SNP determination from allozymes

Many genes in natural populations are polymorphic, with two or more relatively frequent alleles; approximately 20 percent of enzyme genes are polymorphic in plants and vertebrates. The alternative forms of an enzyme coded by different alleles at a single gene are known as allozymes (Hartl & Jones, 2008). Allozyme polymorphism between species can be analysed to detect the responsible SNPs. SNP candidates responsible for the allozyme differences are identified by first distinguishing the non-synonymous SNPs responsible for the amino acid substitutions, then eliminating those substitutions that did not result in amino acid charge differences (Brunelli et al., 2008).

The allozymes have been used at the lactate dehydrogenase (*LDH-B2**)

and superoxide dismutase loci (*sSOD-1**) which distinguishing between species of the inland native populations of rainbow trout *Oncorhynchus mykiss gairdneri* and introductions of the widely cultured subspecies, *O. m. irideus* (Brunelli et al., 2008). However, one study characterized both coding and non-coding regions of lactate dehydrogenase-B (ldh-B) where this locus is highly conserved between two tropical perciformes, *L. calcarifer* and *L. niloticus*, with just 2.9% divergence of coding regions and five amino acid differences (Edmunds et al., 2009). Differences in the electrophoretic mobility of allozyme alleles on cellulose acetate gels are primarily due to charge variation, with minor effects of structural modifications affecting mobility (Barbadilla et al., 1996). Sequence variability in the gene and the percentage of variation in some species are displayed in Table 3.1.

Table 3.1 Variation of cDNA sequences in SNP determination from allozyme

No	Gene/Species	cDS Size (bp)/∑ aa	∑ Locus	∑ SNP	∑ aa substitution	% of aa substitution in SNP/gene
1	LDH-B2/ Rainbow trout[a)]	1,002/334	2 (LDH-B2*76, LDH-B2*100)	3	1	33.33/0.3
2	sSOD-1*/Rainbow trout[a)]	462/154	2 (sSOD-1*152, sSOD-1*100)	9	4	44.44 /2.60
3	LDH-C1[b)]	440(partial)/146.67	2 (LDH-C1*90, LDH-C1*100)	1	1	100/0.23
4	LDH-B/*L calcarifer*[c)]	1,005/335		29	5	17.24/1.49
5	LDH-B/*L niloticus*[c)]	1,005/335		29	5	17.24/1.49

Note:
[a)]Brunelli et al., 2008; [b)] Chat et al., 2008); [c)] Edmunds et al., 2009
Aa, amino acids; LDH, Lactate dehydrogenase; sSOD, Superoxide dismutase; PGI, Phosphoglucose isomerase.

3.2.4 Objectives of the study

The current work is focused on developing species-specific DNA markers for tilapia based on variation in a specific gene, adenosine deaminase (*ADA*). A key question that will be studied is how polymorphic sequences of *ADA* in tilapia and in particular Single Nucleotide Polymorphisms (SNPs) can be applied as a specific marker to distinguishing tilapia species. Initially, sequence variation in this gene was analysed through Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) and Sanger DNA sequencing. This allowed the design of SNP assays that can distinguish alleles varying between species.

The main objectives of this chapter were:

1. To look for species-specific markers that distinguish between tilapia species using the candidate gene *ADA*.
2. To develop such sequence differences into allelic discrimination assays.
3. To demonstrate the potential of the markers in hybridization and introgression studies in aquaculture and/or wild populations.

3.3 Materials and Methods

3.3.1. Tissue collection

Liver from individuals of five different tilapia species were collected and stored in RNA Later. After 24 hours, the samples were drained and kept in the freezer at -20°C until they were used in RNA extraction. The five species, consisting of *O. niloticus* (n=5), *O. mossambicus* (n=2), *O. karongae* (n=2), *O.*

aureus (n=2) and *T. zillii* (n=2), were originally from the Institute of Aquaculture live collection. The list of chemicals that were used in the RNA extraction, reverse transcription and polymerase chain reaction can be seen in Appendix III-1.

The five afore mentioned species and another five other species of tilapia (*O. urolepis, S. melanotheron, S. galilaeus, O. andersonii* and *O. macrochir*) (Table 2.1) and some populations of *O. niloticus* and/or hybrids were also used as a case study (Appendix III-2). In order to get fin samples for the SNP assay, all fish collected from Stirling were anesthetized using Ethyl 4-aminobenzoate (Benzocaine) at a concentration of 1:10,000, then fish were killed under schedule I using brain destruction. Fin samples were placed in 1.5 ml screw cap tubes containing 100% ethanol and kept in cold storage at 4°C until they were used in DNA extraction. DNA was extracted based on Realpure Genomic DNA Extraction Kit with some modifications in chemical reaction volume, centrifugation time and ethanol washing (See Chapter 2).

3.3.2. RNA extraction, Reverse Transcription, Amplification and Purification

RNA was extracted using TRI Reagent/Trizol and BCP (1-bromo-3-chloropropane). It produces very pure, high molecular weight RNA. Initially, liver was cut and weighed. Approximately 50 mg was placed in a 1.5 ml screwcap tube containing 1 ml of TRI Reagent. The samples were homogenized using a Mini-Beadbeater for 40 seconds until the tissue was significantly disrupted. Next, the homogenized samples were removed to a new flip cap tube and incubated at room temperature (RT) for 5 minutes. 100 µl of BCP (per ml TRI Reagent) was added and the tube was shaken vigorously by hand for 15 seconds, incubated at RT for

15 min, then centrifuged at 20,000 g for 15 min at 4°C. The aqueous (upper) phase was transferred to a new tube using a wide-bore pipette tip (as much as 2x150 µl). To precipitate the RNA, ½ volume (per aquaeous phase volume) of RNA precipitation solution and ½ volume of isopropanol were added to the tube, gently inverted 4-6 times and incubated for 10 min at RT. Next, samples were centrifuged at 20,000 x g for 10 min at 4°C. The supernatant was removed by pipetting and the pellet was washed in 1 ml of 75% ethanol for 15 min at RT. Then, the tube was flicked to detach the pellet, inverted a few times and centrifuged at 20,000 x g for 5 min at RT. Most of the supernatant was removed carefully with a pipette (1 ml), the samples were centrifuged again briefly (2s pulse) and all remaining ethanol was removed with a small volume pipette (20 µl). The RNA pellet was dried at RT for 3-5 min, until all visible traces of ethanol were gone, then resuspended in an appropriate amount of RNAse free water (i.e. 20-50 µl). The samples were incubated at RT for 30-60 minutes with gentle flicking of the tubes every 10 minutes to aid resuspension, then placed on ice for 30 min.

High quality RNA is required for optimal reverse transcription reaction to obtain cDNA. The quantity of RNA was determined using Nanodrop. The purity of RNA was assessed using the Absorbance ratio A 260/A280, which should be approximately 2.1. The Absorbance ratio A260/A230 should be very close to 2.0 to make sure there is no contamination from proteins, or chaotropic (a molecule in water solution that can disrupt the hydrogen bonding network between water molecules) salts like guanidinium isothiocyanate and phenol. To check the integrity, an aliquot of RNA (~300-500 ng) was heated with the loading buffer for 5 min at 75°C, chilled briefly on ice and run on a 1.2% agarose gel along with a

molecular marker standard from λ*Hind*III. A good intact RNA will produce sharp, clear 28S and 18S rRNA bands on a denaturing gel.

Reverse Transcription

Complementary DNA (cDNA) was synthesized from purified and concentrated RNA using High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Paisley, U.K.), following manufacture's instructions, but using a mixture of random primers (1.5 µl as supplied) and anchored oligo-dT (0.5 µl at 400 ng/µl, Eurofins MWG Operon, Ebersberg, Germany). RNA samples were adjusted to 200 ng/µl, mixed and heated up in 70°C for 5 minutes prior to cDNA synthesis. A 20 µl total reaction volume was made with 10 µl RNA, 2 µl 10x RT buffer, 0.8 µl 25x dNTP Mix (100 mM), 1.5 µl 10x RT random hexamer primers, 1.0 µl MultiScribeTM reverse transcriptase (50 U/µl), 0.5 µl oligo dT primers and 4.2 µl nuclease-free water. Synthesis was carried out in ABI 9700 Thermocycler (Applied Biosystems, Foster City, CA) and reaction conditions were 25°C for 10 min, 37°C for 120 min, and 85°C for 5 min. cDNA samples were stored at -20°C, or 4°C for immediate PCR.

Amplification of cDNA product using PCR (Polymerase Chain Reaction)

Amplification of the *ADA* gene from tilapia liver cDNA was accomplished with specific primers by aligning the *ADA* coding sequences from several fish species from the National Center for Biotechnology Infromation (NCBI) sequence database (GeneBank). The following species sequences were chosen to identify conserved regions: (*Danio rerio*, *Ictalurus furcatus*, *Gasterosteus aculeatus* and *Oreochromis niloticus)*. The sequences were aligned using the ClustalX program, however the specific primers were designed based on *O. niloticus* using primer design from DNASTAR. The primers were as follows: Forward primer (*ADA*_ON_For) GGCCGATCGCTCTTCTG (17mer), reverse primer 1 (*ADA*_ON_Rev1) CAGTGCTCTGGATCATCTC (19mer) and reverse primer 2 (*ADA*_ON_Rev2) CGTACCGCTTCCTCATAG (18mer) (Figure 3.1). Two reverse primers were design to reach the length size of sequencing products. In the end, only reverse primer 1 was used because it amplified almost the entire cDNA.

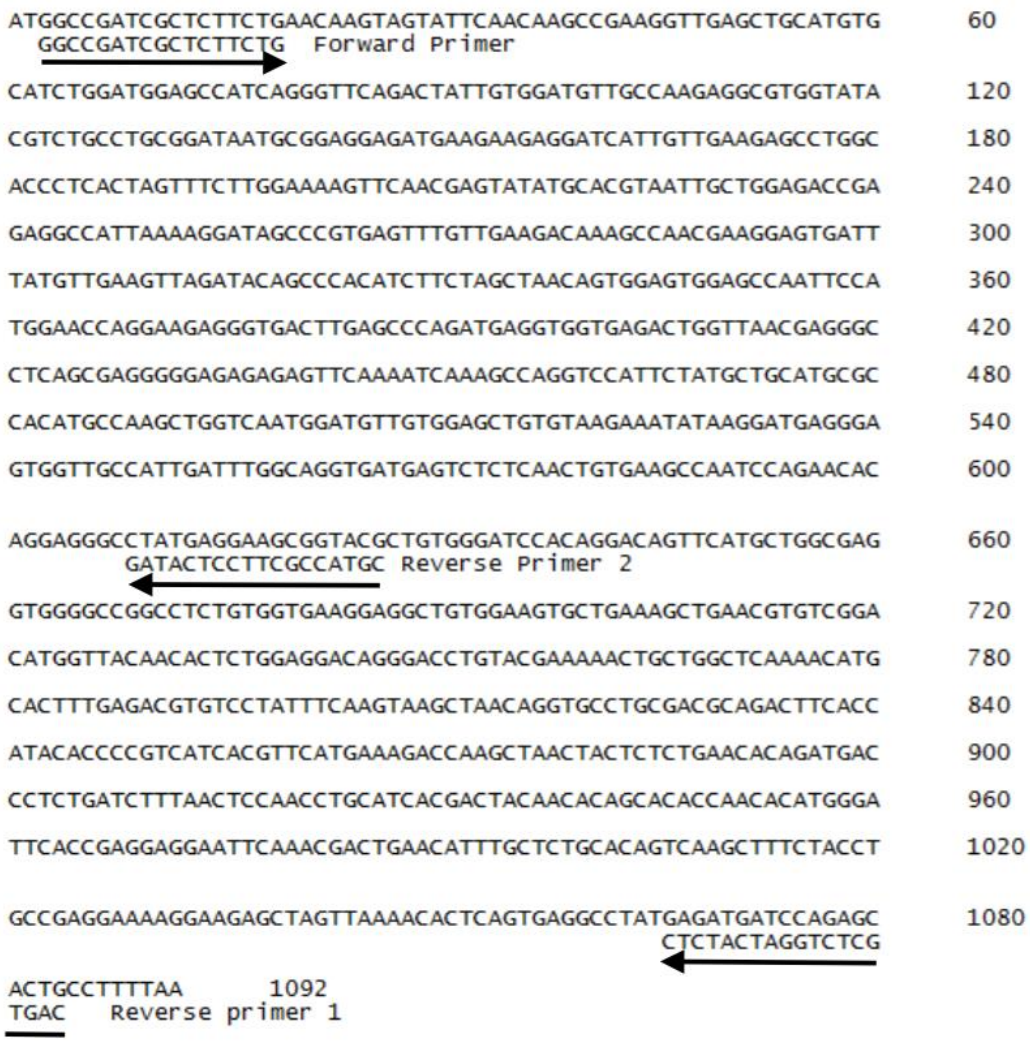


Figure 3.1 Forward and reverse primer position in the *ADA*-like sequences of reference from *O. niloticus*

PCR was performed in 10 µl (final volume) containing 5 µl My Taq (buffer, $MgCl_2$, Taq polymerase mix), 0.7 µl primer mix, 3.3 µl nuclease free water, and 1 µl cDNA template. The amplification conditions were; initial cycle of 95°C for 1 min followed by 32 cycles of 95°C for 15 sec (denaturation), 56°C for 15 sec (annealing), 72°C for 35 sec, final extension at 72°C for 2 min and finally cool down to 15°C. PCR products were checked for expected size on 1.2% agarose gels by loading 2 µl of the PCR reaction and 7 µl of 1x loading dye along

with molecular marker of 1 kb ladder. Gel runs were carried out at 75 volts for 35-40 min.

Purification of PCR Products

PCR products were purified using the Qiagen purification kit protocol (See Chapter 2.)

3.3.3. Sequencing of PCR products

A 5 µl aliquot of purified PCR product from each sample and 5 µl primer (5 pmol/ µl) were sent to GATC Biotech for sequencing. The concentration of PCR product was approximately 20-80 ng/ µl.

3.3.4. Sequences Data Analysis

The coding sequences were aligned in SeqMan software from DNA star. The sequences were analyzed using DNA star to determine the quality of sequences and the possibility of gaps and to compare to the reference sequence predicted *ADA*-like from *O. niloticus* (XM_003457049.1 LOC100705718). The nucleotide sequence was translated to the amino acid sequence using software from http://web.expasy.org/translate. Sequence analysis was performed by a Blast search on the sequence database http://www.ncbi.nlm.nih.gov to compare the similarity of the *ADA*-like gene between the tilapiine species to those in the GenBank database. Each sequence from each individual was put together in the alignment file from Mega5 software. The SNPs were determined by aligning the sequence and looking for nucleotide variations between species. Furthermore, the variation both in nucleotide and amino acid sequence was compared within and

between species. Molecular weight and isoelectric point values were calculated based on nucleotide sequence by free online resource in http://web.expasy.org/compute_pi/.

3.3.5. SNP Assay Design

The exon position in *ADA* was determined by alignment to the coding sequences of the reference *ADA* genome *O. niloticus* using Artemis software, a step to determine exon-intron position. They were picked up and saved in fasta format, thus the exon position can be mapped among tilapia species. A SNP was determined based on base substitutions between the tilapia species. The SNP assays then were developed by designing the specific primer by picking 15 nucleotides before and after the SNPs location that distinguished between species. For the exon sequences that did not have enough nucleotide length before and after a SNP, the adjacent intron sequence was added and the primer was designed based on exon-intron boundaries (Table 3.2). The 10 SNP primers were then designed into SNP assays using the KBioscience Competitive Allele-Specific PCR (KASP-PCR) genotyping system (KBioscience UK Ltd, UK) for routine analysis of wild and and farmed tilapia.

Table 3.2 Sequence submission for designing SNP assay to LGC Genomic. All sequences are from exons apart from the underlined intronic sequences.

No	SNP ID	Sequence
1.	Oaur_3_R122	TTTTTGCCCTATGTTAGGAGGCGTGGTATA C[R]TCTGCCTGCGGATAATGCGGAGGAGA TGAAGCAGAGGATCATTGTTGAAGAGCCT GGCACCCTCACTAGTTTCTTGGAAAAGTT CAACGAGTATATGCACGTAATTGC
2.	Tzil_3_M170	GAGGCGTGGTATACCTCTGCCAGTGAATA CTGTGGAGGAGATGAAGCAGAGGATCATT GTTG[M]AGAGCCTGGCACCCTTACTAGTT

		TCTTGGAAAAGTTCAACGAGTATATGCATG TAGTTGC
3.	Okar2_6_S492	CTGTTTAGTAACACATCTGTTTTAAATTCT CCAGG[S]TGGTCAAKGGATGTTGTGGAGC TGTGTAAGAAATATAAGGATGAGGGAGTG GTTGCCATTGATTTGGCAGGTGATGAGTCT CTCAACTGTGAAGCCAATCCAGAACACAG GAGGGCCTATG
4.	Okar2_6_K500	CTGTTTAGTAACACATCTGTTTTAAATTCT CCAGGSTGGTCAA[K]GGATGTTGTGGAGC TGTGTAAGAAATATAAGGATGAGGGAGTG GTTGCCATTGATTTGGCAGGTGATGAGTCT CTCAACTGTGAAGCCAATCCAGAACACAG GAGGGCCTATG
5.	Tzil_6_Y580	GCTGGTCAATGGATGTTTTGGAGCTCTGTA AGAAATATAAGGAGAAGGGAGTGGTTGCC ATTGATTTGGCAGGTGATGAGTCTCTCAAC [Y]TTGAAGCCAGTCCTGAACACAAGAAG GCCTATG
6.	Okar2_8_M770	GCTGTGGAAGTGCTGAAAGCCGAACGTGT CGGACATGGTTACAACACTCTGGAGGACA GGGACCTGTACGAAAAACTGCTGG[M]TC AAAACATGCACTTTGAGGTAAAGAACTGT GG
7.	Oaur_7_R626	TGATTGATTTTGTTCCCAGGAAGCGGTAC[R]CTGTGGGATCCACAGGACAGTTCATGCT GGCGAGGTGGGGCCGGCCTCTGTGGTGA AGGAG
8.	Onil_8_Y756	GCTGTGGAAGTGCTGAAAGCTGAACGTGT CGGACATGGTTACAACACTCTGGAGGACA GGGACCTGTA[Y]GAAAAWCTGCTGGCTC AAAACATGCACTTTGAGGTAAAGA
9.	Onil_8_Y762	GCTGTGGAAGTGCTGAAAGCTGAACGTGT CGGACATGGTTACAACACTCTGGAGGACA GGGACCTGTAYGAAAA[W]CTGCTGGCTC AAAACATGCACTTTGAGGTAAAGA
10.	Omoss_10_Y879	TCTTTCGTGCAGGTTCATGAAAGACCAAG CTAA[Y]TACTCTCTGAACACAGATGACCC TCTGATCTTTAACTCCAACCTGCATCACGA CTACCACACAGCACGCCAACACATGGGAT TCACCGAGGAGGAATTCAAACGACTG

Note:
SNP ID contains four details: the name of species, exon location in the gene structure, the SNP type together with position of SNP in the gene.
The International Union of Pure and Applied Chemistry (IUPAC) code for nucleotide M: A or C; R: A or G; S: G or C ;K: G or A and Y: C or T.

3.3.6. Phylogenetic tree

Phylogenetic trees between tilapia species were produced based on the coding sequences (11 exons) of adenosine deaminase from 5 tilapia species. Additionally, *ADA* sequence from *Haplochromis burtoni* was also included in the tree reconstruction as a species outgroup. Blast software was used to retrieve the sequences from NCBI Genbank database. The sequences were picked up, standardized in length (1,092 bp) with all coding sequences from five tilapia species (1,076-1,092 bp) using alignment from Mega 5 software and saved in fasta format. All the sequences were aligned in Phylip format by Clustal Omega program, and the trees constructed using RAxML and visualized by FigTree software.

3.3.7. SNP assays

Ten SNP markers derived from *ADA* sequences were validated in five tilapia species (*O. niloticus*, *O. mossambicus*, *O. karongae*, *O. aureus* and *T. zillii*) and extended to five other species of tilapia (*O. urolepis*, *S. melanotheron*, *S. galilaeus*, *O. andersonii* and *O. macrochir*), where each species consisted of 1-3 different populations.

Each SNP derived from the *ADA* sequence was tested using KASP assay. PCR reaction was performed with a T-Gradient Thermoblock (Biometra GmbH, Germany) (Robinson & Holme, 2011). The final PCR 10µL reaction mixture consisted of 1µL DNA sample 45ng/µL, 3.86µL ddH2O, 5.0µL KASP master mix, and 0.14µL of KASP assay primer. DNA samples were arrayed into 96-wells PCR plates, using either dried or wet DNA (directly added to the plate after

loading PCR reactions, then used within 0.5 hrs or dried for future use). The dried DNA samples gave more consistent result in comparison to liquid DNA samples. The plates were sealed with a clear seal to avoid contamination between samples. Furthermore, NTCs (Non Template Controls) were included in each assay. Dried DNA can be stored in the freezer. If working with dried DNA, an additional 1.00µL of distilled water was added to the dried samples along with the other PCR reaction mixs to make up 10.00µL of PCR final volume. To re-suspend the dried DNA, these samples were left for approximately 30 minutes. Everything was kept on ice while dispensing and under sterile conditions to avoid any contamination. The chemicals in the KASP assay are light vulnerable; therefore the preparation was also done in low light conditions. The thermal cycling conditions were as follows: initial cycle of 94°C for 15 min (hot start enzyme activation) followed by 94°C for 20sec, touchdown over 65°C to 57°C for 60 sec (10 cycles dropping 0.8°C each cycle) and an extra 34 cycle at 94°C for 20 sec, and 57°C for 60 sec.

Genotyping data was analysed using Techne Quantica® Quantifiable Realtime PCR Thermal Cycler (Techne Cambridge Ltd, UK) and viewed graphically to know the allele discrimination for particular SNPs. KASP uses the fluorophores FAM and HEX for distinguishing genotypes. In KBioscience's software, the FAM and HEX data are plotted on the x- and y- axes, respectively. The passive reference dye ROX is also used to allow normalization of variation in signal caused by differences in well-to-well liquid volume (Figure 3.2). If the genotype at a given SNP is homozygous, only one or other of the possible fluorescent signals will be generated. If the individual is heterozygous, the result will be a mixed fluorescent signal. The name of assay, the allele variation and

number of individual tested in each species was recorded and the allele frequency calculated for each SNPs assay.

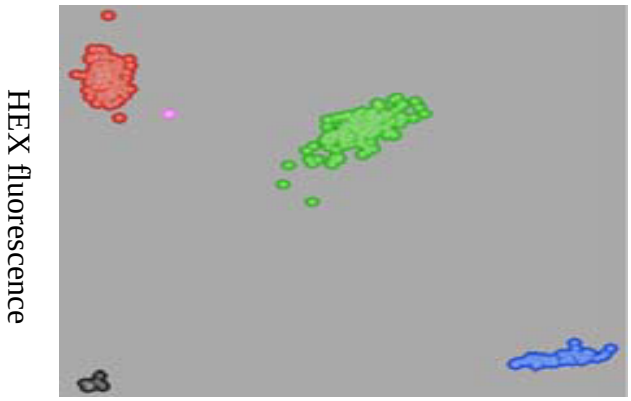


Figure 3.2 Genotyping data plotted using KBiosciences KlusterKaller software. Genotyped samples marked red are homozygous for the allele reported with HEX, those marked blue are homozygous for the FAM allele, those marked green are heterozygous, and those marked black are non template controls (NTC).

3.4 Results

3.4.1. Isolation, Reverse Transcription and Amplification

The results of RNA isolation from liver tissues of 5 different tilapia species indicated two prominents bands, 28S and 23S rRNA are shown in Figure 3.3. Sample number 3 indicated less-intact RNA, probably due to decreasing quality tissue sample during storage.

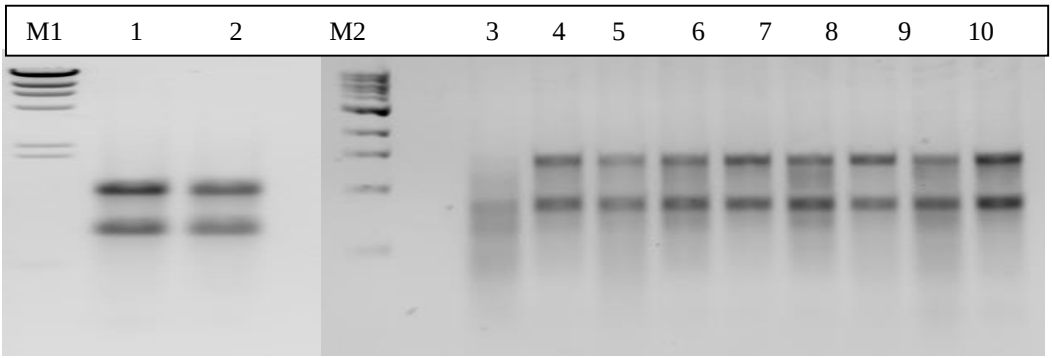


Figure 3.3 Gel electrophoresis of total RNA from 5 tilapia species. The gel was run in 1.2% agarose at 75 V for 45 min.
M1: Hind III ladder; M2: 1 kb ladder; 1: *O. niloticus* (liver); 2: *O. niloticus* (muscle), 3: *O. niloticus* 2, 4-5: *O. aureus*; 6-7: *O. karongae*; 8-9: *O. mossambicus* and 10-11: *T. zillii*. Samples 3-10 were from liver.

The concentration of purified complementary DNA (cDNA) produced from the RNA templates ranged from 11.9-56.5 ng/µl (Appendix 3.3). These values fulfilled the minimum criteria for sequencing by GATC Biotech Ltd., i.e. at least 10 ng/µl derived from PCR product. DNA quantification after purification from 10 samples ranged from 1.54-2.17 (OD 260/280) and 0.48-2.26 (OD 260/230). Samples from *O.aureus* 2 and *O. mossambicus* 1 had the lowest cDNA purity, albeit the sequencing results were still good for SNP calling.

3.4.2. Gene Characterization

The nucleotide Blast to the GeneBank database from NCBI reference genome showed that all the sequences of adenosine deaminase matched with the predicted adenosine deaminase-like gene of *O. niloticus* isolate 000638D3DF (accession NC_022218.1), consisting of 10 introns and 11 exons encoding 364 amino acids. Based on the reference, this gene was located in the linkage group LG20 with locus size 22,776 bp. The exon vary in size: the longest was number IV (144 bp), and the shortest was number I (45 bp). The two largest introns are number 5 (5,350 bp) and 1 (4,453 bp) respectively and the smallest was number 9 (105 bp) (Figure 3.4).

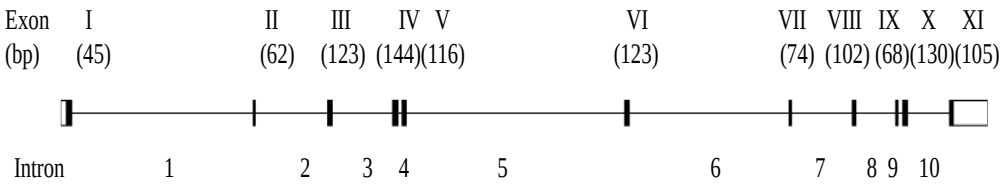


Figure 3.4 Gene structure of adenosine deaminase in tilapia. The vertical bars are the exons and the gaps between denote introns.

The total exon sequence length showed some polymorphisms between tilapia species in comparison to the predicted *O. niloticus ADA* sequence (ranged from 1,076 to 1,092 bp), but one of the sequences from *O. niloticus* was identical to the reference (Table 3.3). *T. zillii* was the most different from the reference, both for non-synonymous (3.11%) and synonymous (2.20%) changes, while *Oreochromis* spp. were generally more similar compared to the reference. The majority of

polymorphisms in *O. niloticus* were synonymous, while in other *Oreochromis* and *T. zillii* non-synonymous polymorphisms were more frequent than synonymous (Figure 3.5 & 3.6). The ratio dS/dN denoted that only *O. niloticus* had a value >1, while others species <1, with *O. karongae* showing the lowest ratio.

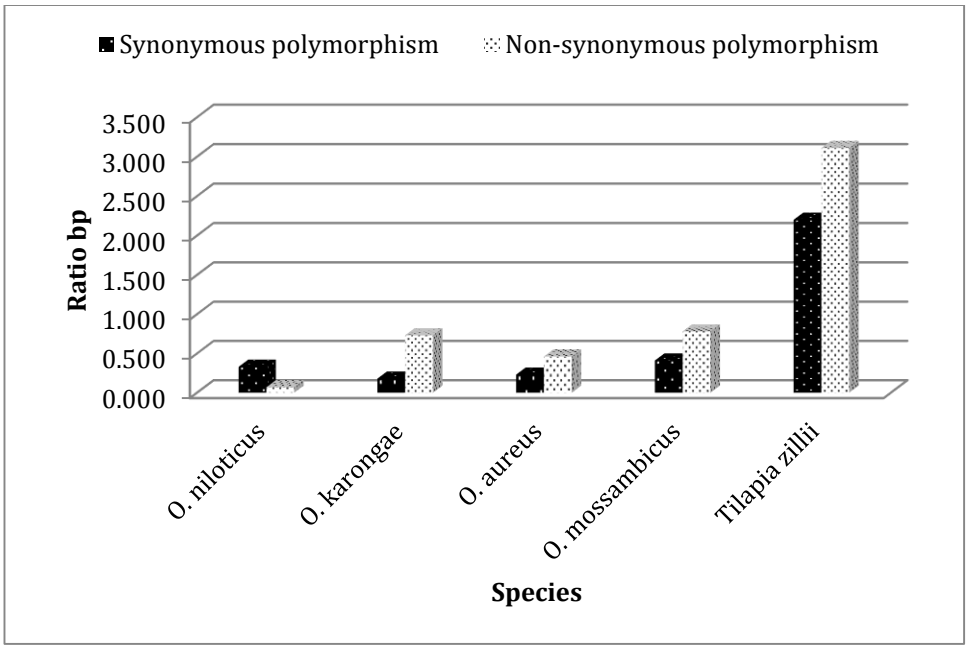


Figure 3.5 The average of polymorphism in *ADA* gene from five tilapia species.

Table 3.3 Synonymous (dS) and non-synonymous polymorphisms (dN) in exon regions of *ADA* sequences in tilapia species (ref = accession NC_022218.1).

No	Species	Synonymous polymorphism (dS) (% of bases)	Non-synonymous polymorphism (dN) (% of bases)	dS/dN ratio
1	*O. niloticus* 2	0.46	0.09	5
	O. niloticus 3	0.18	0.00	NA
	O. niloticus 4	0.55	0.18	3
	O. niloticus 5	0.46	0.09	5
	O. niloticus 6	0.00 (identical to reff.)	0.00 (identical to reff.)	Identical to reff
	Average	0.41	0.09	4.5
2	*O. karongae* 1	0.18	0.64	0.29
	O. karongae 2	0.18	0.82	0.22
	Average	0.18	0.73	0.25
3	*O. aureus* 1	0.27	0.46	0.60
	O. aureus 2	0.18	0.46	0.40
	Average	0.23	0.46	0.50
4	*O. mossambicus* 1	0.46	0.64	0.71
	O. mossambicus 2	0.37	0.92	0.40
	Average	0.41	0.78	0.53
5	*Tilapia zillii* 1	2.20	3.11	0.71
	Tilapia zillii 2	2.20	3.11	0.71
	Average	2.20	3.11	0.71

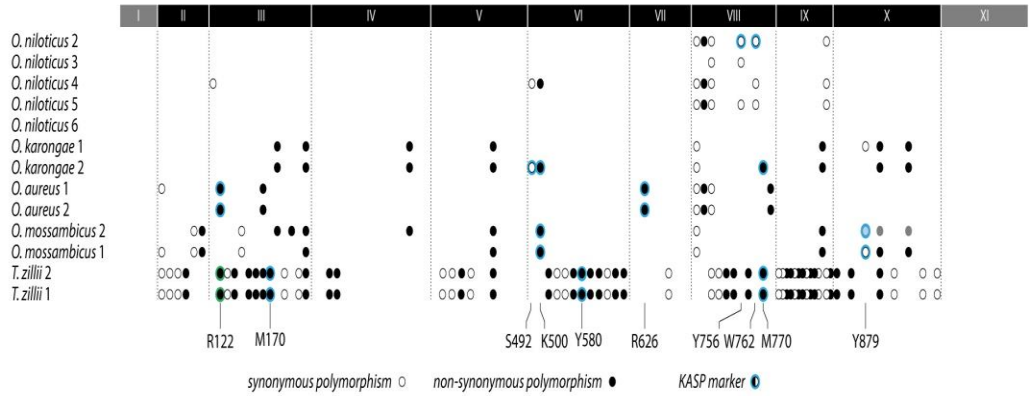


Figure 3.6 Mapping polymorphisms in *ADA* gene to reference genome of *O. niloticus.*

Numerous polymorphisms in the coding region of the adenosine deaminase gene can be seen in the Table 3.4. Non-synonymous polymorphisms impacted on side chain polarity and hydropathy index. For instance, nucleotide changed in position 106 resulted in a change from Lysine to Glutamic Acid, and as a result side chain polarity changed from basic polar to acidic polar with the side chain charge (pH 7.4) changing from positive to negative. Conversely, in nucleotide position 439, side chain polarity changed from negative to positive due to a change from glutamic acid to lysine. Some changes, for instance arginine to histidine did not impact on side chain polarity (both basic polar) and side chain charge (both are positive).

Table 3.4 Characteristics of synonymous and non-synonymous polymorphism in *ADA* sequences.

No	Nucleotide		Exon	Amino Acid Change	Side chain polarity	Charge	HI
1	106	A	II	Lysine	basic polar	+	-3.9
		G		Glutamic acid	acidic polar	_	-3.5
2	122	G	III	Arginine	basic polar	+	-4.5
		A		Histidine	basic polar	+	-3.2
		C		Proline	non polar	neutral	-1.6
3	170	A	III	Glutamine	polar	neutral	-3.5
		C		Alanine	non polar	neutral	1.8
4	172	G	III	Glutamine	polar	neutral	-3.5
		A		Lysine	Basic polar	+	-3.9
5	364	A	IV	Asparagine	polar	neutral	-3.5
		G		Aspartic acid	acidic polar	_	-3.5
6	439	G	V	Glutamic acid	acidic polar	_	-3.5
		A		Lysine	basic polar	+	-3.9
7	492	C	VI	Serine	polar	neutral	-0.8
		G		Arginine	Basic polar	+	-4.5
8	500	G	VI	Metionine	non polar	neutral	1.9
		T		Arginine	basic polar	+	-4.5
9	580	T	VI	Cysteine	polar	neutral	2.5
		C		Leucine	non polar	neutral	3.8
10	626	G	VII	Arginine	basic polar	+	-4.5
		A		Histidine	basic polar	+	-3.2
11	756	C	VIII	Tyrosine	polar	neutral	-1.3
		T		Tyrosine	polar	neutral	-1.3
12	762	A	VIII	Lysine	Basic polar	+	-3.9
		T		Asparagine	polar	neutral	-3.5
13	770	C	VIII	Alanine	non polar	neutral	1.8
		A		Asparagine	polar	neutral	-3.5
14	842	T	IX	Isoleucine	non polar	neutral	4.5
		A		Lysine	basic polar	+	-3.9
15	879	C	IX	Asparagine	polar	neutral	-3.5
		T		Asparagine	polar	neutral	-3.5
16	947	A	X	Histidine	basic polar	+	-3.2
		G		Arginine	basic polar	+	-4.5
17	1024	G	XI	Glutamic acid	acidic polar	_	-3.5
		A		Lysine	Basic polar	+	-3.9

Note:
N, nucleotide; HI, Hydrophaty Index (Kyte and Doolittle, 1982).

The variation of isoelectric point and molecular weight for each species can be seen in Figure 3.7. The isoelectric point (pI), sometimes abbreviated to IEP, is the pH at which a particular molecule or surface carries no net electrical charge. At a pH below their pI, proteins carry a net positive charge; above their pI they carry a net negative charge.

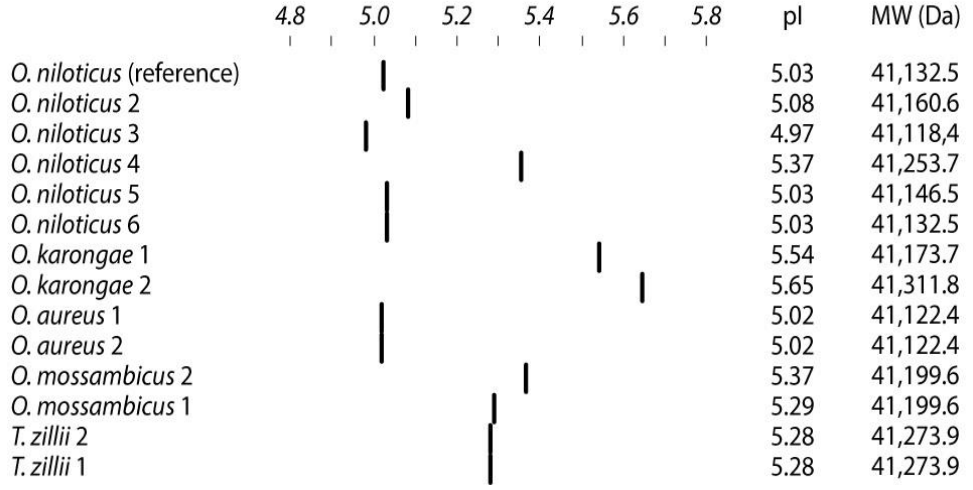


Figure 3.7 Isoelectric point and molecular weight of adenosine deaminase from 5 tilapia species.

The highest pI values occurred in *O. karongae* 2 (5.54-5.65) resulting in individuals with different polypeptide amino acid sequences or alleles that would show up in gel electrophoresis as different allozymes. Conversely, the lowest pI value was in *O. niloticus* 3 (4.97). Interestingly, *O. niloticus* 5 and 6 had the same pI value with the sequence reference (5.14), however, molecular weight in *O. niloticus* 5 was slightly higher than the reference and *O. niloticus* 6. So, potentially four different allozymes were found in five individuals of *O. niloticus*. Apparently, *O. aureus* and *T. zillii* had the same amino acid sequences within the individuals

analysed. The molecular weight of adenosine deaminase sequences varied between species. The highest was found in *karongae* 2 (41,312 .76 Da), while the lowest was found in *O. niloticus* 3 that reach 41,118.4 Da. It can be seen that individual *O. niloticus* 6 has similar molecular weight to the reference.

3.4.3. *ADA* gene tree

Nucleotides of *ADA* gene in *O. niloticus*, *O. mossambicus*, *O. karongae*, and *O. aureus* showed high sequence identities (99%) compared to the *ADA*-like gene reference from *O. niloticus* (accession number: XM_003457049.2), while a lower similarity (94%) was observed with *T. zillii*. Blast searches of the GenBank confirmed that the coding sequences also showed extensive similarities (96-78%) to other fish species (*Pundamilia nyererei*, *Maylandia zebra*, *Haplochromis burtoni*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Xiphophorus maculatus*, *Takifugu rubripes* and *Salmo salar*).

The gene tree in Figure 3.8 shows the tilapia species grouped into two main clades, *Oreochromis* and *Tilapia* (bootstrap value/bv= 100). Within the *Oreochromis sp*, *O. niloticus* and *O. aureus* form one clade and *O. mossambicus* and *O. karongae* form another (bv= 100). The *T. zillii*, *O. aureus* and *O. mossambicus*, clusters showed little differentiation (bv= 100, 100 and 98 respectively). *O. karongae* (bv= 39) and *O. niloticus* (bv= 76) showed more intraspecific variation. An additional *ADA* sequence from *H. burtoni* (Accession number: XM 005918779.1) was included in the gene tree, and indicated a species outgroup compared to the tilapia species, as expected.

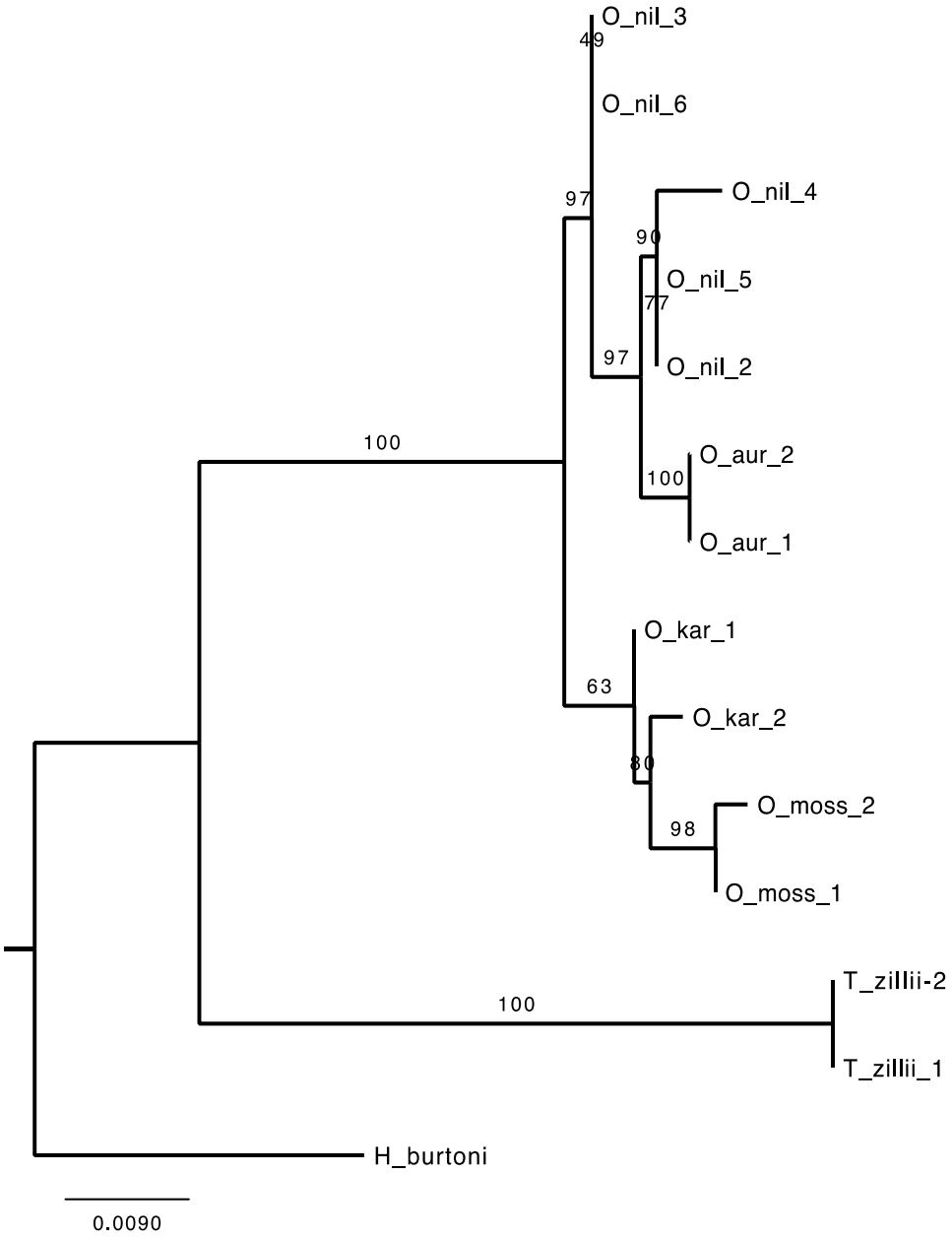


Figure 3.8 The gene tree of tilapia species inferred from adenosine deaminase *(ADA*) coding sequence, and rooted to *H. burtoni* as outgroup. All the sequences were constructed using RAxML software and visualized by FigTree. The numbers on the branches denote the frequencies (%) which describing the tree topology after bootstrapping (100 iterations). O_nil, *O. niloticus*; O_aur, *O. aureus*; O_kar, *O. karongae*; O_moss, *O. mossambicus*.

3.4.4. SNPs: Marker Application

SNP markers derived from adenosine deaminase nucleotide sequence were validated to more individuals of the tilapia species using KASP PCR genotyping assays. A preliminary matrix was made for particular evaluation of SNPs that could potentially be used for particular species pairs (Table 3.5).

a. SNP 1 (Oaur_3_R122)

SNP 1 was verified in 9 species, involving 28 individuals. *O. aureus* in nucleotide position 122 was homozygous G/G, whereas *O. karongae, O. mossambicus, O. niloticus, O. andersonii, O. macrochir, O. u. hornorum, S. melanotheron and S. galilaeus* were homozygous A/A (Figure 3.9).

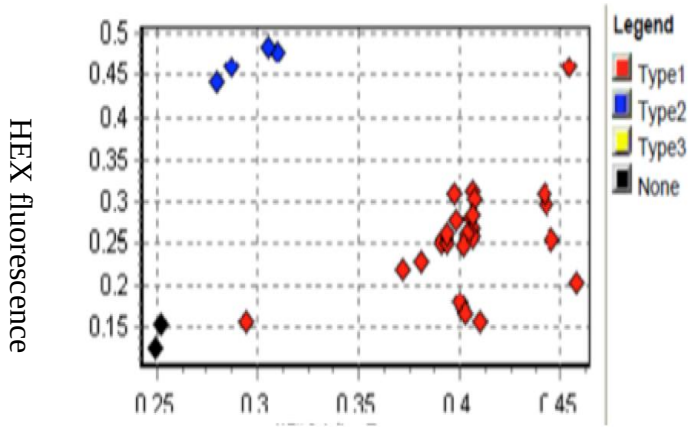


Figure 3.9 Genotypic discrimination graph of SNP 1 (Oaur_3_R122).
Type 1 (homozygous A/A): *O. karongae* (n=7), *O. mossambicus* (n=3), *O. niloticus* (n=4), *O. andersonii* (n=2), *O. macrochir, O. u. hornorum* (n=2), *S. melanotheron* (2) *and S. galilaeus* (n=2); Type 2 (homozygous G/G.): *O. aureus* (n=4).

Table 3.5 Preliminary matrix of SNP markers for distinguishing tilapiine species developed from *ADA* sequences
(N1, N2): number of individuals that were genotyped per each SNP assay. N1= number of individuals in the top row, and N2= number of individuals in the first column.

Note	
SNP 1	Oaur_3_R122
SNP 2	Tzil_3_M170
SNP 3	Okar2_6_S492
SNP 4	Okar2_6_K500
SNP 5	Tzil_6_Y580
SNP 6	Oaur_7_R626
SNP 7	Onil_8_Y756
SNP 8	Onil_8_Y762
SNP 9	Okar2_8_M770
SNP 10	Omoss_10_Y879

*One individual coding OK33 showed different allele

	Tzil	Smel	Sgal	Omac	Oan	Okar	Oh	Omos	Onil	Oaur
Tzil		SNP 8 (2;4)	SNP 2 (2;14)	SNP 2 (2;14)	SNP 2 (2;4)	SNP 2 (2;4)	SNP 2 (2;14)		SNP 2 (4,14)	
Smel				SNP 8 (2;2)	SNP 8 (1;2)	SNP 3 (5;2)	SNP 8 (2;2)	SNP 3 (3;2)		SNP 1 (4;2)
						SNP 8 (2;2)		SNP 8 (4;2)		SNP 8 (2;2)
Sgal				SNP 7 (2;1)	SNP 2 (2;2)	SNP 3 (8;2)	SNP 10 (5;9)	SNP 10 (6;9)		SNP 6 (8;2)
								SNP 3 (3;2)		SNP 1 (4;2)
Omac						SNP 4 (4,2)		SNP 10 (6;6)		SNP 1 (4;2)
Oan										SNP 2 (4;2)
							SNP 2 (2;2)	SNP 10 (6;2)		SNP 6 (8;4)
										SNP 1 (4;2)
										SNP 3 (8;3)
Okar										SNP 6 (8;6)
							SNP 4 (4,5)*	SNP 10 (6;7)	SNP 3 (4,4)	SNP 1 (4;7)
								SNP 4 (5;5)		SNP 3 (9;4)
Oh										SNP 6 (8;2)
										SNP 1 (4;2)
Omos										SNP 10(4;6)
										SNP 6 (8;4)
									SNP 10 (4;6)	SNP 1 (4;3)
										SNP 4 (8;3)
										SNP 6 (8;4)
										SNP 1 (4;4)
Oaur										

b. SNP 2 (Tzil_3_M170)

SNP 2 was validated in 8 species, totalling 37 individuals. The result of KASP-PCR clearly distinguished alleles to specific species. *ADA* sequences of *T. zillii* in nucleotide position 170 was fixed for allele C (Cytosine), while *O. karongae*, *O. aureus*, *O. mossambicus*, *O. niloticus, O. macrochir, O. u. hornorum* and *S. galilaeus* were fixed for allele A (Adenine) (Figure 3.10). *O. andersonii* (n=2), and *O. karongae* (n=1) appear to be heterozygous C/A at this locus.

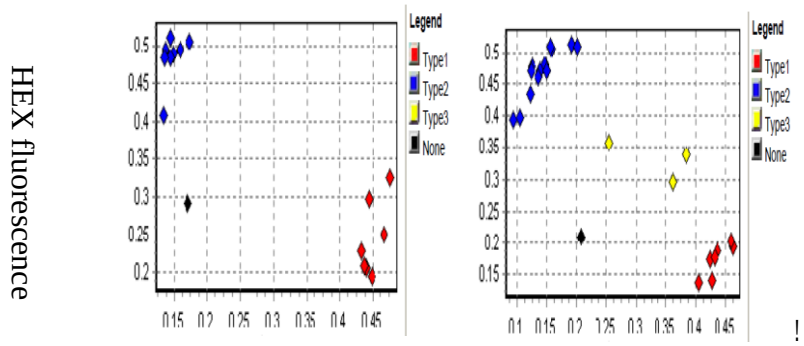


Figure 3.10 Genotypic discrimination graph of SNP 2 (Tzil_3_M170).
Type 1 (homozygous C/C): *T. zillii* (n=14); Type 2 (homozygous A/A): *O. karongae* (n=2), *O. aureus* (n=4), *O. mossambicus* (n=4), *O. niloticus* (n=4), *O. macrochir* (n=2), *O. u. hornorum* (n=2) and *S. galilaeus* (n=2), and Type 3 (heterozygous C/A) : *O. andersonii* (n=2) and *O.karongae* (n=1).

Further assays were conducted to more thoroughly test discrimination between *O. niloticus* vs *O. u. hornorum*, and *O. niloticus* vs *O. andersonii* (Table 3.6). The three species were selected for further assays due to hybridization occurrence between two species pairs. There was an introduction of *O. niloticus* into the Kafue River, Zambia, leading to hybridization with native species *O.*

andersonii (Deines et al., 2014). In addition, testing SNP 2 was also conducted to species pairs *O. niloticus – O. u. hornorum* due to their potential involvement in the multiple hybrids, e.g. Florida reds. DNA samples of *O. andersonii* (n=10) and *O. niloticus* (n=12) were tested using SNP 2. *O. andersonii* had A/A, A/C and C/C genotypes with allele frequencies of A = 0.5 and C = 0.5. While, all *O. niloticus* indicated homozygous A/A (allele frequencies A= 1). Apparently, *O. u. hornorum* showed similar genotype homozygous A/A to *O. niloticus*. SNP 2 can be used to distinguish *T. zillii* from the other tilapia species tested, however this marker did not clearly distinguish between *O. niloticus - O. andersonii* and *O. niloticus - O. u. hornorum*.

Table 3.6 Genotype and Allele frequencies from *O. niloticus* vs *O. u. hornorum* and *O. niloticus* vs *O. andersonii* using SNP marker SNP 2 (Tzil_3_M170).

No	Species	Population	Genotype (n)				Allele Frequency	
			AA	A/C	CC	∑n	A	C
1	*O. andersonii*	ITC, Zambia	4	2	4	10	0.500	0.500
2	*O. niloticus*							
		Stirling	2	0	0	2	1.000	0.000
	strain Kpandu	Ghana	3	0	0	3	1.000	0.000
	strain Nyinuto	Ghana	3	0	0	3	1.000	0.000
	Cancellatus	Hora, Ethiopia	1	0	0	1	1.000	0.000
		Koka, Ethiopia	1	0	0	1	1.000	0.000
		Metahara, Ethiopia	2	0	0	2	1.000	0.000
	∑		12	0	0	12	**1.000**	**0.000**
3	*O. hornorum*		5	0	0	5	1.000	0.000

c. SNP 6 (Oaur_7_R626)

SNP 6 was verified in 7 species, involving 31 individuals. *ADA* nucleotide 626 in *O. aureus* was homozygous A/A, whereas in *O. karongae, O. mossambicus, O. niloticus, O. u. hornorum, T. zillii, O. andersonii and S. galilaeus* it was homozygous G/G. Based on the matrix (Table 3.5), SNP 6 could distinguish *O.*

aureus - *O. niloticus* and *O. aureus* - *O. u. hornorum*. Further assays were conducted using more individuals of these species (Table 3.7).

The majority of *O. niloticus* were homozygous G/G (frequency allele G=0.885), except Nyinutho population from Ghana that shared the genotype A/A with *O. aureus* (n=4) from Ain Faskha Israel and Stirling (n=4). While, one individual of Nyinuto population from Ghana were heterozygous A/G with *O. aureus* (n=3) from Ain Faskha, Israel. Evidently, *O. aureus* both from Stirling and Ain Faskha can be discriminated from *O. hornorum* using SNP 6. Seven individuals of *O. hornorum* were homozygous G/G (n=12), while *O. aureus* consistently indicated homozygous A/A for Stirling (n=4). However, population Ain Feskha, Israel indicated homozygous A/A (n=4), and heterozygous A/G (n=3) with total frequency of allele A = 0.864.

Table 3.7 Genotype and allele frequencies from *O. aureus* - *O. niloticus* and *O. aureus* - *O. hornorum* using SNP 6 (Oaur_7_R626).

No	Species/strain	Population	Genotype (n)				Allele Frequency	
			AA	A/G	GG	∑n	A	G
1	*O. aureus*	Stirling	4	0	0	4	1.000	0.000
		AF, Israel	4	3	0	7	0.786	0.214
	∑		**8**	**3**	**0**	**11**	**0.864**	**0.136**
2	*O. niloticus*							
		Stirling	0	0	3	3	0.000	1.000
	Kpandu	Ghana	0	0	2	2	0.000	1.000
	Nyinuto	Ghana	1	1	0	2	0.750	0.250
	Cancellatus	Hora, Ethiopia	0	0	2	2	0.000	1.000
		Koka, Ethiopia	0	0	1	1	0.000	1.000
		Metahara, Ethiopia	0	0	3	3	0.000	1.000
	∑		**1**	**1**	**11**	**13**	**0.115**	**0.885**
3	*O. hornorum*		**0**	**0**	**12**	**12**	**0**	**1.00**

d. SNP 7 and 8 (Onil_8_Y756 and Onil_8_Y762)

SNP 7 was verified in 10 species, involving 29 individuals. The result of KASP-PCR did not clearly discriminate *O. niloticus* from the other species. There were three genotypes observed in *O. niloticus* at the *ADA* nucleotide position 756; C/C, T/T and T/C with frequency of allele C =0.6, and allele T = 0.4. Meanwhile *S. galilaeus* had both homozygous T/T and heterozygous T/C genotypes with allele frequency of T = 0.68. *O. karongae, O. mossambicus, O. hornorum, T. zillii, O. aureus, O. andersonii* and *O. macrochir* were all homozygous C/C, and *S. melanotheron* were all homozygous T/T. This marker could not be used to distinguish *O. niloticus* with other species due to the highly sequences variation within species. However this marker has potential to distinguish between *S. melanotheron* (n=2) against *T. zillii* (n=4) and *S. melanotheron* (n=2) against all *Oreochromis* sp (n=1-5 per spesies) apart from *O. niloticus* (Table 3.5).

SNP 8 was verified in 9 species, involving 26 individuals. The result of KASP-PCR did not clearly distinguish *O. niloticus* from other species. There were two genotypes of *O. niloticus* in *ADA* nucleotide position 762, homozygous T/T and A/A with allele frequency of T = 0.4, and A =0.6 respectively. Meanwhile species *O. karongae, O. mossambicus, O. hornorum, T. zillii, O. aureus, O. andersonii, S. melanotheron and S. galilaeus* showed only the homozygous T/T genotype.

e. SNP 10 (Omoss_10_Y879)

SNP 10 was verified in 10 species, consisting of 28 individuals (Figure 3.8). The result of KASP-PCR in *O. mossambicus* at nucleotide position 879 showed all

but one were T/T homozygous, while OM-13 was a C/C homozygote. Two other species, *O. andersonii* and *O. u. hornorum,* also showed only T/T homozygotes. *O. karongae*, *O. aureus*, *O. niloticus*, *O. macrcochir, S. melanotheron, T.zillii* and *S. galilaeus* were all C/C homozygotes.

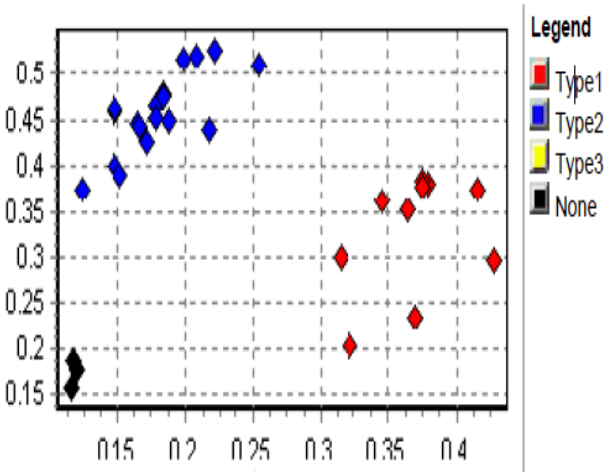


Figure 3.11 Genotypic discrimination graph of SNP 10 (Omoss_10_Y879)

Type 1 (homozygous TT): O. mossambicus (n=6), O. u. hornorum (n=2) and O. andersonii (n=2); Type 2 (homozygous CC): O. karongae (n=2), O. aureus (n=4), T. zillii (n=4), O. niloticus (n=4), O. macrochir (n=2), and S. galilaeus (n=2), S. melanotheron (n=2) and O. mossambicus (n=1).

This marker can be applied in hybrid identification between the most common *O. mossambicus* vs *O. niloticus* and *O. mossambicus* vs *O. aureus* hybrid combinations (Matrix in Table 3.5). Therefore, the locus was tested against a larger number of individuals (Table 3.8). The genotypic discrimination graph and allele frequency indicated that all *O. mossambicus* but one were a fixed homozygote T/T in nucleotide position 879, and *O. niloticus* was fixed for the homozygote C/C.

Table 3.8 Genotype and allele frequencies of *O. mossambicus* VS *O. niloticus* using SNP 10 (Omoss_10_Y879).

No	Species/sub sp	Genotype			∑n	Allele Frequency	
		TT	T/C	CC		T	C
1	*O. mossambicus*						
	Stirling	5	0	0	5	1.000	0.000
	Nathal, SA	6	0	0	6	1.000	0.000
	∑	**11**	**0**	**0**	**11**	**1.000**	**0.000**
2	*O. niloticus*						
	a. *niloticus*						
	Stirling	0	0	3	3	0.000	1.000
	Kpandu, Ghana	0	0	6	6	0.000	1.000
	Nyinuto, Ghana	0	0	5	5	0.000	1.000
	b. *cancellatus*						
	Hora, Ethiopia	0	0	3	3	0.000	1.000
	Koka, Ethiopia	0	0	1	1	0.000	1.000
	Metahara, Ethiopia	0	0	1	1	0.000	1.000
	∑	**0**	**0**	**19**	**19**	**0.000**	**1.000**

Hence, SNP 10 can be widely used to distinguish between *O. mossambicus* and *O. niloticus*. SNP 10 can also discriminate between *O. mossambicus* and *O. aureus* (Table 3.9).

Table 3.9 Genotype and allele frequency of *O. mossambicus* - *O. aureus* using SNP 10 (Omoss_10_Y879).

No	Species	Genotype			∑n	Allele Frequency	
		TT	T/C	CC		T	C
1	*O. mossambicus*						
	Stirling	5	0	0	5	1.000	0.000
	Nathal, SA	6	0	0	6	1.000	0.000
	∑	**11**	**0**	**0**	**11**	**1.000**	**0.000**
2	*O. aureus*						
	Stirling	0	0	7	7	0.000	1.000
	Ain Faskha, Israel	0	0	9	9	0.000	1.000
	∑	**0**	**0**	**16**	**16**	**0.000**	**1.000**

All *O. mossambicus* were homozygous T/T, while all *O. aureus* from both Stirling and Ain Faskha were homozygous CC.

f. SNP 3, 4, 9 (Okar2_6_S492, Okar2_6_K500, and Okar2_8_M770)

Three SNP markers in nucleotide position 492, 500 and 770 from *O. karongae* were verified in a number of other tilapia species (Table 3.10). SNP 3 was tested in 10 species comprising 44 individuals. *O. karongae* position 492 showed both the G/G and C/C genotype with allele frequencies G = 0.83 and C = 0.17. *O. aureus, O. mossambicus, O. niloticus, O. hornorum, O. macrochir, S. melanotheron, T. zillii, O. andersonii* and *S. galilaeus* showed the C/C or G/C genotype. Furthermore *O. niloticus, O. mossambicus* and *O. macrochir* were homozygous G/G. There was no specific genotype in the nucleotide 492 that distinguishing between tilapia species.

Table 3.10 Genotyping tilapia species using SNP 3,4 and 9 (Okar2_6_S492, Okar2_6_K500, and Okar2_8_M770)

No	Species/ population (n)	Genotype with SNP 3			Genotype with SNP 4			Genotype with SNP 9		
		G/G	G/C	C/C	G/G	G/A	A/A	A/A	A/C	C/C
1	*O. karongae* (6/5/6)	5	0	1	4	1	0	4	0	2
2	*O. aureus* (7/8/4)	0	0	7	0	0	8	2	0	2
3	*O. mossambicus* (6/6/6)	2	1	3	3	3	0	0	0	6
4	*O. niloticus* (9/5/9)	1	2	6	0	2	3	1	0	8
5	*O. andersonii* (4/3/2)	0	2	2	0	3	0	0	0	2
6	*O. macrochir* (2/0/4)	2	0	0	-	-	-	2	0	2
7	*O. u. hornorum* (2/3/2)	0	2	0	1	2	0	0	0	2
8	*S. galilaeus*	0	0	2	0	0	2	0	0	2

	(2/2/2)									
9	*S. melanotheron* (2/2/1)	0	0	2	0	0	2	0	0	1
10	*T. zillii* (4/8/4)	0	2	2	0	3	5	2	0	2

SNP 4 was verified in 9 species, involving 42 individuals. The result showed homozygous G/G and heterozygous G/A genotypes shared among *O. karongae*, *O. mossambicus*, and *O. hornorum*, meanwhile other species displayed homozygous A/A genotype. Therefore this marker was not applied in the further study. In the SNP 9, *ADA* nucleotide in position 770 denoted two different homozygous A/A & C/C genotypes, both in *T. zillii* and *O. karongae*, meanwhile *O. aureus, O. niloticus, O. mossambicus, O. macrcochir, O. hornorum, S. melanotheron, O. andersonii* and *S. galilaeus* indicated homozygous C/C genotype. *T. zillii* only showed genotype A/A homozygous, however the SNP assay indicated A/A and C/C homozygous A/A with frequency 0.5 for both allele A and C. *O. aureus* (n=7) and *O. mossambicus* (n=6) showed consistently the genotype C/C homozygous. Therefore, there was no specific genotype that distinguishing tilapiine using this marker.

Hence four SNP markers (some of them based on small sample sizes) could be potentially used to test into expanded species and populations (Table 3.11):

Table 3.11 SNP marker potentially applied to wider population

SNP Marker	Distinguishing ability	
	with species studied	With others species pairs
Tzil_3_M170	except heterozygous for *O. karongae* and *O. andersonii*	Not applicable between *O. niloticus* - *O. andersonii* and *O. niloticus* - *O. u. hornorum*
Oaur_3_R122	√	Have not been tested (see its potential in Table 3.6)
Oaur_7_R626	except heterozygous for *O. aureus* and *O. niloticus*	*O. aureus* - *O. u. hornorum*
Omoss_10_Y879	√	Distinguishes *O. mossambicus*, *O. u. hornorum* and *O. andersonii* (all TT) from all other spp (CC), apart from one *O. mossambicus* that was homozygous for CC. Expanded species pairs: *O. mossambicus* - *O. niloticus* and *O. mossambicus* - *O. aureus*.

g. SNP marker application in known tilapia hybrid

SNP markers 6 and 10 were tested to known tilapia hybrids originally from Indonesia (Table 3.12). The result based on SNP 6 indicated that defining Pandu (n=1) population from Freshwater Research Center Unit in Klaten, Indonesia had heterozygous A/G, similar to three individuals of *O. aureus*, while the species reference *O. niloticus* (n=4) had homozygous G/G. SNP 10 indicated that known

O. niloticus and hybrid populations from Indonesia are still identified as *O. niloticus*.

Table 3.12 SNP markers application in known hybrids using SNP 6 (Oaur_7_R626) and 10 (Omoss_10_Y879).

No	Species/population (n)	Genotype with SNP 6			Genotype with SNP 10		
		A/A	A/G	G/G	T/T	T/C	C/C
1.	Species references						
	a. *O. niloticus* (4/3)	0	0	4	0	0	3
	b. *O. aureus* (11/16)	8	3	0	0	0	16
	c. O. mossambicus (4/11)	0	0	4	11	0	0
2.	Known tilapia hybrids						
	a. Pandu (2/2)	0	1	1	0	0	2
	b. Larasati (2/2)	0	0	2	0	0	2
	c. Sulthana (4/3)	0	0	4	0	0	3
	d. YY (3/3)	0	0	3	0	0	3

3.5 Discussion

3.5.1. Polymorphisms in *ADA* gene

The variation of some nucleotides between tilapia species in coding sequences (exon) of the *ADA* gene indicated synonymous and non-synonymous substitutions. Non-synonymous polymorphisms impact on side chain polarity and hydropathy index. The ratio of synonymous polymorphisms in *O. niloticus* was higher in comparison to other tilapia species. Synonymous substitutions accumulate much more rapidly than non-synonymous substitutions because they are far less likely to cause phenotypic changes. Mutations tend to accumulate more rapidly in introns compared to exons. Pseudogenes appear to have higher substitution rates compared to functional genes (Li, 1997). The most significant finding, that the *ADA* gene in *O. niloticus* indicated dN/dS is greater than 1, suggests that the region in this species is under positive selection although functional impacts are unknown. In contrast the dN/dS others species showed the value less than 1, indicated that in these species they might be selectively neutral.

The different polymorphisms between tilapia species influence amino acid change, side chain polarity, side chain charge and also hydropathy index. Ten amino acids, glycine, alanine, valine, leucine, isoleucine, methionine, proline, phenylalanine and tryptophan are classified as having nonpolar side chains. Meanwhile, five amino acids, serine, threonine, asparagine, glutamine, and tyrosine are commonly classified as having un-charged polar side chains (Alberts et al., 2008). Five amino acids (glutamic acid, aspartic acid, arginine, lysine, and histidine) have ionisable side chains that give a protein characteristic net charge,

depending on the pH of the surrounding medium. An amino acid substitution may directly replace one of these charged amino acids, or a non-charged substitution near one of them in polypeptide chain may affect the degree of ionization of the charged amino acid, or a substitution at the junction between two α helices may cause a slight shift in the three-dimensional packing of the folded polypeptide. Obviously, the net charge on the polypeptide is altered because the net charge on a protein is not simply the sum of all individual charges on its amino acids, however this will depend on their exposure to the liquid medium in its surrounding (Griffith et al. 2002). An important factor governing the folding of any protein is the distribution of its polar and nonpolar amino acids. The nonpolar (hydrophobic) side chains in a protein tend to cluster in the interior of the molecule (just as hydrophobic oil droplets coalesce in water to form one large droplet). This enables them to avoid contact with the water that surrounds them inside a cell. By contrast, polar groups tend to arrange themselves near the outside of the molecule, where they can form hydrogen bonds with the water and with other polar molecules (Alberts et al., 2008).

The highest pI values occurred in *O. karongae* 2 (5.54-5.65) resulting individuals with different polypeptide amino acid sequences. Conversely, the lowest pI value was in *O. niloticus* 3 (4.97). Apparently, species *O. aureus* and *T. zillii* denoted the similar type of allelic within species. Molecular weight of adenosine deaminase sequences varied between species. The highest was found in *karongae* 2 (41,312.76 Da), meanwhile the lowest was found in *O. niloticus* 3 that reach 41,118.4 Da. *O. niloticus* 6 had a similar molecular weight to the reference (accession number: XM_003457049.2). The variation of pI value due to amino

acid changes in tilapia's nucleotide lead to allelic differences at the SNP. This supported previous study where monomeric enzyme in ADA has a wide range or mobility between tilapia species (McAndrew & Majumdar, 1983). Alleles can differ within species, for instance in *O. karongae*, however they also can be similar between species as happened in *T. zillii* and *O. aureus*. The atomic weight of an atom, or the molecular weight of a molecule, is its mass relative to that of a hydrogen atom. The mass of this molecule is often specified in dalton, whereas one Dalton being an atomic mass unit approximately equal to the mass of a hydrogen atom (Alberts et al., 2008). Obviously, the molecular weight of most protein was well conserved across species boundaries due to the presence of shared domains that classify the protein as functionally identical. Conversely, protein point isoelectric was frequently not well conserved. Protein for differences species known to have the same function can have isoelectric point, which is greatly different (Wilkins & Williams, 1997).

3.5.2. *ADA* gene tree

The gene tree in Figure 3.7 showed that all individuals were clustered in two different clades of *Oreochromis* sp and *T. zillii*. Each species can be clearly distinguished among others species. However, *O. niloticus* appeared highly polymorphic across the different wild and captive populations sampled but generally each population tended to cluster together. *Tilapia zillii* were highly similar within species and the most distance group between tilapia species (bootstrap value=100) analysed. Nevertheless, the low number of samples analysed might limit the observed variation. The availability of sequence data in

many species provides a database to determine the phylogenetic relationship between species or related taxa (Hedrick, 2005). However, as we are only using a single homologous gene in the different species, we are generating a gene tree, rather than the evolutionary history of a group of species. Where the splitting of one species into two species indicates the time of origin, this can be represented in a species (population) tree (Nei & Kumar, 2000). Therefore, if the genes are polymorphic within the species that diverged, the times of divergence in the gene tree may be greater than the times of divergence of the species.

The majority of genetic polymorphism arises randomly and is maintained or lost as a result of random events. However, the molecular changes might also be retained by selective processes. A combination of stochastic effect and natural selection means that a proportion of mutations will inevitably be maintain within a species and this accumulation of mutation, along with recombination, means that even the same species frequently have divergent genomes, for instance in the human genome, approximately 0.1% is variable. Sequence divergence tends to be higher between more distantly related groups (Li and Sadler, 1991 *in* Freeland et al., 2001). The branch lengths are proportional to the amount of genetic change that has occurred (Freeland, 2011).

3.5.3. SNP Genotyping by KASP assay

The SNP discovery between tilapia species based on *ADA* gene sequence indicated that four out of 10 SNP markers work well to distinguish tilapia species. The SNP markers are *T. zillii* (Tzil_3_M170), *O. aureus* (Oaur_3_R122 and Oaur_7_R626) and *O. mossambicus* (Omoss_10_Y879), which could be more

widely tested to investigate hybridization and introgression. All these SNP markers indicated specific homozygous genotypes based on KASP assay.

Marker Tzil_3_M170 was tested to distihguish between *O. niloticus* - *O. andersonii* and *O. niloticus* - *O. hornorum*, but it did not show clear discrimination. Marker Oaur_7_R626 was tested to distinguish *O. aureus* - *O. niloticus* and *O. aureus* - *O. hornorum*. Generally, vast majority of strains in *O. niloticus* indicated homozygous G/G, except one individual of Nyinutho stock from Ghana, which shared homozygous A/A genotype with *O. aureus* (n=4) from Ain Faskha Israel and Stirling (n=4). This species also shared heterozygous A/G genotype with 3 individuals of *O. aureus* from Ain Faskha, Israel. Initially, the population of Nyinuto, located at the downstream limit of Nile tilapia distribution along the Volta River, is likely to have more limited size and to be easily subject to fluctuation, for instance variation of water level and salinity, beside it was also more isolated than a central basin population. Conversely the population of Kpandu, located at the shore of Lake Volta is likely to have a large census size and/or well be connected to neighbouring populations (Bezault et al., 2011).

SNP marker Oaur_7_R626 was used to discriminate *O. aureus* both Stirling anad Ain Faskha, Israel populations from *O. hornorum*. Seven individuals of *O. hornorum* showed homozygous G/G genotype, meanwhile *O. aureus* consistently displayed A/A homozygous for Stirling (4), and homozygous A/A (4) plus heterozygous A/G (3) with allele frequency 0.786 and 0.429 for Ain Faskha population Israel, respectively. The heterozygote presence in the *O. aureus* population indicated that it is very likely that there will also be some GG homozygotes. Therefore, this marker does not discriminate *O. aureus* from other

species. SNP 10 is able to distinguish species pairs between *O. mossambicus* - *O. niloticus* and *O. mossambicus* – *O. aureus* (Table 3.11).

3.5.4 SNP application in a known tilapia hybrid

Some *O. niloticus* and hybrid populations from Indonesia were also tested with SNP 6 and 10 markers (Table 3.12). SNP marker 6 was tested to distinguish *O. aureus* - *O. niloticus*. Generally, the majority of populations in *O. niloticus* were homozygous G/G, except one individual of Pandhu stock. Based on genotyping, this individual shared heterozygous A/G genotype with 3 individuals of *O. aureus* from Ain Faskha, Israel. Historically, stock Pandhu from Indonesia was Singapore Red tilapia from the National Inland Fisheries Institute (NIFI) resources, a cross between a mutant reddish-orange female *O. mossambicus* and a normal male *O. niloticus.* Based on genotyping using SNP marker 10, known *O. niloticus* and hybrid populations from Indonesia are still identified as *O. niloticus*.

It was becoming evident from the initial sequence analysis that we were unlikely to find species-specific SNP markers that were as clear as the allozyme results. Furthermore, the low sample size (only 2-3 individuals tested per population) remained a limitation to distinguish species and hybrid occurrence. Therefore, it suggests applying the SNP markers to many more individuals with different species/population. With a limited number of species-specific SNP markers based on *ADA* sequence variation, Restriction-site Associate DNA sequencing (RADseq and ddRADseq), described in the following chapters, offered greater potential to generate many SNPs with alleles unique to each species.

3.6 Conclusion

Adenosine deaminase (*ADA*) sequences were polymorphic in species *O. niloticus*, *O. aureus*, *O. mossambicus*, *O. karongae* and *T. zillii*, where *T. zillii* was the most genetically distant from other species and *O. niloticus* showed the highest polymorphism within species. The *ADA* approach was partially successful in developing species-specific SNP markers, where four out of ten SNP markers from *T. zillii* (Tzil_3_M170), *O. aureus* (Oaur_3_R122, Oaur_7_R626) and *O. mossambicus* (Omoss_10_Y879) can be applied in identifying and discriminating among certain tilapia species. Despite the ADA enzyme being variable among tilapia species based on allozyme studies and sequence variation, developing SNPs based only on a single marker (only one locus) limits its application to species discrimination. In addition, SNP markers validation in this experiment used only a few individuals, so less common alleles may not have been found in certain species.

3.7 Acknowledgments

We thank Keith Ranson of the Tropical Aquarium Facility at the Institute of Aquaculture, University of Stirling, for help in rearing fish. The authors acknowledge the support of the Director General of Higher Education, Ministry of Research, Technology and Higher Education (*Kemenristek Dikti*), Indonesia for funding PhD scholarship of MS at the University of Stirling. We also thank Prof. Slamet Budi Prayitno (University of Diponegoro Semarang) and Toni Kuswoyo, S.Pi, M.Si (Freshwater Research Center Unit, Klaten) Indonesia for *O. niloticus* sample.

4. SPECIES-SPECIFIC SNP MARKERS AND THEIR GENOMIC DISTRIBUTION IN TILAPIA SPECIES

Mochamad Syaifudin[1,2], Michaël Bekaert[1], John B. Taggart[1], Karim Gharbi[3], Gideon Hulata[4], David J. Penman[1,#], Brendan J. Mcandrew[1]

[1] Institute of Aquaculture, School of Natural Sciences, University of Stirling, Stirling FK9 4LA, Scotland, United Kingdom

[2] Program Study of Aquaculture, Faculty of Agriculture, University of Sriwijaya, South Sumatera, Indonesia

[3]Edinburgh Genomics, Ashworth Laboratories, King's Buildings, University of Edinburgh, Edinburgh EH9 3JT, Scotland, United Kingdom

[4]Dept. of Poultry and Aquaculture, Institute of Animal Science, Agricultural Research Organization, The Volcani Center, Bet Dagan, Israel

[#] Corresponding author: Tel +44 (0)1786 467901; Fax +44 (0)1786 472133; Email d.j.penman@stir.ac.uk

Keywords: Tilapia, Species-specific markers, Phylogeny, RADseq, SNP.

Abbreviations :

RAD: restriction-site associated DNA; SNP: single nucleotide polymorphism; DBA: de novo-based analysis; RBA: reference-based analysis; LG: Linkage group.

Contributions: The first draft of the present manuscript was compiled and written in full by the author of this thesis, who was also fully involved in all subsequent

editing. The DNA extraction, RAD libraries preparation (under John B. Taggart's assistance), PCR, Cytochrome c oxidase subunit I sequence analysis and physical mapping of species-specific diagnostic SNP markers were conducted by the candidate. The other co-authors contributed towards the experimental design, the analysis of sequenced reads derived from RADseq, phylogenetic tree reconstruction, sequences alignment, the SNP positioning across the reference genome of *O. niloticus* and editing.

4.1 Abstract

Background

Identification of tilapia species, hybrids and introgressed populations is of importance in aquaculture and in areas where introductions of tilapias have occurred. Many species are morphologically similar, particularly juveniles and females. Using restriction site associated DNA (RAD) sequencing, we aimed to find single nucleotide polymorphisms (SNP) in nuclear DNA that distinguish between tilapia species, and to analyse the distribution of such markers in the genome.

Results

Analysis of sequence data from RBA detected 1,613 shared SNP markers in 1,002 RAD loci among seven tilapia species, *Oreochromis aureus*, *O. karongae*, *O. mossambicus*, *O. niloticus*, *O. urolepis*, *Sarotherodon galilaeus*, *Tilapia zillii* and an outgroup cichlid species, *Pelvicachromis pulcher*. A phylogenetic tree based on these markers showed a very similar pattern to the consensus derived from earlier molecular marker-based analyses. Further analysis detected 677

species-specific SNP markers for the seven tilapia species (i.e., allele[s] unique to a single species). Physical mapping of these species-specific SNP markers onto the *O. niloticus* genome assembly showed that they were relatively evenly distributed across the genome, ranging from 0.47 SNPs/Mb in linkage group 3 to 1.53 SNPs/Mb in linkage group 9. Analysis of cytochrome oxidase unit I DNA sequence in each tilapia species resulted in a similar phylogenetic tree to that generated using SNP markers from RAD sequencing.

Conclusions

The large number (677) of species-specific SNP markers that were identified suggests that further studies including more species and populations should identify robust species-specific markers for this group of fish. The results demonstrate the potential of RADseq-based analyses for phylogeny reconstruction.

4.2 Introduction

The tilapias are a group of African and Middle Eastern cichlid fish that are widely cultured in both developed and developing countries (major producers include China, Egypt, Indonesia, Philippines, Thailand and Brazil), with total world aquaculture production of 4.5 million t and total value of 7.6 billion USD in 2012 (FAO, 2014). Of this, 3.8 million t was *O. niloticus*, representing 84.13% of the total. With many different species and sub-species of tilapia, extensive introductions (into approximately 140 countries) and use of interspecies hybrids in aquaculture, it is often difficult to differentiate these, or ascertain contribution to hybridized/introgressed stocks. The published descriptions of the species based on

meristic and morphometric characters show considerable variation and broad interspecific overlaps (B-Rao & Majumdar, 1998; Trewavas, 1983; Wohlfarth & Hulata, 1983).

In some cases tilapia species have been introduced into water bodies where other tilapia species exist. For example, the introduction of *O. niloticus* (Linnaeus) into Lake Victoria and some of its satellite lakes led to hybridization and introgression with the native species (*O. variabilis*, Boulenger, and *O. esculentus*, Graham) and eventually to the loss of these species from Lake Victoria (Agnése et al., 1999). Introduction of *O. niloticus* into the Kafue River, Zambia, led to introgression among this species and two native tilapias, *O. andersonii* (Castelnau) and *O. macrochir* (Boulenger) (Deines et al., 2014). Introgression between indigenous tilapia species has also occurred in instances where the environment has been significantly altered by man without introduction of non-native species, e.g. between at least two species of *Tilapia* when the River Bia was dammed to make Lake Ayamé in the Ivory Coast (Adépo-Gourène et al., 2006), and introgression has also been seen where more than one tilapia species has been introduced into water bodies outside of the natural distribution, e.g. there has been a high degree of mixing between *O. mossambicus* (Peters) and *O. niloticus* in Southern Sri Lanka (De Silva & Ranasinghe, 1989).

Within aquaculture, hybridization and introgression among tilapia species has sometimes been the result of poor management (McAndrew & Majumdar, 1983), but hybrids have also been produced intentionally. Historically, several hybrids were produced to try to generate all-male fish for aquaculture (Mcandrew, 1993). Today, F1 hybrids between *O. niloticus* and *O. aureus* (Steindachner)

account for a significant proportion of tilapia production in China, the largest global producer (Thodesen (Da-Yong Ma) et al., 2013). Red tilapia hybrids are commonly used in brackishwater environments: these may include genetic contributions from as many as four species (*O. mossambicus*, *O. niloticus*, *O. aureus* and *O. urolepis hornorum* (Norman) and are likely to have been introgressed for several generations (Penman & McAndrew, 2000). The base population for the GIFT (Genetically Improved Farmed Tilapia) strain, now widely distributed in many countries, contained stocks of *O. niloticus* collected directly from the wild in Africa and also Asian stocks of *O. niloticus* that are thought to have been introgressed with feral *O. mossambicus* (Acosta & Gupta, 2010; Macaranas et al., 1995; McKinna et al., 2010).

Genetic markers offer a practical means to confidently differentiate among tilapia species, to assess the genetic composition of established, feral hybrids in new environments, and to assess the composition of cultured stocks (Costa-Pierce, 2003). Different marker technologies have been tested as tools in species identification. Allozyme loci have been shown to distinguish between a number of tilapia species (Deines et al., 2014; De Silva & Ranasinghe, 1989; Penman & McAndrew, 2000; Costa-Pierce, 2003; Barriga-Sosa et al., 2004; Sodsuk & McAndrew, 1991), but samples need to be kept frozen and fish often need to be killed to obtain tissues for analysis. The number of detectable markers is also limited (<50 in most studies). Mitochondrial DNA (mtDNA) haplotypes differ among species of tilapia (D'Amato et al., 2007; He et al., 2011; Nagl et al., 2001; Rognon & Guyomard, 1997; Shirak et al., 2009) but mtDNA is in effect a single locus, haploid and maternally inherited, and thus of limited use in the analysis of

hybridization and introgression. West African *O. niloticus* exhibit mtDNA haplotypes typical of *O. aureus*, although nuclear markers (allozymes) clearly indicated the differences between these two groups, and shared identity of W. African *O. niloticus* with *O. niloticus* samples from other parts of the species' distribution (Agnése et al., 1997; Rognon & Guyomard, 2003). Microsatellite markers are often highly polymorphic, show overlapping allele size ranges between species and PCR products of the same size may actually be different alleles, which restricts their applications in species discrimination and phylogenetic studies, but some do exhibit diagnostic alleles for some tilapia species pairs, e.g. two microsatellite markers were diagnostic between *O. niloticus* and *O. esculentus* (Agnése et al., 1999), and private microsatellite alleles were found in *O. niloticus* that were not found in *O. leucostictus* (Trewavas) from Lake Baringo, Kenya (Nyingi & Agnèse, 2007). A few species-specific markers have also been identified from Randomly Amplified Polymorphic DNA (RAPD) screening (Bardakci & Skibinski, 1994; Dinesh et al., 1996; Hassanien et al., 2004) and restriction fragment polymorphisms in ribosomal DNA (Dinesh et al., 1996; Hassanien et al., 2004).

The genetic diversity in tilapias has been researched over many years; however an accurate assessment of hybridisation and introgression still remain problematic. With many different species and many captive aquaculture stocks of tilapia around the world, there is a pressing need for better tools for such analyses. In particular, larger numbers of species-specific markers would be very useful. High-throughput sequencing offers new opportunities to rapidly isolate and genotype very large numbers of genetic markers, primarily SNPs and

microsatellites, and to do so from structured samples (families, populations or species) in a way that will identify markers associated with specific traits or differences between populations or species (Baird et al., 2008). The objective of the research described here was to test the potential of this approach, exploiting restriction site associated DNA sequencing (RADseq) to isolate SNP markers to distinguish between seven tilapia species and to analyse the distribution of such markers in the genome. DNA sequence of the mtDNA COI gene was also analysed, as a reference for comparison. The conserved sequence of the 5' region of the mitochondrial gene cytochrome oxidase subunit I (COI or Cox1), a platform for the universal DNA barcoding of life, has been widely used for distinguishing, for example, species in the Persian Gulf (Asgharian et al., 2011), Australian fish (Ward et al., 2005), marine fishes in the northwest Atlantic Ocean, Canada (McCusker et al., 2013) and tilapia species (Shirak et al., 2009; Wu & Yang, 2012).

4.3 Materials And Methods

4.3.1 Ethics Statement

All working procedures complied with the UK Animals Scientific Procedures Act (Parliament of the United Kingdom 1986).

4.3.2 Biological Materials

Fin samples were collected from seven different tilapia species (from 5 to 9 individuals per species, from a single population in each case) and two individuals of *Pelvicachromis pulcher* (Boulenger) as an out-group.

Oreochromis niloticus, *Oreochromis aureus* and *Tilapia zillii* (Gervais) were collected from Lake Manzala, Egypt in 1979 (McAndrew et al., 1988). *O. mossambicus* was collected from the Zambezi River, Zimbabwe in 1985 (Majumdar & McAndrew, 1986). *Oreochromis karongae* (Trewavas) was collected originally from Lake Malawi, Tanzania in 1994. Populations of these fish were maintained in the Tropical Aquarium Facilities of the Institute of Aquaculture, University of Stirling. Samples for two species, *Oreochromis urolepis hornorum* (Norman; originally from Tanzania) and *Sarotherodon galilaeus* (Linnaeus) were sourced from Israel, while the out-group species, *P. pulcher*, was obtained from a local aquarist supplier. Samples were stored in 99% ethanol at -20°C until required.

4.3.3 Genomic DNA extraction

Total genomic DNA was extracted using the Realpure Genomic DNA Extraction Kit (Durviz S.L) following the manufacturer's protocol. An RNAse incubation step was included to minimise RNA contamination, with each precipitated DNA sample being finally resuspended in 5 mM Tris, pH8.5. Extracted DNA was quantified by spectrometry (Nanodrop ND 1000 Spectrophotometer, NanoDrop Technologies Inc., Montchanin, DE). Both 260/280 and 260/230 ratios were > 1.8 for all samples. Sample integrity was checked by agarose gel (0.8%) electrophoresis. Those samples that passed quality control (no observable RNA and comprising predominantly high molecular weight DNA) were selected for use and diluted to a concentration of 50 ng/L in 5 mM Tris; pH 8.5.

4.3.4 RAD library preparation and sequencing

Initially four RAD libraries were constructed, each comprising pooled DNA from 11-12 individually barcoded fish. Later a fifth library comprising DNA from a further 7 individuals was made. The steps of standard RAD experiment are described in the Figure 4.1, while details of species composition of the five libraries are given in Appendix 4.1. RAD library preparation followed the methodology of Etter et. al (Etter et al., 2011), with some modifications of Houston et. al (Houston et al., 2012). Each sample (0.25 µg DNA) was digested at 37°C for 45 min with *Sbf*I high fidelity restriction enzyme (New England Biolabs, NEB) using 6U *Sbf*I per g genomic DNA in 1× Reaction Buffer 4 (NEB) at a final concentration of 1 g DNA per 50 L reaction volume. The reactions (12.5 L final volume) were then heat inactivated at 65°C for 20 minutes. Individual specific P1 adapters, each with a unique 5 or 7 base barcode, were ligated to the *Sbf*I digested DNA at 22°C for 45 minutes by adding 1 L 100 nM P1 adapter, 0.15 L 100 mM rATP (Promega), 0.25 L 10× Reaction Buffer 2 (NEB), 0.125 L T4 ligase (NEB, 2 M U/mL) and reaction volumes made up to 15 L with nuclease free water for each sample. Following heat inactivation at 65°C for 20 minutes, the ligation reactions were then combined in appropriate library pools. Each library pool was physically sheared (using Covaris sonication) and size selected (190 – 510 bp) using agarose gel electrophoresis (Houston et al., 2012). The remainder of the library construction (i.e. gel purification; end repair, dA overhang addition, P2 paired-end adapter ligation and library amplification) followed the original RAD protocol (Etter et al., 2011). A total of 250 L of each amplified library (14 PCR cycles) was prepared, column purified, eluted in 35 L EB buffer and subjected to a second

round of size selection (c. 300–550 bp) by gel electrophoresis. Following a final gel elution step into 20 L EB buffer (MinElute Gel Purification Kit, Qiagen), the libraries were QCed by gel electrophoresis (1.5% gel) and Bioanalyser (Agilent Technologies) and accurately quantified by fluorimetry.

4.3.5 RAD libraries sequencing

Equimolar amounts of libraries 1-4 were combined and sequenced on a single lane of the Ilumina HiSeq 2000platform (100 bases, paired-end reads) at the GenePool Genomics Facility, University of Edinburgh. Library 5 was sequenced in house at University of Stirling using two runs of Illumina Miseq (v3 chemistry, 100 bases Read 1 / 75 bases Read 2). The diagram of standard RAD sequencing experiment and genotyping RAD alleles was described in Figure 4.1.

4.3.6 Genotyping RAD Alleles

The construction of RAD tags was based on the Read 1 sequence data only. Reads of low quality (QC values under 30), missing the expected restriction site or with ambiguous barcodes were discarded. Retained reads were sorted into loci using a reference-based analysis (RBA) (including only loci found in the Broad Institute of MIT and Harvard *O. niloticus* genome assembly Orenil1.1, NCBI assembly GCF_000188245.2 (Brawand et al., 2014) and genotyped using Stacks software 1.13 (Catchen et al., 2013). The likelihood-based SNP-calling algorithm (Hohenlohe et al., 2010) implemented in Stacks evaluates each nucleotide position in every RAD-tag of all individuals, thereby statistically differentiating true SNPs from sequencing errors. A minimum stack depth of at least 20 and a maximum of 2 mismatches were allowed in a locus in

an individual, with an additional mismatch allowed between individuals. Polymorphic RAD-tags may contain more than one SNP, but the vast majority (over 99%) showed only two allelic versions; RAD-tags with more than two alleles or shared by less than 75% of the samples were excluded. RAD loci that were shared among all tilapia species in the RBA, exhibited no intraspecific polymorphism but showed interspecific polymorphism were identified using *find_pattern.pl* (a bespoke Perl script to find fix allele patern) with grouping between individuals. The same analysis was used to retrieve species-specific markers with only one fixed allele and grouping between species. Data from the RBA was used for this purpose to ensure consistency of homologous loci across all species.

4.3.7 Phylogenetic Reconstruction

The phylogeny of the tilapia species was inferred from the shared SNP markers, and rooted to *P. pulcher* as an out-group. These SNP were concatenated and trees were constructed using RAxML v8 (Stamatakis, 2014) and PhyloBayes 3.3 (Lartillot, Lepage, & Blanquart, 2009). This analysis produced the Best-scoring ML tree to be constructed, with support values from 1000 bootstrap replicates. Bayesian posterior probabilities using GTR+CAT and GTR-Gamma using RAxML and GTR+CAT model using PhyloBayes were computed for each branch.

4.3.8 Physical Mapping

The locations of the shared SNP markers exhibiting interspecific polymorphism but no intraspecific polymorphism and specific of one species only

(*e.g.*, *O. niloticus* exhibiting a 'C' allele of a SNP and all other species a 'G') were extracted from the *O. niloticus* genome assembly and visualised using Genetic-Mapper v0.6 (Bekaert, 2014).

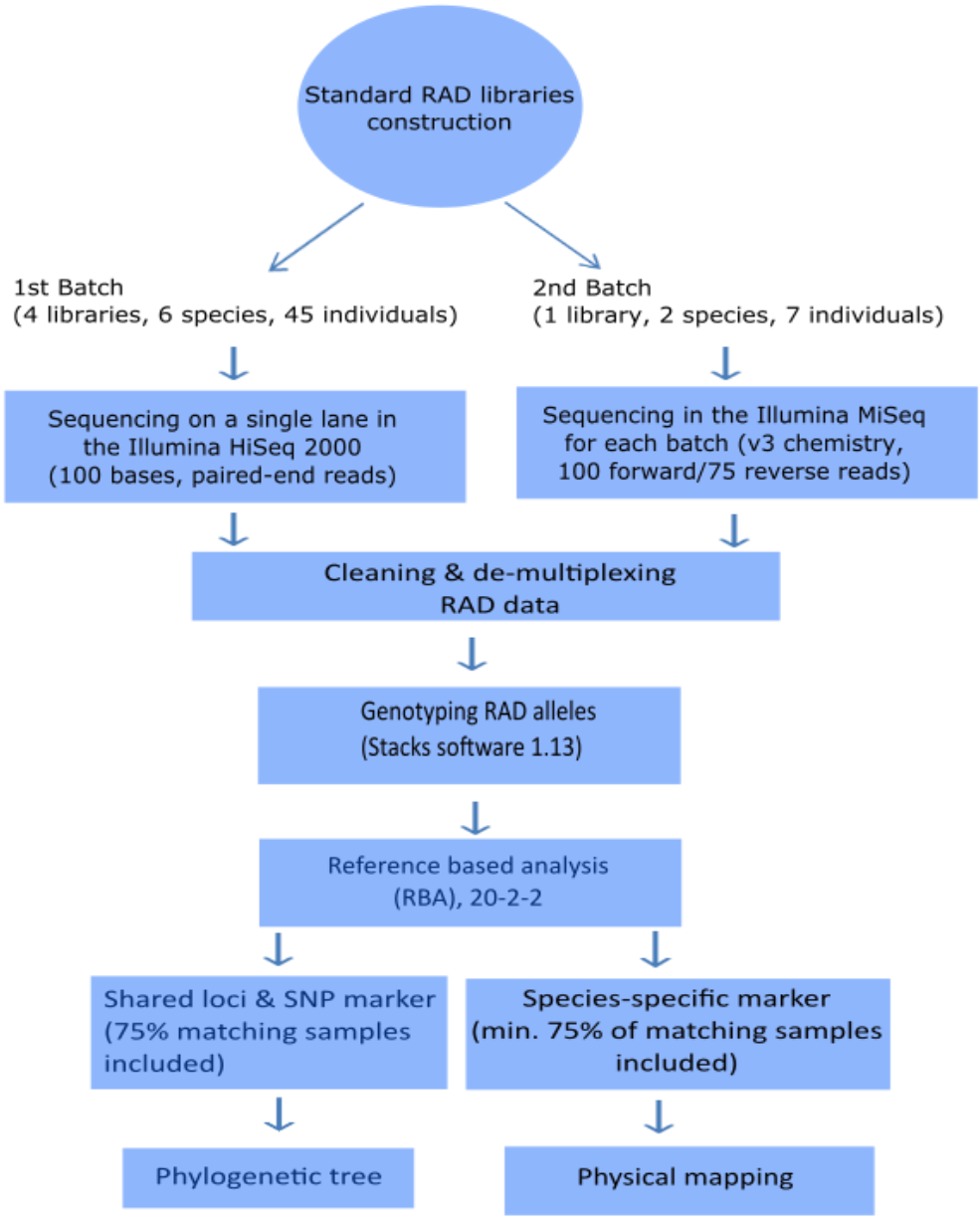


Figure 4.1 Flow diagram of standard RAD sequencing experiment and genotyping RAD alleles.
* Indicating the parameter applied in the stacks analyses, which are m : minimum stack depth; M : distance allowed between stacks; n : distance allowed between catalog loci.

4.3.9 COI DNA Barcoding

The DNA from 7 tilapia species (3 individuals each) and 2 individuals of *P. pulcher* were also used in targeting approximately 655 bp of the CO-I gene from mitochondrial DNA with primer pairs FishF2-5' TCGACTAATCATAAAGATATCGGCAC 3' and FishR2-5' ACTTCAGGGTGACCGAAGAATCAGAA 3' (Appendix IV.1) (Ward et al. 2005). PCR was performed in 20 µl final volumes using Phusion High-fidelity DNA Polymerase from New England Biolabs. Each reaction contained 4 µl 5x Phusion HF buffer, 0.4 µl 10 mM dNTPs, 1 µl 10 µM FishF2 primer, 1 µl 10 µM FishR2 primer, 12.35 µl nuclease-free water, 0.25 µl Phusion DNA polymerase (2000 units/ml) and 1 µl DNA template (c. 50 ng). The amplification conditions were: initial cycle of 98°C for 30 s followed by 33 cycles of 98°C for 10 sec, 59°C for 30 sec, 72°C for 30 sec, 72°C for 30 sec and final extension at 72°C for 10 mins. The amplification products were purified by spin column following the manufacturer's instructions (QIAquick PCR Purification kit). The purified samples were commercially sequenced (Sanger sequencing, GATC Biotech Ltd.). CO-I sequences from seven tilapia species and *P. pulcher* were aligned and a gene tree constructed using *RAxML* and visualized using *FigTree*.

4.3.10 Data Access

All species names used are in accordance with The Catalogue of Life (Roskov et al., 2014). The raw sequence data from this study have been submitted to the EBI Sequence Read Archive (SRA) study ERP006545.

4.4 Results

4.4.1 RAD Sequencing

In total, 347,145,626 raw reads were produced (173,572,813 paired-end reads, EBI SRA study ERP006545) from two batches of RADseq libraries construction. After removing low quality sequences (quality score <30), ambiguous barcodes and orphaned paired-end reads, 87.26% of the raw reads were retained (302,904,154 reads), with analysis being performed on Read 1 sequence data (151,452,077 reads). In total the stacks analysis identified 51,750 unique RAD-tags (i.e. the total number of loci across all species, with overlapping subsets of loci among species) (Figure 4.2). The number of reads and RAD-tags for each sample are reported in Appendix IV.2.

4.4.2 SNP-based phylogenetic tree

1,613 SNP in 1,002 RAD loci were identified that were common to all individuals across 8 species based on reference based analyses (RBA). Phylogenetic analysis using these markers (Figure 4.3, A) showed *P. pulcher* furthest from all of tilapia species, appropriate as the out-group species, followed by *T. zilli*, then *S. galileus*, with the *Oreochromis* species clustered more closely together. The probability values across the branches (approximately between 0.99/96 and 1/100: CAT/Gamma model) gave the highest level of confidence for species discrimination.

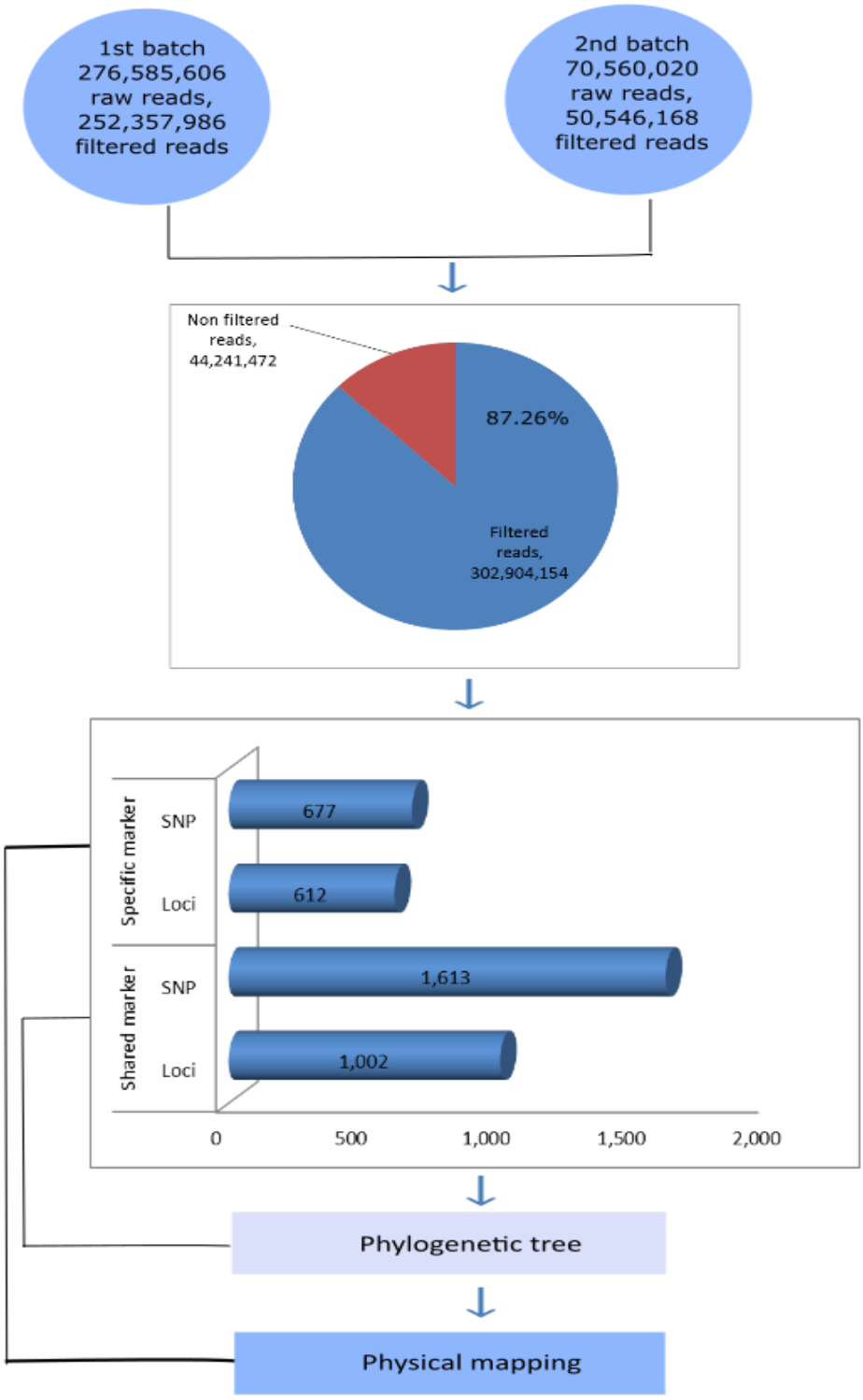


Figure 4.2 The number of retained reads, shared loci and specific SNP maker among 7 tilapia species.

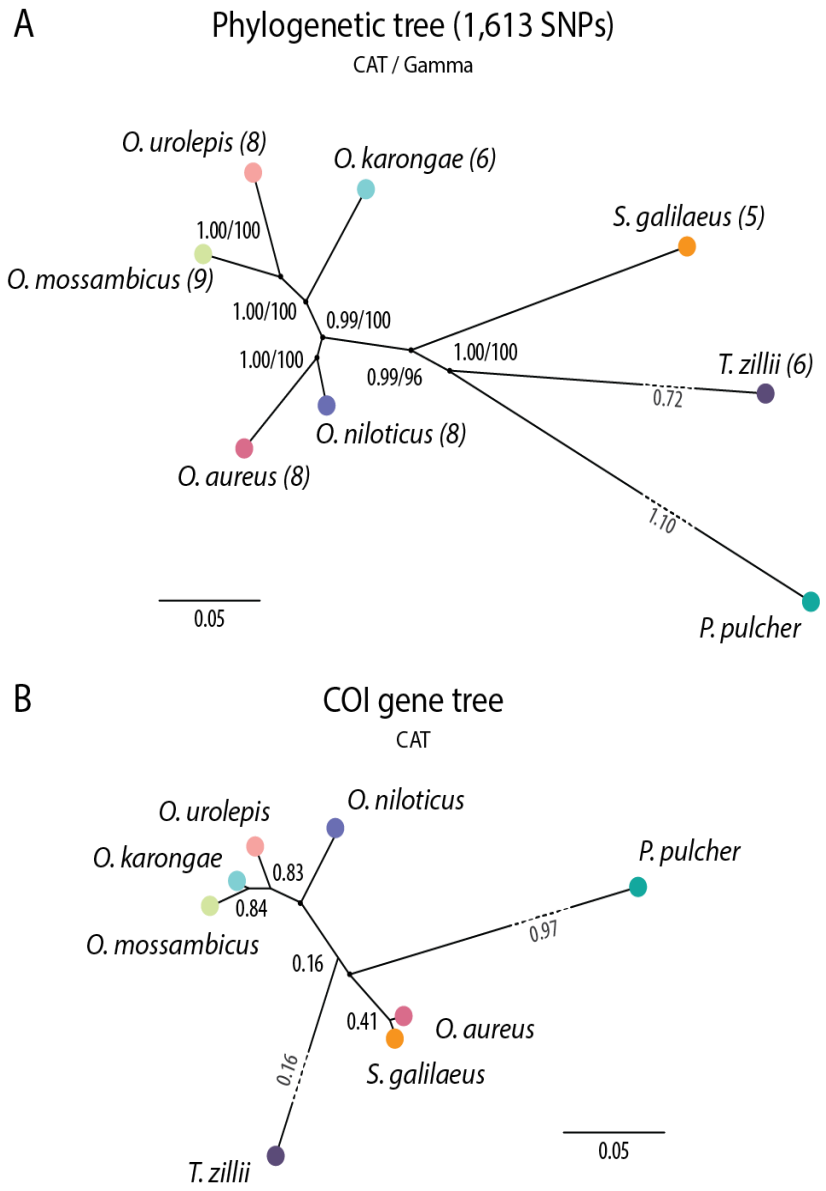


Figure 4.3 Trees of tilapia species inferred from common marker retrieved from reference, and rooted to *P. pulcher* as out-group.
(A) Phylogenetic tree. Best-scoring ML tree with support values are from 1000 bootstrap replicates with values along branches reported as Bayesian posterior probabilities for GTR-CAT model (RAxML and PhyloBase) and GTR-Gamma model (RAxML). *T. zillii* and *P. pulcher* real branches were shortened for visual purpose; real distance is indicated below the branches. (B) Gene tree of tilapia species derived from Cytochrome c oxidase subunit I sequences. Best-scoring ML tree with support values are from 1000 bootstrap replicates with values along branches reported as Bayesian posterior probabilities for GTR-CAT model (RAxML).

4.4.3 Species-specific SNP markers and physical map

In total, 677 species-specific SNP markers (i.e. with an allele that was unique to one species) were found across the seven tilapia species using bespoke based on analysis of the 1,613 SNP showing fixed differences among species). The *Oreochromis* species had 24-43 species-specific SNPs, with 106 for *S. galilaeus* and 400 for *T. zillii* (Table 4.1), most likely reflecting that there were five species from within the *Oreochromis* genus in the dataset and only one each from the other two genera.

The species-specific markers were mapped in the *O. niloticus* genome to determine their distribution across the linkage groups. The markers span the 22 linkage groups. LG 7, 15, 16-21, and 20 contained at least one diagnostic marker for all species. LG 9 exhibited the highest density of 1.53 markers/Mb, while LG 3 contained the fewest species-specific markers (9, or 1.3% of the total; 0.47 SNP per Mb). As an example, Figure 4.4 illustrates the physical map of LG 20 and the distribution of SNP markers (two for *O. niloticus*, four for *O. aureus*, three for *O. karongae*, three for *O. mossambicus*, one for *O. urolepis*, eight for *S. galilaeus* and 16 for *T. zillii*). The complete mapping of species-specific markers can be seen in Appendix IV.3.

Table 4.1 SNP-based diagnostic species-specific markers for tilapia.

LG	Physical size (bp)	*Oau*	*Oka*	*Omo*	*Oni*	*Our*	*Sga*	*Tzi*	Total	Number of SNP/Mb
1	31,194,787	2	0	3	3	1	4	34	47	1.51
2	25,048,291	0	2	1	0	0	0	22	25	1.00
3	19,325,363	0	0	0	0	0	3	6	9	0.47
4	28,679,955	6	1	0	1	2	7	16	33	1.15
5	37,389,089	2	6	1	0	0	3	25	37	0.99
6	36,725,243	3	2	0	1	0	7	29	42	1.14
7	51,042,256	2	1	3	1	3	11	22	43	0.84
8-24	29,447,820	1	1	1	0	3	6	18	30	1.02
9	20,956,653	3	2	1	0	0	5	21	32	1.53
10	25,048,291	0	0	0	2	1	3	8	14	0.56
11	33,447,472	2	5	1	3	0	3	18	32	0.96
12	34,679,706	0	2	1	1	3	10	29	46	1.33
13	32,787,261	3	3	0	0	3	5	19	33	1.01
14	34,191,023	2	0	6	0	5	6	19	38	1.11
15	26,684,556	1	2	1	2	3	1	10	20	0.75
16-21	34,890,008	3	3	1	3	1	4	15	30	0.86
17	31,749,960	1	4	0	0	3	2	21	31	0.98
18	26,198,306	0	2	2	0	2	8	9	23	0.88
19	27,159,252	1	2	0	1	2	4	12	22	0.81
20	31,470,686	4	3	3	2	1	8	16	37	1.18
22	26,410,405	0	1	5	2	2	6	18	34	1.29
23	20,779,993	0	1	0	2	3	0	13	19	0.91
Total	665,306,376	36	43	30	24	38	106	400	677	1.02

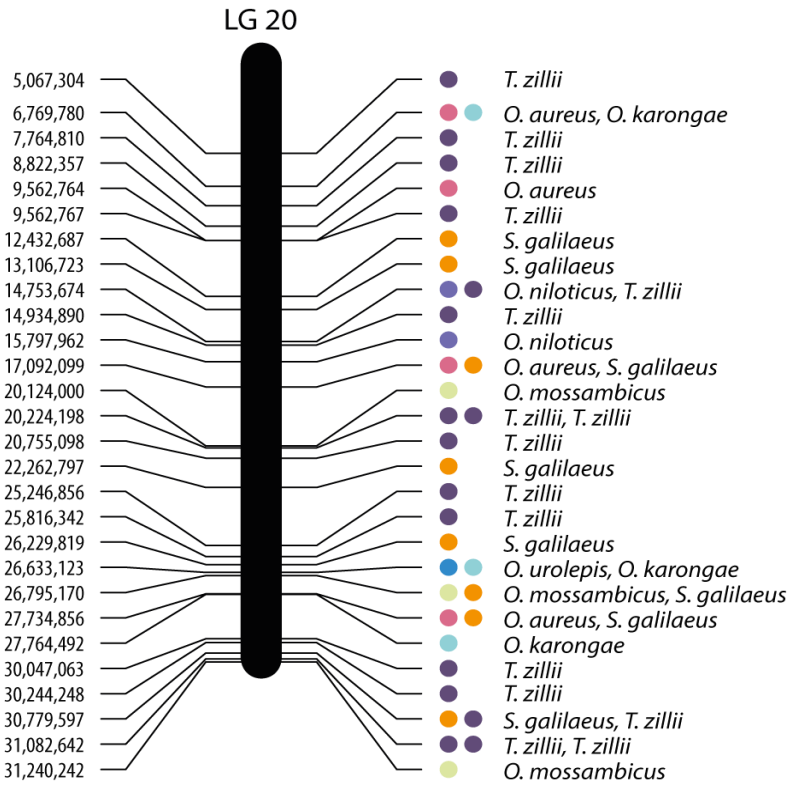


Figure 4.4 The diagnostic SNP markers in LG 20 of tilapia species. Numbers on the left side show the SNP position (bp), while those on the right side denote the name of the species for which an allele of the SNP is unique.

4.4.4 COI

The length of the COI partial sequences that were retrieved varied between 395-631 bp in *P. pulcher* and 477-689 bp for the tilapia species. The COI sequences of tilapia species and *P. pulcher* in the present study agreed with those in the Barcode of Life Data System (BOLD) and the NCBI GenBank Database, with a single exception. A sequence labelled "*O. niloticus* GIFT" in NCBI (GU477624.1) matched the three *O. karongae* from our study. However, the COI gene tree indicated a genuine discrimination between *O. karongae* and

O. niloticus (Figure 4.3, b), and all other samples labelled *O. niloticus* in BOLD and NCBI matched the *O. niloticus* in our study. In contrast to the SNP-based phylogenetic tree, the tree based on COI partial sequence placed *O. aureus* close to *S. galilaeus*, apart from the other *Oreochromis* species.

4.5 Discussion

4.5.1 RAD marker and quality of sequence data

In general, the level of retained reads (87.26%) between the two subsets of RAD sequencing was high enough to produce reliable SNP markers for this study. Standardization of DNA was a crucial step in making DNA libraries with samples from several different species. Balancing the DNA quantities from different samples influences sampling depth (Andolfatto et al., 2011). The variation in filtered reads between samples might arise not only from genomic attributes, for instance the number of loci identified, the level of polymorphism and divergence among species, but also the factors that interact with sequencing characteristics such as the quality of DNA and degree of sample multiplexing, the total numbers and length of reads, and the sequencing error rate (Catchen et al., 2011).

In this study, RAD sequencing identified 51,750 unique RAD-tags, resulted 1,613 shared SNP marker at 1,002 loci across 7 tilapia species plus species outgroup *Pelvicachromis pulcher*. Another technique that has been used to generate large numbers of SNP in tilapias is reduced representation library (RRL), which generated 3,569 markers in the widely cultured Genetically Improved Farmed Tilapia (GIFT) strain (Van Bers et al., 2012). In comparison to

RADseq, the RRL method has an advantage in minimizing repetitive elements in the sequences without reducing SNP distribution over the entire genome (Du et al., 2012; Sánchez et al., 2009) and more effective removal of putative paralogous variants (Houston et al., 2014). However, SNP genotyping from RRL (or any defined set of SNPs) becomes less efficient when applied to species other than that from which the SNP set was isolated. For instance, the observed low percentage of polymorphic SNP in *O. aureus* and *O. mossambicus* is likely to have been influenced by the SNPs being originally isolated in *O. niloticus* (Van Bers et al., 2012). In contrast, the shared SNP marker identified in this study involved multiple species.

Even though RAD–seq is a highly valuable tool for SNP discovery and genotyping, there are some limitations of this technique. Read depths at RAD loci are highly variable because of bias in restriction fragment length caused by incomplete shearing of shorter restriction fragments (Davey et al., 2013). Furthermore, sampling based on restriction digestion may introduce a bias in allele frequency estimation (Arnold et al., 2013). There is a fundamental limitation found both in the RADseq and RRL because of the steep drop in the number of loci shared across distant phylogenetic scales (Lemmon & Lemmon, 2013). Therefore, when applying these techniques to make phylogenetic and/or population genetic inferences, it is important to consider the specific goals of the study. For example, in SNP discovery, one may prefer to select an enzyme with an AT-rich recognition sequence; conversely, for distinguishing divergent populations, GC-rich recognition sequences will generally access a higher proportion of conserved regions of the genome and may increase overlap in

sampled loci between populations (Arnold et al., 2013).

4.5.2 Phylogenetic tree

A phylogenetic tree based on common SNP markers (Figure 4.3) showed the *P. pulcher* as the most distant, as expected, while *T. zillii* and *S. galilaeus* were the most divergent among the tilapias (CAT/Gamma value = 0.99/96). The species of the *Oreochromis* genus were more closely related in comparison to *S. galilaeus* and *T. zillii*. There was a significant distance between *O. karongae*, *O. urolepis* and *O. mossambicus* (CAT/Gamma value = 1.00/100), meanwhile *O. niloticus* and *O. aureus* were closer but still significantly distinct (CAT/Gamma value = 1.00/100). This result was in agreement with earlier phylogenetic trees derived from allozymes (Pouyaud & Agnese, 1995) and mtDNA (Nagl et al., 2001). Twenty-four enzyme loci indicated a clustering of tilapia species according to their genera (*Tilapia*, *Sarotherodon*, *Oreochromis*) with the exception to *S. melanotheron* which was closer to species of the genus *Oreochromis* than to those of *Sarotherodon* (Pouyaud & Agnese, 1995); likewise the sequences of mitochondrial control region gave a phylogeny in concordance to the systematics proposed by Trewavas (Nagl et al., 2001). The current study supports earlier research in concluding that mouthbrooding in tilapias evolved only once from a single substrate-brooding *Tilapia*-like ancestor, while the mouthbrooding branch itself subsequently divided into two with an independent evolution of biparental and paternal mouthbrooders on the one hand and maternal mouthbrooders on the other hand (Trewavas, 1983). This was supported with partial tilapiine phylogenies (Dunz & Schliewen, 2013) and a family wide reconstruction of the

transitions in parental care behaviour (Goodwin et al., 1998), in contrast to the multiple speciation hypothesis of Peters & Berns (Peters & Berns, 1982). In addition, a novel phylogeny of tilapiine cichlid fish on the basis of complete mitochondrial NADH dehydrogenase subunit 2 (*ND2*) gene sequence indicated several independent origins of derived mouthbrooding behaviors in the family Cichlidae (Klett & Meyer, 2002).

The COI sequence did succeed in separating the species studied here, but could not resolve certain species relationships, for instance *O. aureus* clusters with *S. galilaeus* instead of with other members of the *Oreochromis* genus, in this and earlier (Shirak et al., 2009) studies. One marker cannot be as powerful for phylogenetic studies as a large number of SNP, such as the set used here for the SNP-based phylogeny. When only a single gene, for instance COI, is used, then what is generated is a gene tree, rather than the broader evolutionary history of a group of species (Rubin, Ree, & Moreau, 2012) .

In contrast, constructing phylogeny from RADseq data from many loci across the entire genome is generally high accurate across a wide range of clustering and filtering parameters (Rubin et al., 2012). In addition, RADseq data gives promising results for reconstructing phylogenetic relationship only in younger clades in which sufficient numbers of orthologous restriction sites are retained across species. Further study indicated that high-throughput sequencing of RAD tags capable to resolve fine-scale genetic divergence among intraspecific population that have been separated for less than 20,000y (Emerson et al., 2010).

4.5.3 Species-Specific SNP Markers and their Genomic Distribution

RADseq analysis retrieved SNP markers that were diagnostic across the samples from the seven tilapia species included in the study (24-400 such markers per species), supported by the known history of the populations concerned (pure as far as could be determined) and COI analysis. While the study demonstrated the power of RADseq analysis to isolate such markers, each species was represented by only a limited number of fish from a single population. In consequence, more extensive analysis of a wider variety of species and multiple populations per species (representing the diversity of each species) would be required to demonstrate that robust species-specific markers can be isolated in this way for tilapias.

Given the number of potentially useful species-specific SNP markers isolated (hundreds in total), consideration should be given to the most cost-efficient way of marker discovery and genotyping in future studies of this kind. Genotyping of hundreds of SNP for hundreds of individuals using individual SNP assays would be laborious. At present, SNP chips are not available for tilapia, and would be costly if developed. RADseq allows isolation and genotyping of a few thousand SNP at a cost of approximately £100 per individual, comparable to many SNP chips. A variation on RADseq, double-digest RADseq (Peterson et al., 2011), typically reduces the number of SNP per individual by approximately five-fold, but allows more individuals to be sequenced and reduces the cost per individual to around £20.

Mapping species-specific SNP markers to the reference genome resulted in the loss of many loci/SNP from the initial DBA. The density of the species-

specific SNP markers ranged from 0.47 SNPs/Mb in LG 3 to 1.53 SNPs/Mb in LG 9. The physical chromosome that is equivalent to LG3 is the largest, but in terms of the genome assembly and the number of SNP markers mapped in LG3 in the current study, it is the smallest and had the lowest number of SNP markers. It is likely that this disparity is due to the high density repetitive DNA in this chromosome (Cnaani et al., 2008; Harvey et al., 2003). Despite the largest in the genome assembly, LG7 indicated the third biggest number of SNP markers.

Analysis of tilapia karyotypes indicated that the 22 pairs of chromosomes seen in most tilapia were derived from the 23 pairs in the Cichlid ancestor by a fusion which led to the biggest chromosome pair, LG3 (Mota-Velasco et al., 2010). The nucleotide sequencing of sub cloned DNA segments of BAC-C4E09 identified the repeated DNA sequences CiLINE2 transposon in LG3 of *O. niloticus* (Oliveira et al., 1998). Another study also indicated that the largest chromosome of *O. niloticus* evidenced labelling signals of eight BACs of LG3 and the LG3 mapping confirmed the presence of a lot of repetitive DNA in the end of the largest chromosome (Mazzuchelli et al., 2012). This LG has high heterochromatin (Guyon et al., 2012). LG 7, 15, 16-21 and 20 contained at least one SNP marker for each of the 7 tilapia species. The relatively even spread of these SNP markers across the *O. niloticus* genome suggests that RADseq can perform well to increase the resolution of population structure and phylogenetic relationships in tilapias and presumably in other species groups (Rubin et al., 2012).

4.6 Conclusion

In summary, RADseq analysis retrieving 677 species-specific SNP markers from seven tilapia species each represented by a sample from a single population. These markers were distributed fairly evenly across the genome, from 0.47 SNP/Mb in linkage group 3 to 1.53 SNP/Mb in linkage group 9. This large number of markers suggests that further studies including more species and populations should identify robust species-specific markers for this group of fish, and that RADseq is a suitable technology for both isolation and genotyping of such markers for a variety of applications requiring such markers.

4.7 Acknowledgments

We thank staff at Edinburgh Genomics Facility, especially Urmi Trivedi and Marian Thomson, for assistance with RAD library sequencing, and Keith Ranson of the Tropical Aquarium Facility at the Institute of Aquaculture, University of Stirling, for help in rearing fish. The authors acknowledge the support of the MASTS pooling initiative (The Marine Alliance for Science and Technology for Scotland). MASTS is funded by the Scottish Funding Council (grant reference HR09011) and contributing institutions. We also thank to Director General of Higher Education, Ministry of Research, Technology and Higher Education (*Kemenristek Dikti*), Indonesia for funding PhD scholarship of MS at the University of Stirling.

5. IDENTIFICATION OF SPECIES-SPECIFIC SNP MARKERS IN TILAPIA USING DOUBLE-DIGEST RAD SEQUENCING (DDRADSEQ)

Mochamad Syaifudin [a, b], Michaël Bekaert [a], John B. Taggart [a], Christos Palaiokostas [a], Gideon Hulata [c], Helena D'Cotta [d], Jean-Francois Baroiller [d], David J. Penman [a, #], Brendan McAndrew [a]

[a] Institute of Aquaculture, University of Stirling, Stirling FK9 4LA, Scotland, UK.
[b] Program Study of Aquaculture, Agriculture Faculty, University of Sriwijaya, Indonesia.
[c] Institute of Animal Science, Agricultural Research Organization, The Volcani Center, Israel
[d] Cirad, Persyst, Umr Intrepid, Campus International De Baillarguet, Montpellier, France

[#] Corresponding author: Tel +44 (0)1786 467901; Fax +44 (0)1786 472133; Email d.j.penman@stir.ac.uk

Contribution: The first draft of the present manuscript was compiled and written in full by the author of this thesis, who was also fully involved in all subsequent editing. The DNA extraction, RAD library preparation (under John B. Taggart's assistance), PCR, Cytochrome c oxidase subunit I sequence analysis, phylogenetic tree reconstruction, and physical mapping of species-specific diagnostic SNP markers were conducted by the candidate. The other co-authors contributed towards the experimental design, the samples analysis of sequenced reads derived from ddRADseq, sequence alignment, SNP position across the reference genome – *O. niloticus* and editing.

Keywords: Tilapia, population, SNP markers, ddRADseq

Abbreviations: ddRAD: double digested restriction-site associated DNA; SNP: single nucleotide polymorphism; DBA: de novo-based analysis; RBA: reference-based analysis; LG: linkage group.

5.1 Abstract

Background

A previous study using standard Restriction-site Associated DNA sequencing (RADseq) enabled us to retrieve hundreds of SNP markers for each species, however the number of samples that can be used in a single sequencing lane and retain this coverage is restricting for large surveys. Recently it became possible to sequence many species with hundreds of samples at reasonable cost, but with fewer markers, using double digest Restriction-site Associated DNA (ddRADseq). We aimed to test the potential of ddRADseq for discovering SNP markers to distinguish between 10 tilapia species, including 2 sub-species, and to analyse the distribution of such markers in the genome.

Results

Analysis of ddRAD sequencing data detected 1,358 SNP (all shared RAD loci) in the *de novo* based analysis (DBA) and 938 SNP (all shared RAD loci) in the reference based analysis (RBA) among 10 tilapia species. A phylogenetic tree based on the two analyses indicated very similar patterns. Further analysis in the RBA using in house *perl scripts* ascertained 38 species-specific SNP markers (i.e. with an allele unique to that single species) as follows: *Oreochromis aureus* (1), *O. karongae* (2), *O. macrochir* (1),

Sarotherodon galilaeus (3), *S. melanotheron* (5) and *Tilapia zillii* (26), but no species specific markers were obtained for *O. mossambicus*, *O. niloticus*, *O. urolepis hornorum* or *O. andersonii*. A larger set of diagnostic SNP markers was identified (37 SNP at 35 loci) from a subset of four economically important species which are often involved in hybridization in aquaculture: *O. niloticus* (7 SNPs), *O. aureus* (13), *O. mossambicus* (10) and *O. u. hornorum* (7). The analysis also identified three species-specific SNP markers between sub-species *O. niloticus niloticus* and *O. niloticus cancellatus*. Physical mapping of the species specific markers, where 26 out of 38 are for *T. zillii* onto the *O. niloticus* genome (Orenil1.1, NCBI assembly GCF_000188245.2) showed that the diagnostic markers were distributed evenly across chromosomes in the genome.

Conclusions

38 Species-specific SNP markers identified across 10 tilapia species were distributed evenly across the genome. A large number of SNP markers were obtained in a subset of four commercial tilapia species with many more SNPs identified between species pairs. These would be beneficial for investigating hybridization and introgression, not only in the species but also in the sub-species levels in tilapia.

5.2 Introduction

The tilapias are a group of African and Middle Eastern cichlid fish, which are now widely cultured across 140 countries. Recent natural species distributions are the result of many processes, including population demography, phylogeographic history, behaviour, physiological tolerances, competition, response to human land use change and adaptation to the environment (Gaston, 2003). The exploitation of wild tilapia and habitat destruction are still commonly occurring across their natural habitat. The genetic resources of many farmed strains have been poorly managed and there have been a large number of serial introductions from a relatively small number of commercial strains, often hybridized (McAndrew, 1993), rather than the use of local wild species in many countries. Distinguishing tilapia species, hybrids and introgressed populations is of utmost importance for both farmed and wild populations.

Introgression has occured among wild tilapia species, e.g. *T. zillii* and *T. guineensis* following damming of a river in the Ivory Coast to form the man made lake Ayame (Adépo-Gourène et al., 2006). Introgression has been observed between introduced *O. niloticus* and native *Oreochromis* spp., e.g. *O. esculentus* in L. Victoria, eventually leading to loss of native species (Angienda et al., 2011). One study also reported a high degree of mixing between *O. mossambicus* and *O. niloticus* in Southern Sri Lanka – both introduced, outside of their native ranges (De Silva & Ranasinghe, 1989).

Where a mixture of tilapia species has been stocked, reproductively viable hybrids have often resulted, making the use of external morphometrics for

hybrid or species determination difficult (Wohlfarth & Hulata, 1983). The identification of many tilapia species, both wild and farmed, has become difficult with the extensive introduction of alien species and transfer of native forms outside of their natural ranges, therefore, genetic markers offer a method to discern the presence/absence of distinct tilapia species, the genetic composition of established, feral hybrids in new environments, and the composition of mixed species in culture (Costa-Pierce, 2003). Different marker technologies have been applied as tools for species identification, such as allozymes (McAndrew & Majumdar, 1983; Sodsuk & McAndrew, 1991; Sodsuk et al., 1995; B-Rao & Majumdar, 1998), microsatellite markers (Agnése & Adépo-Gourène, 1999; Nyingi et al., 2009), Randomly Amplified Polymorphic DNA (RAPD) (Bardakci & Skibinski, 1994; Dinesh et al., 1996; Hassanien et al., 2004) and restriction fragment polymorphisms in ribosomal DNA (El-Serafy et al., 2007; Toniato et al., 2010). To date no study has used single nucleotide polymorphism difference to identify tilapia species at the sub-species and/or population level.

Currently, improvements in NGS technologies now make it possible to carry out genotyping-by-sequencing (GBS) in many species using many individuals, whether the reference genome is known or unknown (*de novo*). Various studies have illustrated some of the differences between GBS methods, in particular, aligning paired-end reads to achieve longer consensus sequences in contrast to single-end reads with shorter alignments, and double digest versus sonication methods to fragment DNA (Campbell et al., 2012). One method that allows a maximum of 50 samples to be analysed in a run and generates

thousands of SNP markers is Restriction-site Associated DNA sequencing (RAD-seq). GBS methods using restriction enzymes are simpler and less expensive in comparison to other reduced representation methods. However elimination of size selection steps results in libraries of more variable fragment size, where tag and single nucleotide polymorphism (SNP) counts are limited by number of cut sites rather than by read number (Hamblin & Rabbi, 2014). In a previous study, as preliminary result, we used RAD-seq to infer phylogenetic relationships and identify SNP markers from seven tilapia species, which generated 24 to 400 for each species (Chapter 4). Despite its advantages in discovering and genotyping hundred or thousands of SNPs at reasonable cost (Baird et al., 2008; Davey et al., 2011), the limited number of samples (individuals) that can be used in a single sequencing lane using RAD-seq becomes a drawback.

A new technique, named double digest Restriction-site Associated DNA sequencing (ddRADseq), was developed by eliminating random shearing and explicitly using size selection to recover a number of regions across individuals (Peterson et al., 2012). Digesting a genome with two REs can generate a wide range of fragment sizes, depending on the frequency of the enzyme recognition sites. Enzymes with short (4 or 6bp) recognition sequences will cut frequently, therefore generating the potential to sample a larger fraction of the genome than enzymes cleaving at less abundant (8bp) sequences. Generally, pairs of enzymes with common cut sites can be used to sample many loci in each individual, but in fewer individuals; by contrast a pair of enzymes with rare sites will generate fewer loci but will allow a greater number of

individuals to be analysed at the same cost. Very large genomes e.g mammals (De Donato, 2013), requires use of an enzyme that cuts infrequently, while for high levels of multiplexing, combining a 6-cutter with a 4-cutter give the best results (Poland et al., 2012) to limit the fragment number and reduce fragment size (Hamblin & Rabbi, 2014). Theoretically, ddRADseq reduces the bias of fragment length coverage, however statistically, the effects of restriction-site polymorphism on summary statistics are more prominent with this method (Arnold et al., 2013). Therefore, when we have many samples from different species and populations, the advantages of ddRADseq is that it allows more individuals to be sequenced at fewer, but still informative, sequence stacks. An added advantage is that this technique will effectively reduce the cost of such an analysis per individual to around £20.

The objective of the research described here was to test the potential of ddRADseq for discovering SNP markers to distinguish between 10 tilapia species (including sub-species and different populations where available) and analyse the distribution of such markers in the genome. In parallel, DNA sequence of the mtDNA COI gene was also analysed, as a reference for comparison to ddRADseq. The conserved sequence of the 5' region of the mitochondrial gene cytochrome oxidase subunit I (COI or Cox1), a platform for the universal DNA barcoding of life, used for distinguishing tilapia species (Shirak et al., 2009; Wu & Yang, 2012).

5.3 Materials and Methods

5.3.1 Ethics statement

All working procedures complied with the UK Animals Scientific Procedures Act (Parliament of the United Kingdom 1986)

5.3.2 Biological Materials

Fin samples were collected from ten different tilapia species (4 to 13 individuals per species or sub species). The *Oreochromis niloticus* samples consisted of two sub-species (*O. n. niloticus* and *O. n. cancellatus*) from three populations in each case; *Oreochromis aureus*, *O. mossambicus* and *Tilapia zillii* (Gervais) comprised samples from two populations each, while *O. karongae* (Trewavas), *O. urolepis hornorum* (Norman; originally from Tanzania), *O. andersonii*, *O. macrochir*, *Sarotherodon galilaeus* (Linnaeus) and *S. melanotheron* consisted of one population each. Samples (fin tissue) were stored in 99% ethanol at -20°C until required. Details of samples and origins for the three batches of libraries are listed in Table 5.1.

Table 5.1 Species, strain, geographic origin and number used for three batchs of libraries in ddRAD sequencing.

No	Species/sub species	Strain/ Population	Origin	n
1.	*O. niloticus*			
	a. niloticus	a. Stirling	L. Manzala, Egypt	6
		b. Kpandu	Ghana	12
		c. Nyinuto	Ghana	12
	b. *cancellatus*	a. Hora	Ethiopia	13
		b. Koka	Ethiopia	12
		c. Metahara	Ethiopia	8
	Sub total 1			**63**
2.	*O. mossambicus*	a. Stirling	Zimbabwe	5
		b. Natal	South Africa	10
	Sub total 2			**15**
3.	*O. aureus*	a. Stirling	L. Manzala, Egypt	5
		b. Ain Faskha	Israel	10
	Sub total 3			**15**
4.	*O. karongae*	Stirling	L. Malawi, Tanzania	5
5.	*O. u. hornorum*	Israel	Israel	5
6.	*T. zillii*	a. Stirling	L. Manzala, Egypt	5
		b. Ghana	Ghana	5
	Sub total 6			**10**
7.	*S. galilaeus*	Israel	Israel	5
8.	*O. andersonii*	Itezhi-tezhi	Zambia	6
9.	*O. macrochir*	Itezhi-tezhi	Zambia	4
10	*S. melanotheron*	Ghana		4
	Total samples			**132**

5.3.3 Genomic DNA Extraction

Total genomic DNA was extracted using the Realpure Genomic DNA Extraction Kit (Durviz S.L) following the manufacturer's protocol. An RNAse incubation step was included to minimise RNA contamination, with each precipitated DNA sample being finally resuspended in 5 mmol/L Tris, pH8.5. Extracted DNA was quantified by spectrometry (Nanodrop ND 1000 Spectrophotometer, NanoDrop Technologies Inc., Montchanin, DE). Both A 260/280 and 260/230 ratios were > 1.8 for all samples. Sample integrity was

checked by agarose gel (0.8%) electrophoresis. Those samples that passed quality control (no observable RNA and comprising predominantly high molecular weight DNA) were selected for use and diluted to a concentration of 50 ng/µL in 5 mM Tris; pH 8.5.

5.3.4 ddRAD library preparation and sequencing

Initially one ddRAD library was constructed from seven tilapia species with 36 individually barcoded fish (four replications of each sample), followed by a second library consisted of 132 individuals from a wider range of species and populations. Later a third library comprising 52 individual derived from pooling several low quality DNA samples in the second library was made. The diagram of ddRAD sequencing experiments and genotyping RAD alleles is described in Figure 5.1. Details of species composition of the three libraries are given in Appendix V.1. The ddRAD library preparation protocol followed Palaiokostas et al. (2015), a modified version of the methodology described by Peterson et al. (2012). High quality genomic DNA with a concentration approximately 21 ng/µl based on fluorometry was digested using restriction enzymes *Sbf*I (recognizing the CCTGCA⏐GG) and *Sph*I (recognizing the GCATG⏐C motif). In a 96 well plate format a 6 µl reaction volume containing 3 µl (21ng) DNA, 0.6 µl 10x CutSmart Buffer, 0.010 µl 10 units (U)/µg *Sbf*I, 0.010 µl 10 units (U)/µg *Sph*I and 2.380 µl double distilled water (ddH_2O) was mixed well, incubated for 40 min at 37 ^{0}C, then cooled to room temperature.

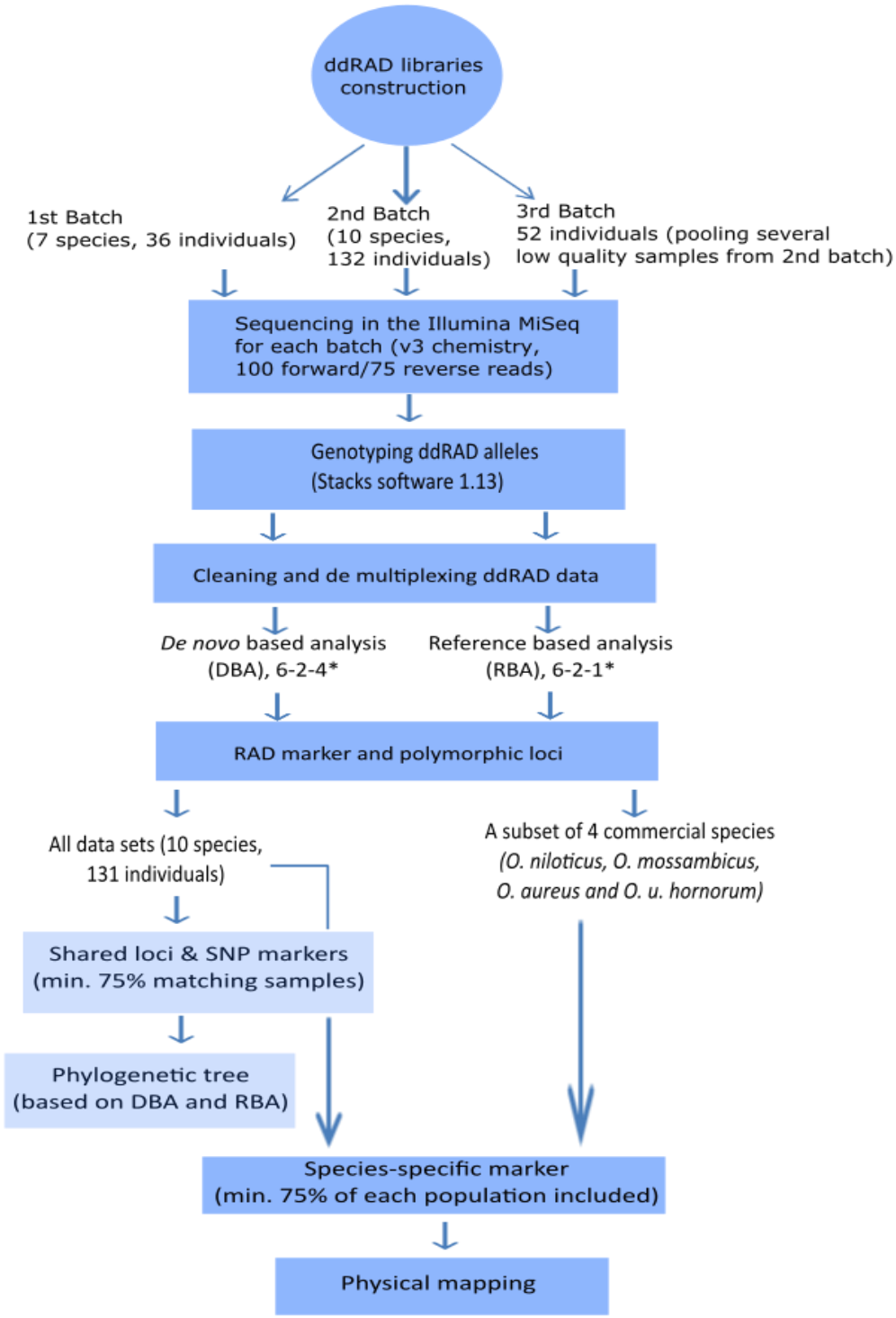


Figure 5.1 Flow diagram of ddRAD sequencing experiment and genotyping ddRAD alleles.

* Indicating the parameters applied in the Stacks analyses, which are m: minimum stack depth; M : the number nucleotide mismatches allowed between stacks; n : the number nucleotide mismatches allowed between catalog loci.

A 3µl aliquot of barcode mix (*Sbf*I:*Sph*I 1:10), an individual specific combination of P1 (25 nM) and P2 adapter (100 nM), each with unique 5 or 7 bp barcode, was added to the digested DNA, then ligated by adding 0.3µl 1x CutSmart Buffer, 0.120 µl rATP (1mM), 0.020 µl T4 Ligase (2 K ceU/ug), and 2.560 µl ddH_2O, which were mixed well, and incubated in thermocycler (heated lid off) for 2 hours 30 min at 22^0C. The ligation reactions were heat inactivated by incubating at 65^0C for 20 minutes in thermocycler (heated lid set to 70^0C, briefly spin down the plates, then combined in a single pool (for one sequencing lane) and purified using MinElute PCR purification kit (see chapter 2). Size selection (320-590 bp) was performed by agarose gel separation and followed by gel purification with Qiagen MinElute Gel Clean Up and PCR amplification (see chapter 2). Initial PCR was conducted to determine optimal PCR condition, while bulk PCR amplification was carried out with at least 350 µl of amplified library (see chapter 2). A total of 100 µl of the amplified libraries (13-14 cycles) was purified using an equal volume of AMPure beads. The libraries were sequenced in three different lanes in house at the University of Stirling using two runs of an Illumina Miseq (v3 chemistry, 100 forward / 75 reverse reads).

5.3.5 Genotyping RAD Allele

Reads of low quality (QC values under 30), missing the expected restriction site or with ambiguous barcodes were discarded. Retained reads were sorted into loci using either a *de novo* based analysis (DBA; all loci included) or a reference-based analysis (RBA) (including only loci found in the Broad

Institute of MIT and Harvard *O. niloticus* genome assembly Orenil1.1, NCBI assembly GCF_000188245.2 (Brawand et al., 2014) and genotyped using Stacks software 1.13 (Catchen et al., 2013). 131 out of 132 samples were used for further analyses due to very low quality read in one *O. andersonii* (Md100). The stack parameters applied in the DBA were 6-2-4, i.e. a value of m (minimum stack depth), M (the number nucleotide mismatches allowed between stacks) and n (the number nucleotide mismatches allowed between catalog loci) in the DBA, while parameters 6-2-1 were used in the RBA. The reference-based analysis was used to eliminate potential contamination from human or bacterial DNA and to know the position of SNP markers in the reference genome. The likelihood-based SNP-calling algorithm (Hohenlohe et al., 2010) implemented in Stacks evaluates each nucleotide position in every RAD-tag of all individuals, thereby statistically differentiating true SNPs from sequencing errors. Polymorphic RAD-tags may contain more than one SNP, but the vast majority (over 99%) showed only two allelic versions; RAD-tags with more than two alleles or shared by less than 75% of the samples were excluded. RAD loci that were shared among all tilapia species in the DBA and RBA, exhibited no intraspecific polymorphism but showed interspecific polymorphism were identified using *find_pattern.pl* (a bespoke Perl script to find fixed allele paterns) with grouping between individuals. The same analysis was used to retrieve species-specific markers with only one fixed allele in a given species and different allele in other species. The all shared RAD loci are defined as a minimum of 75% of matching samples with a maximum of 2 SNP per locus analysed. While, species-specific marker is defined as one fixed allele only between species with at least 75% loci

present in each species. In addition, a subset population from four commercial tilapia species (*O. niloticus*, *O. mossambicus*, *O. aureus* and *O. u. hornorum*) was created from the dataset to generate species-specific SNP markers among these species. Data from the RBA was used for retrieving species-specific markers to ensure consistency of homologous loci across all species.

5.3.6 Phylogenetic Reconstruction

The phylogeny of the tilapia species was inferred from DBA (all 1,358 SNP markers) and RBA (all 938 SNP markers), and rooted to *T. zillii*. All SNPs were concatenated and trees were constructed using RAxML v8 (Stamatakis, 2014). This analysis allows Best-scoring ML tree with support values from 1000 bootstrap replicates with values along branches reported as Bayesian posterior probabilities using GTR+CAT and GTR-Gamma using RAxML.

5.3.7 Physical Mapping

The shared SNP markers exhibiting interspecific polymorphism but no intraspecific polymorphism and diagnostic of one species only (*e.g., O. niloticus* exhibiting a 'C' allele of a SNP and all other species a 'G') are reported with the SNP marker location extracted from the alignment against the *O. niloticus* genome then visualised using Genetic-Mapper v0.6 (Bekaert, 2014).

5.3.8 COI DNA Barcoding

The DNA from 10 tilapia species, involving 1-3 populations per species (2-3 individuals each population), were also used in targeting approximately 655 bp of the CO-I gene from mitochondrial DNA with primer pairs FishF2-5' TCGACTAATCATAAAGATATCGGCAC 3' and FishR2-5'

ACTTCAGGGTGACCGAAGAATCAGAA 3' (Ward et al., 2005). PCR was performed in 20 µl final volumes using Phusion High-fidelity DNA Polymerase from New England Biolabs. Each reaction contained 4 µl 5x Phusion HF buffer, 0.4 µl 10 mM dNTPs, 1 µl 10 µM FishF2 primer, 1 µl 10 µM FishR2 primer, 12.35 µl nuclease-free water, 0.25 µl Phusion DNA polymerase (2000 units/ml) and 1 µl DNA template (c. 50 ng). The amplification conditions were: initial cycle of 98°C for 30 s followed by 33 cycles of 98°C for 10 sec, 59°C for 30 sec, 72°C for 30 sec, 72°C for 30 sec and final extension at 72°C for 10 mins. The amplification products were purified by spin column following the manufacturer's instructions (QIAquick PCR Purification kit). The purified samples were commercially sequenced (Sanger sequencing, GATC Biotech Ltd.). CO-I sequences from ten tilapia species were aligned using Clustal Omega, then a gene tree constructed using *RAxML* and visualized using *FigTree*.

5.3.9 Data Access

All species names used are in accordance with The Catalogue of Life (Roskov et al., 2014). The raw sequence data from this study have been submitted to the EBI Sequence Read Archive (SRA) study ERP006658.

5.4. RESULTS

5.4.1 Double Digest RAD sequencing

In total, 109,287,766 raw reads were produced (51,181,340 paired-end reads) from three batches of ddRAD libraries construction. After removing low quality sequences, ambiguous barcodes and orphaned paired-end reads, 85.75 % of the raw reads were retained (93,715,389 reads). In total the Stacks analysis identified 72,492 unique RAD-tags (i.e. the total number of loci across all species, with overlapping subsets of loci among species) in the de novo-based analysis (DBA) and 33,216 unique RAD-tags in the reference-based analysis (RBA) (Figure 5.2). However, less common RAD markers (loci retrieved in a minimum 75% of the samples) were found less in the DBA (6,064) than in the RBA (6,646). Polymorphic markers (poly allelic loci) were also found less in the DBA (5,954) than in the RBA (6,536). The number of reads and RAD-tags for each sample are reported in Appendix V.1.

5.4.2 SNP-based phylogenetic tree reconstruction

The phylogeny of the tilapiine species was inferred from all shared SNP markers in DBA and RBA. All 1,358 shared SNP markers in 825 RAD loci were identified across all tilapia species based on *de novo* analysis (DBA), meanwhile 938 shared SNP markers in 571 RAD loci were obtained that were common to all species in the RBA (Figure 5.3).

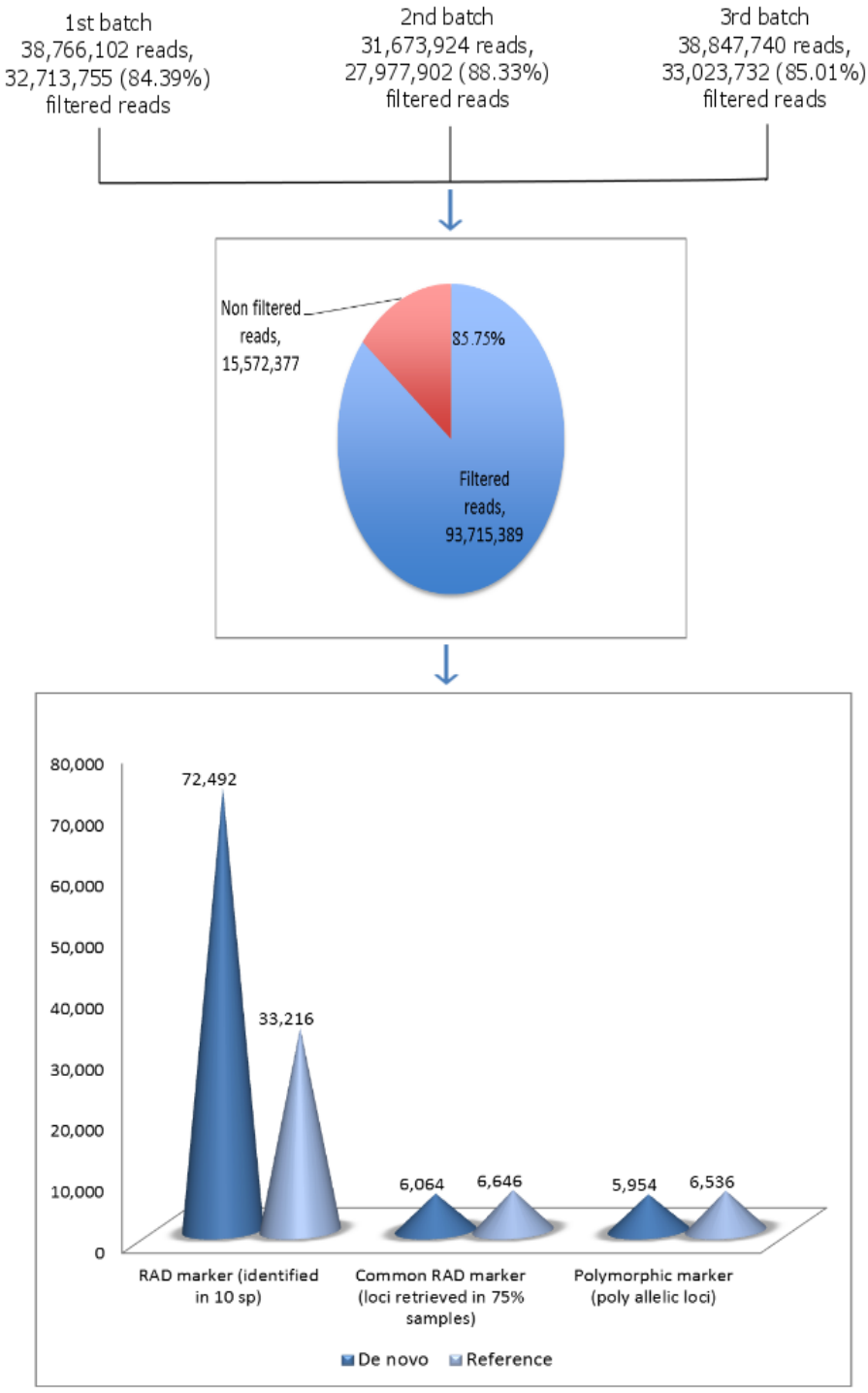


Figure 5.2 The number of retained reads and polymorphic loci between DBA and RBA

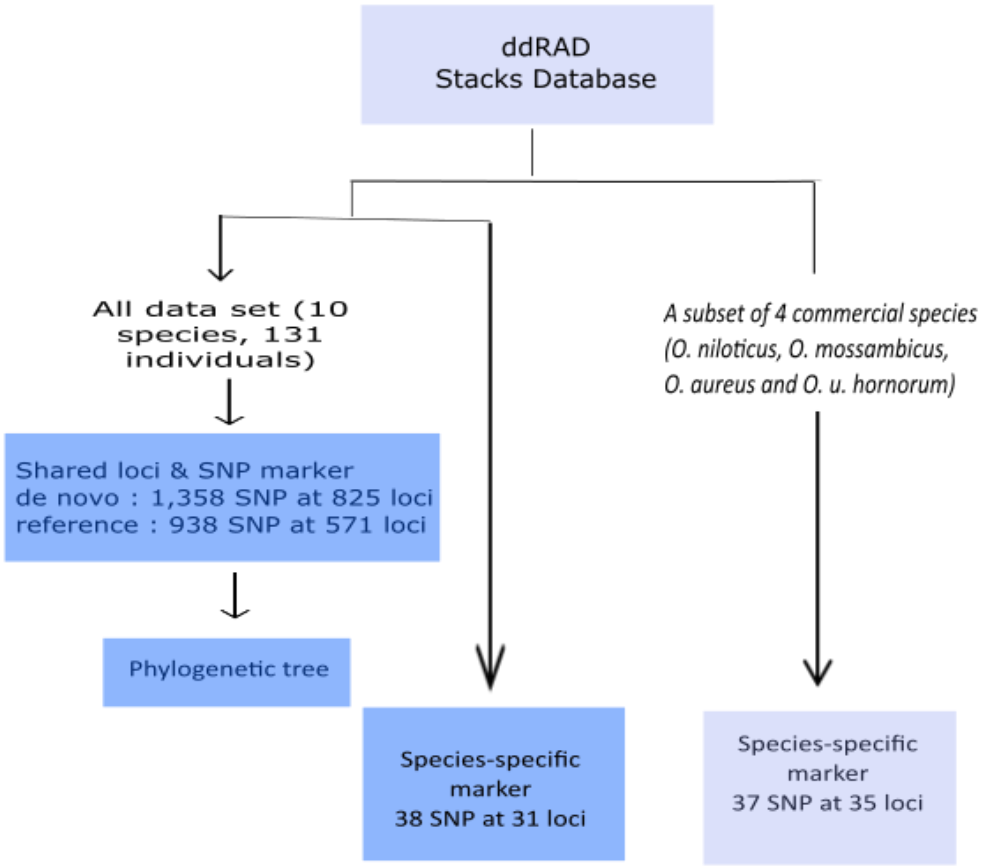


Figure 5.3 Flow diagram for retrieving shared loci and species-specific markers

Phylogenetic analysis of the ten tilapia species using these markers both DBA (Figure 5.4) and RBA (Figure 5.5) showed *T. zillii* furthest from all of the other tilapia species, with *Sarotherodon* (*S. melanotheron* and *S. galilaeus*) closer to *Oreochromis* (*O. niloticus*, *O. aureus*, *O. mossambicus*, *O. karongae*, *O. u. hornorum*, *O. macrochir* and *O. andersonii*). There was no difference in the phylogenetic tree pattern derived from DBA and RBA. The probability values across the branches (approximately between 95 and 100 in the CAT model in the DBA and 90 and 100 in the CAT model in the RBA), gave the highest level of confidence for species discrimination.

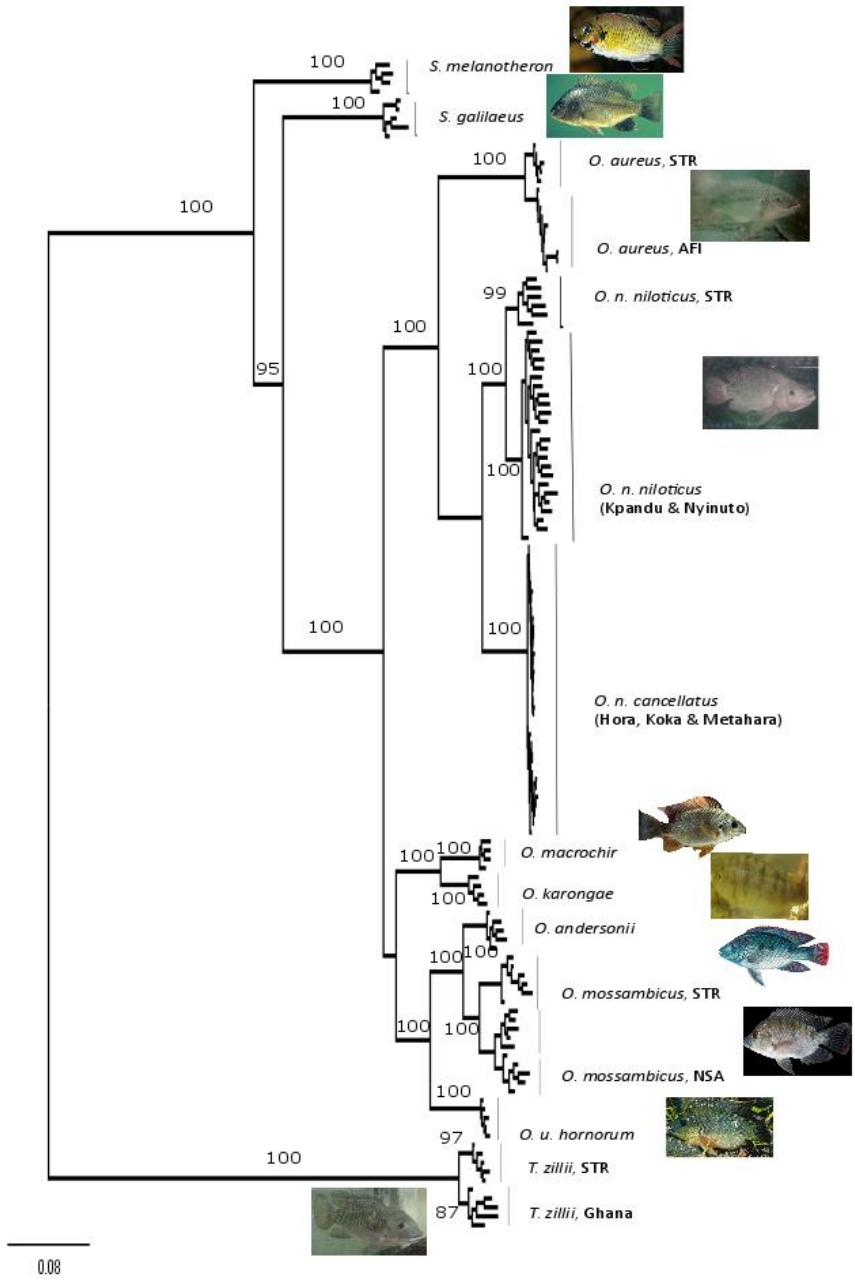


Figure 5.4 Phylogeny tree of tilapia species inferred from 1,358 shared SNP markers developed from de novo-based analysis and rooted to *T. zillii*.

All the sequences were aligned and the tree was constructed using RAxML (Randomized Axelerated Maximum Likelihood) software. Best-scoring ML tree with support values written to: RAxML_bipartitions file then was choosen and run to view the tree using FigTree. Support values are from 1000 bootstrap replicates with values along branches reported as Bayesian posterior probabilities/RAxML standart bbotstrap/RAxML GARLI bootstrap.

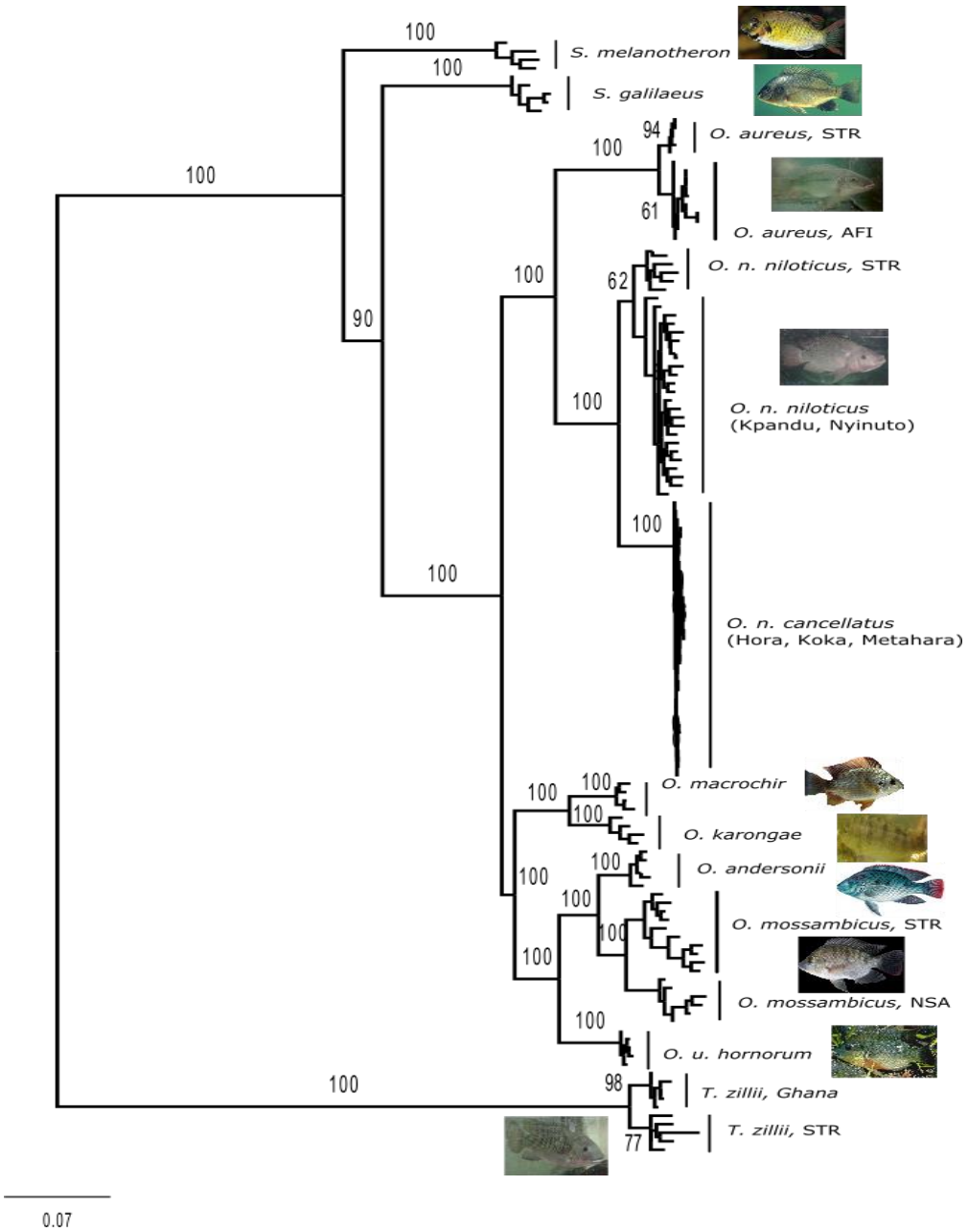


Figure 5.5 Phylogeny tree of tilapia species inferred from 938 shared SNP markers developed from reference-based analysis, and rooted to *T. zillii*.

All the sequences were aligned in Phylip format by a Clustal Omega program and constructed using RAxML (Randomized Axelerated Maximum Likelihood) software. Best-scoring ML tree with support values written to: RAxML_bipartitions file then was choosen and run to view the tree using FigTree. Support values are from 1000 bootstrap replicates with values along branches reported as Bayesian posterior probabilities/RAxML standart bbotstrap/RAxML GARLI bootstrap.

An enlargement of the *O. niloticus* clade from Figure 5.4 and 5.5 indicated that there was no clear difference among fish from Lakes Hora, Koka and Metahara within the subspecies *O. niloticus cancellatus*. In the subspecies *O. n. niloticus*, the Stirling stock (Egyptian origin) can be distinguished from the Volta drainage samples, but there was no discrimination between the two samples (Nyinuto and Kpandu) from the latter (Figure 5.6).

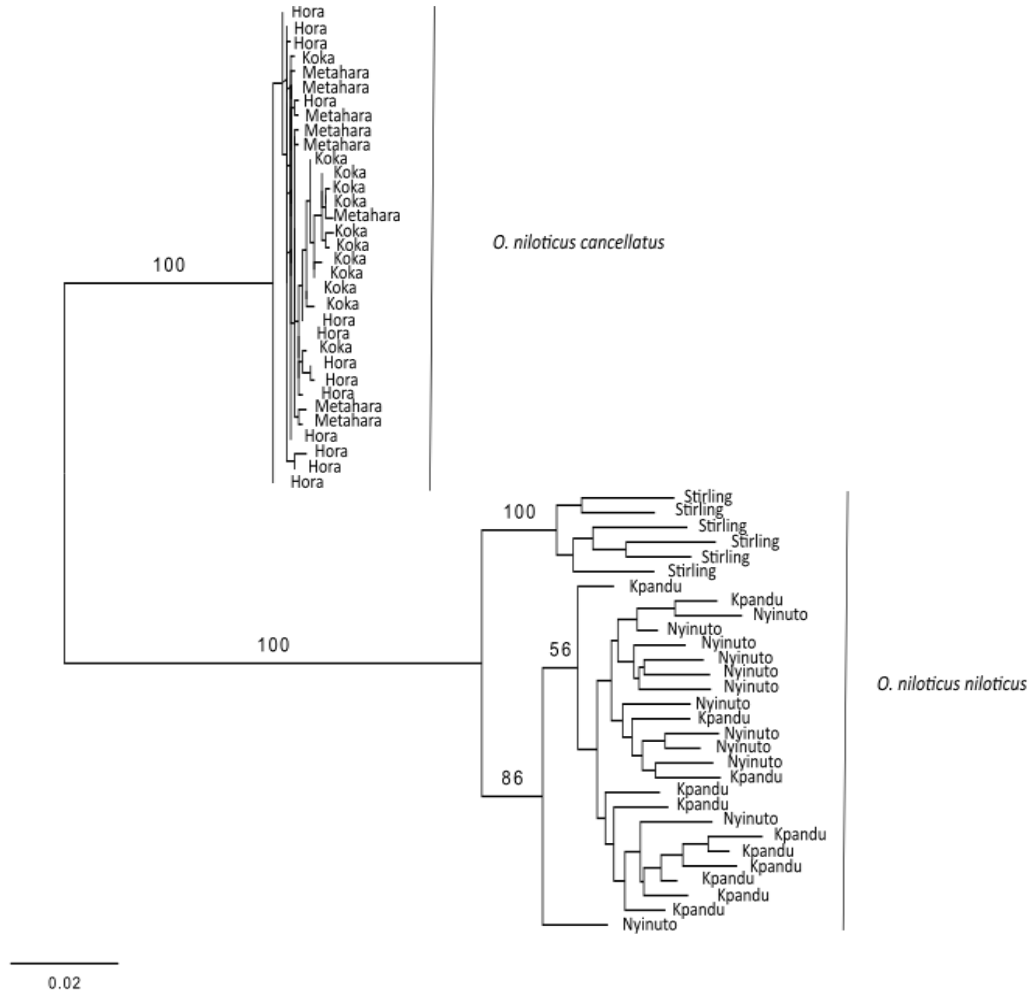


Figure 5.6 An enlarged version of the phylogenetic tree involving two sub species of *O. niloticus* developed from RBA.

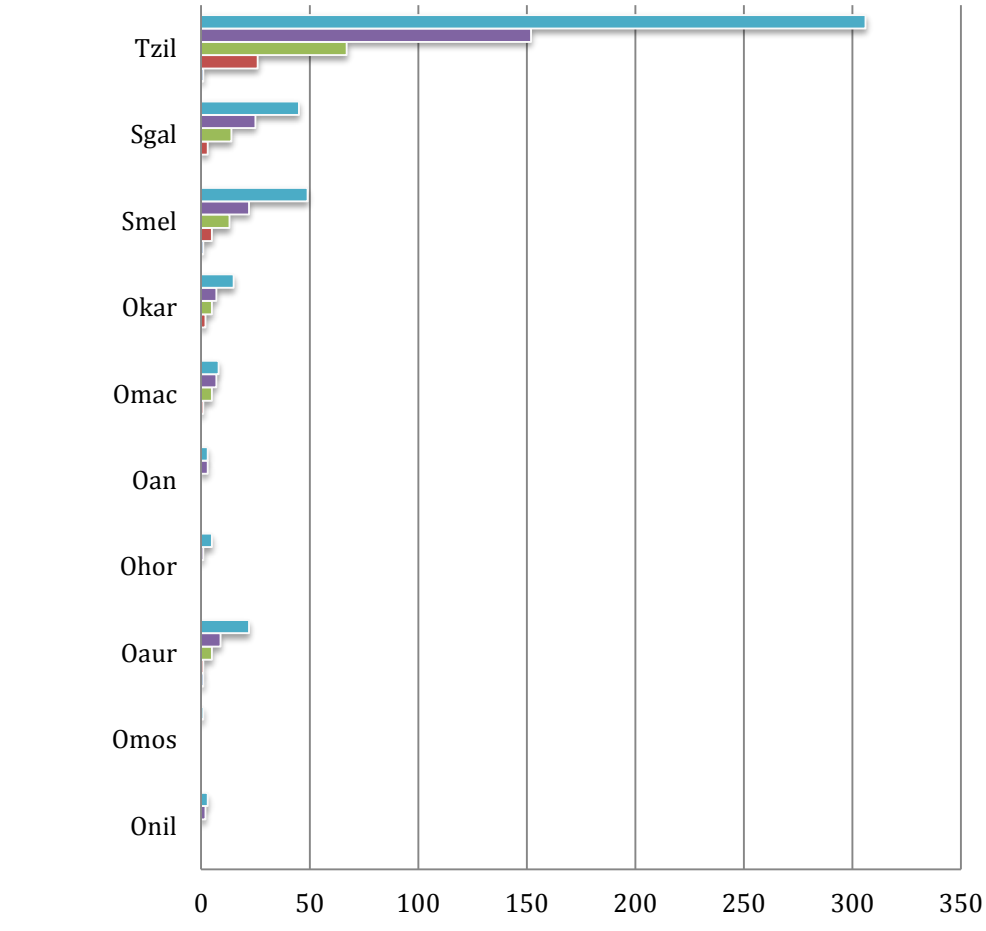
5.4.3 Species-Specific Diagnostic SNP Markers and Physical Map

In total, 38 species-specific SNP markers at 31 loci (i.e. with an allele that was unique to one species) were found across the ten tilapia species (with a minimum 75% of loci present in each species), consisting of one each for *O. aureus* and *O. macrochir*, two for *O. karongae*, three for *S. galilaeus*, five for *S. melanotheron*, 26 for *T. zillii*, but none was found for *O. andersonii*, *O. u. hornorum*, *O. mossambicus* and *O. niloticus*. The *Oreochromis* species (reflecting seven species) had 0-2 species-specific SNP, with 3-5 for *Sarotherodon* (consisting two species) and 26 for *T. zillii* (Table 5.2).

Table 5.2 SNP markers from RBA that are specific for tilapia species

No	Species/sub	Samples		SNP (max 2/locus)	
		Pop/sub-sp	Total	species	sub sp
1	*T. zillii*		10	26	
	a. Stirling	5			
	b. Ghana	5			
2	*S. melanotheron*		4	5	
3	*S. galilaeus*		5	3	
4	*O. niloticus*		63	0	3
	a. *niloticus*	30			
	b. *cancellatus*	33			
	O. n. niloticus		30	N/A	
	a. Stirling	6			
	b. Ghana	24			
5	*O. mossambicus*		16	0	
	a. Stirling	6			
	b. Nathal, SA	10			
6	*O. aureus*			1	
	a. Stirling	5			
	b. Israel	10			
7	*O. karongae*		4	2	
8	*O. u. hornorum*		5	0	
9	*O. macrochir*		4	1	
10	*O. andersonii*		5	0	

When the number of SNPs per locus was allowed to vary in the analysis up to a maximum of five, *T. zillii* also indicated the highest (306), while *O. mossambicus* showed the the lowest with only 1 SNP among tilapia sampled (Figure 5.7). At this level, species-specific SNPs were seen for all species.



	Onil	Omos	Oaur	Ohor	Oan	Omac	Okar	Smel	Sgal	Tzil
1-5 SNP/locus	3	1	22	5	3	8	15	49	45	306
1-4 SNP/locus	2	0	9	1	3	7	7	22	25	152
1-3 SNP/locus	0	0	5	0	0	5	5	13	14	67
1-2 SNP/locus	0	0	1	0	0	1	2	5	3	26
1 SNP/locus	0	0	1	0	0	0	0	1	0	1

Figure 5.7 SNP markers retrieved with up to five SNPs allowed per locus

No species-specific markers were found when a maximum of 3 SNPs/locus was allowed for the three commercial species (*O. niloticus*, *O. mossambicus* and *O. u. hornorum*) in the ten species data set. However, sufficient numbers of species-specific markers (37 SNP at 35 loci) were identified from a subset of the four economically important species, which are often involved in hybridization in aquaculture: both sub-species in *O. niloticus* (7 SNPs), *O. aureus* (13), *O. mossambicus* (10) and *O. u. hornorum* (7). These SNP markers distinguished each species from the other three, and a larger number of SNP markers distinguished between species pairs within this group (Table 5.3).

Table 5.3 The matrix of species-specific SNP markers/loci retrieved from ddRADseq for four commercial tilapia species.

Species/number	*O. niloticus*	*O. aureus*	*O. mossambicus*	*O. u.hornorum*
O. niloticus (63)				
O. aureus (15)	22/20			
O. mossambicus (16)	66/60	86/79		
O. u. hornorum (5)	85/81	109/102	18/18	
All	7/6	13/13	10/10	7/7

Analysis using perl script of find_pattern.pl also succeed in resolving species-specific SNP markers at the sub-species level for *O. niloticus*. Three SNP at LG1, 2 and 23 were identified between sub-species *O. n. niloticus* and *O. n. cancellatus*, representing three natural geographical regions (Table 5.2, Figure 5.8). One SNP marker each was retrieved between Stirling and Ghana population for *T. zillii*, and Stirling and Natal-South Africa population for *O. mossambicus*.

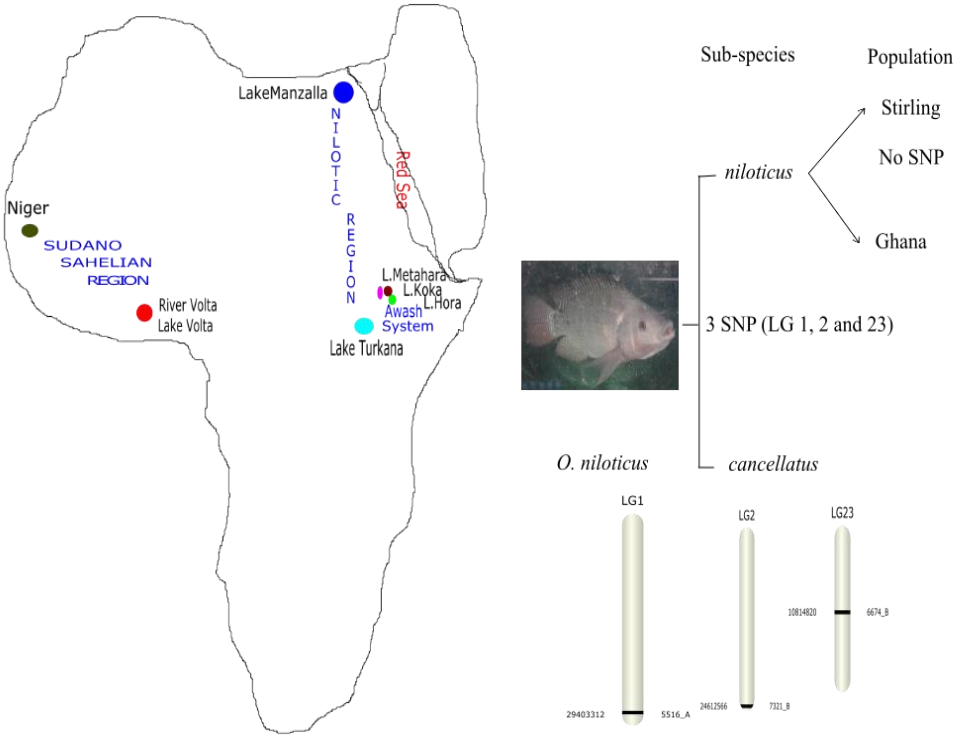


Figure 5.8 Three natural geographical regions of *O. niloticus* and their subset of SNP at sub-species level.

The left figure shows the three natural geographic regions (Nilotic region, Awash system and Sudano-Sahelian region), while the right figure indicates the number of SNPs at sub-species level.

Physical mapping of the 38 species-specific SNP markers (mostly for *T. zillii*) from ddRADseq onto the *O. niloticus* genome assembly showed that the diagnostic markers were distributed randomly across the genome (Table 5.4). As an example, Figure 5.9 illustrates the physical map of LG 19 and the distribution of SNP markers (one each for *T. zillii, S. melanotheron, O. karongae* and *O. macrochir*). The list of species-specific markers can be seen in Appendix V.2, while a complete mapping of SNP in the reference genome located in Appendix V.3.

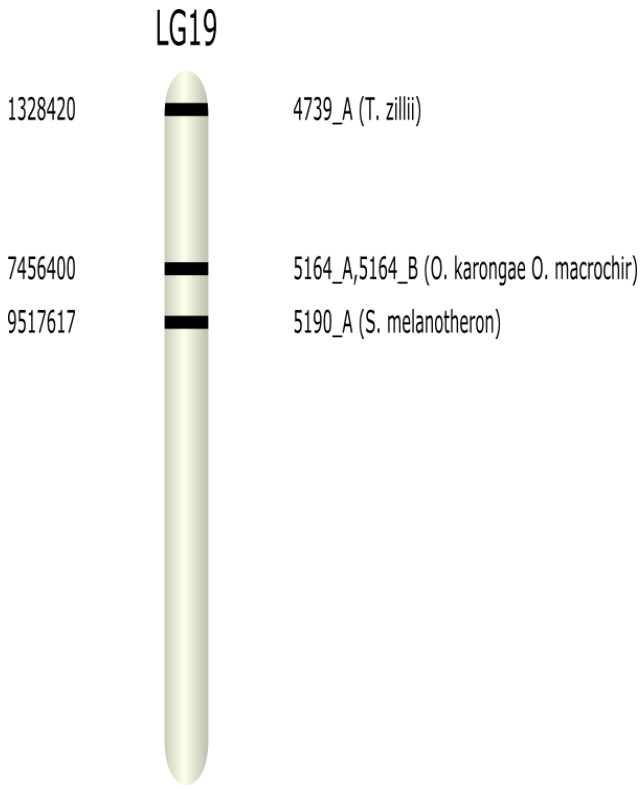


Figure 5.9 The diagnostic SNP markers in LG 19 of tilapia species. Numbers on the left side show the SNP position (bp), while those on the right side denote the catalog id and the name of the species for which an allele of the SNP is unique.

Table 5.4 The genomic distribution of species-specific SNP markers for tilapia, derived from ddRADseq data

LG	Physical size of LG (bp)	*Oau*	*Oka*	*Omo*	*Oni*	*Ohor*	*Oan*	*Omac*	*Smel*	*Sga*	*Tzi*	Total	Number of SNP/Mb
1	31,194,787	0	0	0	0	0	0	0	0	0	1	1	0.03
2	25,048,291	0	0	0	0	0	0	0	0	0	0	0	0.00
3	19,325,363	0	0	0	0	0	0	0	0	0	0	0	0.00
4	28,679,955	1	0	0	0	0	0	0	0	0	0	1	0.03
5	37,389,089	0	0	0	0	0	0	0	0	1	3	4	0.11
6	36,725,243	0	0	0	0	0	0	0	0	0	2	2	0.05
7	51,042,256	0	0	0	0	0	0	0	0	0	0	0	0.00
8_24	29,447,820	0	1	0	0	0	0	0	0	1	1	3	0.10
9	20,956,653	0	0	0	0	0	0	0	0	0	0	0	0.00
10	25,048,291	0	0	0	0	0	0	0	0	0	1	1	0.04
11	33,447,472	0	0	0	0	0	0	0	0	1	1	2	0.06
12	34,679,706	0	0	0	0	0	0	0	0	0	2	2	0.06
13	32,787,261	0	0	0	0	0	0	0	2	0	4	6	0.18
14	34,191,023	0	0	0	0	0	0	0	0	0	2	2	0.06
15	26,684,556	0	0	0	0	0	0	0	0	0	1	1	0.04
16-21	34,890,008	0	0	0	0	0	0	0	0	0	2	2	0.06
17	31,749,960	0	0	0	0	0	0	0	0	0	1	1	0.03
18	26,198,306	0	0	0	0	0	0	0	1	0	0	1	0.04
19	27,159,252	0	1	0	0	0	0	1	1	0	1	4	0.15
20	31,470,686	0	0	0	0	0	0	0	0	0	0	0	0.00
22	26,410,405	0	0	0	0	0	0	0	0	0	1	1	0.04
23	20,779,993	0	0	0	0	0	0	0	1	0	3	4	0.19
Total	665,306,376	1	2	0	0	0	0	1	5	3	26	38	1.02

5.4.4 COI

The COI partial sequences of the tilapia species that were retrieved varied between 395-631 bp, agreed with those in the Barcode of Life Data System (BOLD) and the NCBI GenBank Database. The COI gene tree indicated *Tilapia* genera were separated from *Sarotherodon* and *Oreochromis*, however there were some overlaps between the other two (Figure 5.10). The largest group consisted of most of the *Oreochromis* species i.e. *O. niloticus*, *O. mossambicus*, *O. karongae*, *O. u. hornorum*, *O. andersonii* and *O. macrochir*. However, *O. aureus* was in a group with *S. galilaeus*, while *S. melanotheron* was in a separate group from *S. galilaeus*. West African *O. niloticus* (Onn_Kp and Onn_Ny) exhibited COI haplotypes typical of *O. aureus*, as previously reported in Rognon & Guyomard (2003), although nuclear markers clearly indicated the differences between these two species. The last group consisted of the two populations of *T. zillii*, being the most distant species from the *Oreochromis* genus.

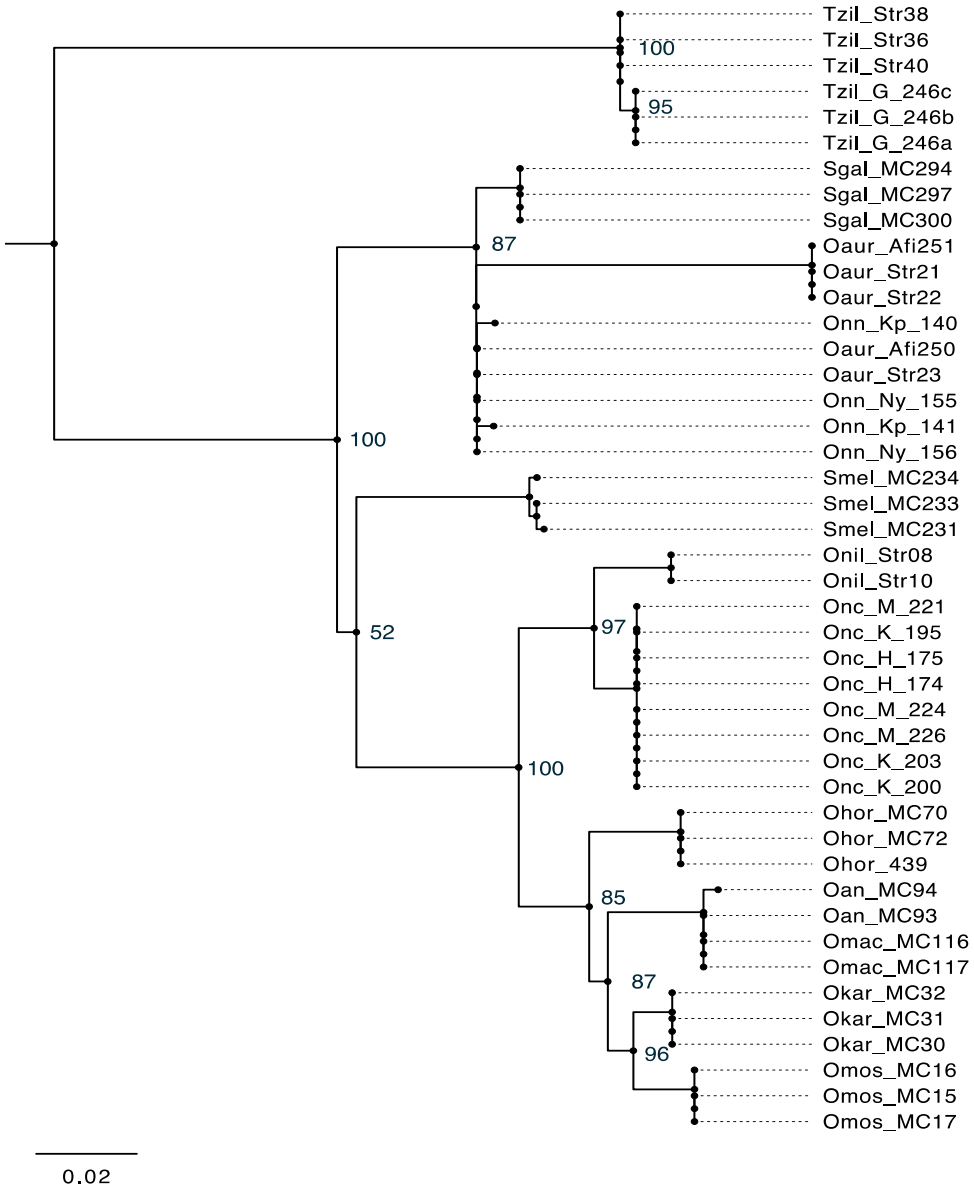


Figure 5.10 A gene tree of tilapia species inferred from COI sequences, and rooted to *T. zillii*.

All the sequences were aligned in Phylip format by a Clustal Omega program and constructed using RAxML (Randomized Axelerated Maximum Likelihood) software. Best-scoring ML tree with support values written to RAxML_bipartitions file then was choosen and run to view the tree using FigTree. Tzil, *T. zillii* (Str, Stirling; G, Ghana); S.gal, *S. galilaeus*; Smel, *S. melanotheron*; Oaur, *O. aureus* (Str, Stirling; Afi, Ain Feskha Israel); Onn, *O. niloticus niloticus* (Kp, Kpandu; Ny, Nyinuto); Onc, *O. niloticus cancellatus* (H, Hora; K, Koka; M, Metahara); Ohor, *O. u. hornorum*; Oan, *O. andersonii*; Omac, *O. macrochir*; Okar, *O. karongae*; Omos, *O. mossambicus*; MCxxx, coding and numbering system.

5.5 DISCUSSION

5.5.1 Unique ddRAD marker in the DBA and RBA

Figure 5.2 shows that the number of unique RAD marker decreased to 45.82% from 72,492 in the de novo based analysis (DBA) to 33,216 in the reference based analysis (RBA). In the DBA the filtered reads were used directly in the Stacks, a software pipeline for building loci from short-reads sequencing, while in the RBA, those consisted only loci found in the Broad Institute of MIT and Harvard *O. niloticus* genome assembly Orenil1.1, NCBI assembly GCF_000188245.2 used in the analyses. The difference in the parameters applied between the DBA and the RBA was in the number of mismatches between catalog loci. In the DBA, a maximum of 4 mismatches were allowed in a locus in an individual while only 1 mismatch was allowed in the RBA. The increasing number of mismatches between catalog loci in the DBA avoids many false loci, which may be created from Stacks, therefore it influences the number of RAD marker identified in the catalog stacks. In the RBA the retained reads have been aligned to the reference genome, so the number of RAD marker identified in the stacks give more stringent results than in the DBA. Some loci were discarded when they were aligned to the reference sequence. Furthermore, the reference genome assembly in *O. niloticus* only represent 70%, so many more loci were also excluded from the analysis. However the number of common RAD marker with minimum 75% of matching samples was still higher in the RBA (6,646) than the DBA (6,064). There was a high drop of unique

RAD markers in the DBA due to many spurious RAD tags in the DBA, where many more markers only appeared in one individual.

Inferring a genetic structure of single nucleotide polymorphism (SNP) variation in wild and farmed populations of tilapia species would require the sampling of tens of species with hundreds of individuals each. There is a trade-off between sample size and number of markers (SNPs in this case), which allows to reduce the sample size per species/sub-species/population to some extent. In addition, adding more than one sample where possible will strengthen analysis. One factor to be considered in the analysis of sequences from distinctly separated genomes is that Stacks may remove a majority of the loci from its analysis, or even divide single loci into two (Eaton, 2014). In the absence of any mismatches between loci (default setting in denovomap.pl), the SNPs could not distinguish between species, however the discrimination is meticulous when mismatches are allowed. Therefore, biological diversity should be considered while generating SNPs in stacks (Chattopadhyay et al., 2014).

5.5.2 Phylogenetic Tree

The phylogenetic tree developed from shared SNP markers (both DBA and RBA) found with ddRADseq showed significant distance measures between the three genera of tilapia: *Sarotherodon*, *Tilapia* and *Oreochromis*. *Sarotherodon* consisted of two species, *S. melanotheron* and *S. galilaeus*. *Oreochromis* consisted of seven species (*O. niloticus*, *O. mossambicus*, *O. karongae*, *O. aureus*, *O. u. hornorum*, *O. andersonii* and *O. macrochir*) while

Tilapia only consisted one species, *T. zillii*. This result was concordant with previous evidence generated with standard RAD-seq (Chapter 4).

The COI gene tree generally agrees with previous publications using allozymes (Sodsuk & McAndrew, 1991; Pouyaud & Agnèse, 1995) and the mitochondrial control region (Nagl et al., 2001), in that *Sarotherodon* species were not clearly separated from *Oreochromis* (unlike our ddRAdseq study using 938 common SNP in the RBA, in which all three genera were separated). We also could not separate *O. andersonii* and *O. macrochir* or West African *O. niloticus* from *O. aureus* using COI sequence. The shortcomings of the COI gene tree are due to it being based on only one marker (with maternal inheritance), so it could not represent the depth expected from multiple nuclear DNA markers.

Meanwhile, the COI data could not separate between West African *O. niloticus* and *O. aureus*. One study reported that nuclear markers (allozymes) showed distinct separation between *O. aureus* and *O. niloticus* in West African populations, although the same sequences in the mtDNA were detected in both species (Rognon & Guyomard, 2003). The current study obviously indicates a very clear differentiation between these two species at the nuclear DNA level. Furthermore, despite common natural distribution between *O. aureus* and *O. niloticus* (Trewavas, 1983), they do not interbreed in nature (Payne & Collinson, 1983).

Phylogenetically, *O. andersonii* can be distinguished from *O. macrochir* using nuclear SNP data (but not using COI sequence). One study reported evidence of hybridization and introgression of both native species (*O. andersonii*

and *O. macrochir*) with *O. niloticus* in the Kafue River fishery resulting in a complex mixed population consisting genetic material from all three species, however, the genetic diversity of the mixed population appears to be lower than that of the parental type (Deines et al., 2014). Another study reported a mating barrier between these two species due to behavioural isolation mechanism (Falter & Dufayt, 1991). In addition, a previous study indicated a low frequency of hybridization between introduced species *O. niloticus* and *O. andersonii* with native *O. mossambicus* in South Africa (Angienda et al., 2010). The current study strongly shows that all four species can be clearly discerned based on the shared SNP markers (Figure 5.4 and Figure 5.5), which is promising for the use of such SNP markers in future studies involving potential hybridization and introgression among these species.

The shared SNP markers also showed a clear distinction between sub-species of *O. niloticus.* The two sub-species, *O. niloticus cancellatus* and *O. niloticus niloticus*, formed two separate clades. The *O. niloticus cancellatus* (found in Lake Hora, Koka and Metahara, all in the Awash System in the Ethiopian Rift Valley) formed a single branch with little discrimination between the different lake populations. *O. niloticus niloticus* was sampled from Lake Manzala of the Nilotic region of Egypt, and the Lake and River Volta (strain Kpandu and Nyinuto) of the Sudano Sahelian Region, Ghana.

5.5.3 Species-Specific SNP Marker

Double digest RADseq analysis retrieved 38 species-specific diagnostic SNP markers across ten tilapia species, which are potentially capable of

distinguishing these species. The analysis identified species-specific markers from 6 species, *T. zillii, S. melanotheron, S. galilaeus, O. aureus, O. macrochir* and *O. karongae*, but none were found for *O. andersonii, O. u. horonorum, O. mossambicus* and *O. niloticus* (with a maximum of 2 SNPs/locus). As a consequence we could not distinguish the four species in the set of 10 species studied. However, when the number of SNP per locus was allowed to increase in the analysis up to a maximum of five, species-specific markers were identified in these species. For instance, we identified three for *O. niloticus* and one for *O. mossambicus*. Therefore, two different methods can be applied for application of these markers in genetic hybridization and introgression occurrence. The species-specific markers consisted 1-2 SNPs can be used for simple and quick assay using KASP assay genotyping, while those contained up to 5 SNPs can be applied using PCR and sequencing.

In comparison to standard RAD, all of the species indicated a high drop in numbers of species-specific SNP markers in ddRAD-seq. In total, the numbers of species-specific markers decreased from 677 in the standard RAD to 38 in the ddRAD, in part because more species/sub-species and populations were added to the analysis. Even none of SNP markers (1-2 SNP) were specific for *O. niloticus*, *O. mossambicus*, *O. andersonii* and *O. u. hornorum* identified in the ddRAD. In *T. zillii*, the numbers decreased from 400 SNP to 26. The reduction is also in line with expectations because of the reduction in specific restriction cut sites, which will also reduce the number of polymorphic loci and SNP for each species. Furthermore, the different criterion was applied between standard RADseq and ddRADseq. In the standard RAD, the analysis included at

least 75% of matching samples (based on the total number of individuals) without considering the species number. Therefore, certain species, especially those consisting of small number of individuals will be not included in the analysis. In contrast, in the ddRADseq, the analysis required at least 75% loci present in each species, which avoid one of species discarded from the analysis. As a consequence, some species-specific markers derived from RADseq were not always diagnostic in the set of 7 species studied, while in ddRADseq, all the species-specific markers were diagnostic in the set of 10 species studied.

Analysis of fixed allele patterns in tilapia species also retrieved species-specific SNP markers at the sub-species level. In *O. niloticus*, it could resolve the two sub-species, with three SNP markers were retrieved between the sub-species *niloticus* and *cancellatus*. The fact that diagnostic SNP markers can be detected between two sub-species suggests that the Nile tilapia is an evolving supra-species (Prunet & Bornancin, 1989) but is also evidence of its high diversity. One of the SNPs was located in LG1 (marker id_5516A), distinguishing specific allele A for *niloticus* from allele G for *cancellatus,* located in the region coding for testis-specific serine/threonine-protein kinase 1-like with 99% locus was identical (NC_022199). There was an insertion polymorphism based on nucleotide BLAST using reference genomic sequences in the NCBI gene bank.

In most species, populations are often subdivided into smaller units because of geographical, ecological or behaviour factors. When a population is subdivided, the amounts of genetic connectedness among the parts of the population can differ, primarily dependent on the amount of genetically effective

gene flow among them. Gene flow has the effect of homogenizing genetic variation over the groups when its value is high, however when it is low, genetic drift, selection and even mutation in the separate groups may lead to genetic differentiation (Hedrick, 2005). A sample size of at least 30 individuals would likely allow for a high probability of detecting even rare alleles (>95% of detecting an allele with a minor allele frequency of 0.05, or 99% with MAF = 0.1) (Allendorf, et al., 2013). Rare SNPs should be divided into two categories dependent on whether the minor alleles are derived or ancestral. The functional and structural consequences are more significant for the rare exonic variants for which the minor alleles are derived (Gu et al., 2015). In the case of low size sample, increasing the number of species-specific markers in a comprehensive SNP panel may reduce the variance in admixture estimates caused by undetected ancestral polymorphism (Hand et al., 2015). Despite a limitation in the number of species, populations and individuals used in the ddRADseq, the number of SNP markers obtained looks promising.

A sufficient number of species-specific SNP markers retrieved from a reduced set of the four economically important tilapia species, ranging from 7 in *O. niloticus*, 10 in *O. mossambicus*, 7 in *O. u. hornorum* to 13 in *O. aureus*, while between species pairs, *O. mossambicus* - *O. u. hornorum* showed the fewest markers (18) and *O. aureus* - *O. u. hornorum* indicated the highest number of SNP markers (109). The number of SNP markers in the species pairs of *O. aureus* – *O. u. hornorum* inflated might be due to the low sample size of *hornorum* (n=5) in comparison to others tilapia species, where they had higher sample sizes and multiple populations. This set of SNP markers will enable

managers to detect hybridization and introgression among these four tilapia species, for instance between *O. niloticus* – *O. mossambicus* there were 66 SNP marker at 60 loci that could be used for species discrimination.

5.6 Conclusions

In summary, the three subsets of double digest RAD-seq using *Sbf*I and *Sph*I as restriction enzyme produced 85.75% retained reads which generated 38 species-specific SNP markers across a set of 10 tilapia species. A sufficient number of SNP markers were identified across four commercial tilapia species (7-13 each species) and its species pairs (18-109). It should be possible to use these species-specific markers to investigate hybridization and introgression, not only at the species level but also at the sub species in tilapia.

5.7 Acknowledgments

We are grateful to Dr. Andrew M. Deines for the *O. andersonii* and *O. macrochir* samples. We thank to Keith Ranson of the Tropical Aquarium Facility at the Institute of Aquaculture, University of Stirling, for help in rearing fish. The authors acknowledge the support of the MASTS pooling initiative (The Marine Alliance for Science and Technology for Scotland). MASTS is funded by the Scottish Funding Council (grant reference HR09011) and contributing institutions. We also thank to Director General of Higher Education, Ministry of Research, Technology and Higher Education (*Kemenristek Dikti*), Indonesia for funding PhD scholarship of MS at the University of Stirling.

6. GENERAL DISCUSSION

6.1 Polymorphism and RAD markers

A range of loci have been analyzed and used for species id and also reconstruction of phylogenetic relationships. Based on these studies, the *ADA* locus showed the greatest level of interspecies polymorphism (McAndrew and Majumdar, 1983; Sodsuk et al., 1995), therefore developing SNP markers based on the *ADA* gene could be a promising new method of distinguishing tilapia species. A part of the coding sequence of the gene was sequenced from five tilapia species, representing four from the *Oreochromis* genus (*O. niloticus*, *O. aureus*, *O. mossambicus*, *O. karongae*) and *Tilapia zillii*. The sequence length spanned from 1076-1092 bp showing nucleotide variations within and between species and matched with the sequence of the predicted adenosine deaminase_like gene of *O. niloticus* (accession NC_022218.1), which consisted of 10 introns and 11 exons encoding 364 amino acids in the NCBI GeneBank database. However sequences in the first and eleventh exons were not retrieved successfully.

The highest level of *ADA* sequences polymorphism was found in *T. zillii*, both non-synonymous (3.11%) and synonymous (2.20%), while *Oreochromis* spp. indicated less polymorphism. The average percentage of synonymous polymorphisms (0.41%) in *O. niloticus* was higher than non-synonymous polymorphisms (0.09). In contrast, all other tilapia species indicated that the percentage of synonymous polymorphisms (0.18-2.20%) was lower than the percentage of non-synonymous polymorphisms (0.46-3.11%). The most

significant findings, *ADA* gene in *O. niloticus* indicated dN/dS is greater than 1, meaning that the region in this species is under positive selection without considering the functions. In contrast the dN/dS others samples showed the value less than 1, indicated these genes are unlikely to be under selection. A gene or a phenotypic trait is said to be polymorphic if there is more than one form of the gene, or more than one phenotype for the character in a population. In some cases, nearly an entire population expresses the phenotype of one form of the gene or character and carries an unusual variant of the gene, while a smaller number of individuals express the rare phenotype (Griffiths, Gelbart, Lewontin, & Miller, 2002). Different parts of the gene are found to evolve at different rates, and those parts of the gene that have the least effect on fitness appear to evolve the fastest. Multi-gene families evolve through repeated duplication of genes, followed by genetic divergence of their sequences. In addition, mitochondrial DNA in animals evolves at a faster rate the DNA of nuclear genes (Russel, 1994).

Sequencing with standard RAD and ddRADseq involved identifying DNA polymorphisms in both coding and non-coding sequences. *Tilapia zillii* as the furthest species indicated the highest polymorphism with 306 species-specific RAD markers, while *O. mossambicus* showed the least polymorphic with only one (up to a maximum of 5 SNPs per locus) in comparison to others tilapia sampled. Generally, the ddRADseq showed a decreased number of SNP markers than standard RAD. Digesting a genome with two restriction enzymes instead of one can generate many distributions of fragment sizes, depending on the frequency of cutting. In this study, enzyme *Sph*I (with a 6 bp recognition site comprising a GCATG|C motif) generates the potential to sample a larger fraction of the genome

than enzyme *Sbf*I (with an 8bp recognition site comprising a CCTGCA|GG motif) because it cuts more frequently than *Sbf*I. Using two enzymes can only analysing fragments with the correct P1 and P2 barcodes, therefore they reduce the fraction of genome analysed. Generally, pairs of enzymes with common cut sites can be used to sample many loci in each individual, but fewer individuals. In contrast, a pair of enzymes with rare sites will sample fewer loci but in a greater number of individuals at the same cost (Arnold et al., 2013). Therefore, combining enzymes between a common (*Sph*I) and a rare cutter (*Sbf*I) will generate an intermediate fraction. RADseq with double digest restriction enzyme (RE) excludes regions flanked by either very close or very distant RE recognition sites, recovering a library consisting of only fragments close to the target size (Peterson et al., 2012).

The number of ddRAD marker in the RBA also decreased in comparison to the DBA. Aligning loci to reference genome sequences will miss particular loci due to the fraction of the genome that has not been captured in the genome assembly (up to about 30% in *O. niloticus*), therefore the analysis of investigating sequences from many species will depend on the criterion implemented in the analysis. For example, in the absence of any mismatches between loci (default setting in denovomap.pl), the SNPs could not distinguish between species, whereas the SNPs can discriminate between species when allelic mismatches are allowed. In addition, Stacks software may remove the majority of the loci from its analysis, or even divide a single locus into two (Eaton 2014). Therefore, biological diversity should be considered while generating SNPs in Stacks software (Chattopadhyay et al., 2014).

6.2 Phylogenetic relationships

Two methods can be used as independent approaches to the reconstruction of phylogeny: cladistics (the assessment of morphological characters for the most parsimonious [shortest] tree linking the species); and molecular phylogeny reconstruction (using protein or RNA/DNA) (Osborne and Benton, 1996). Phenotypes can sometimes be misleading about evolutionary relationships because phenotypic similarities do not necessarily reflect genetic similarities (Russel, 1994), therefore studying sequence data allows for a clearer phylogenetic relationship between individuals to be established. A gene tree based on the *ADA* gene indicated distinct clustering of tilapia species where *O. niloticus* was closer to *O. aureus* and *O. karongae* was closer to *O. mossambicus*, while *T. zillii* had the furthest genetic distance from the other four species. This pattern was concordant with the interspecific polymorphisms derived from standard RAD, ddRADseq and the mitochondrial COI gene. All the trees supported the taxonomy proposed by Trewavas (1983). However, when the number of species and populations were expanded, the COI sequence could not resolve evolutionary relationships of some tilapia species e.g *Sarotherodon* species from *Oreochromis* and *O. andersonii* from *O. macrochir*.

The mitochondrial DNA and nuclear DNA can be used to infer phylogenetic trees from sequence data. The *ADA* and COI genes are both Type I markers that originated in the nucleus and mitochondria respectively, both of which encode a specific function at a specific locus. However, when only a single gene is used to look at evolutionary relationships, it only generates a gene tree

rather than giving information about the broader evolutionary history of a group of species (Nei and Kumar, 2000). In addition, the COI gene is mitochondrial, only inherited via the maternal lineage, so it will not represent a complete evolutionary relationship between species. The complete COI gene might have enabled to resolve fine-scale genetic divergence among species, but the tree would still have been based on variation in a single gene (Emerson et al., 2010). Other studies indicate that mtDNA evolution is non-neutral with sufficient regularity to question its utility as a marker for genomic history. Making inference from mtDNA data could be unreliable due to direct selection (selection on mtDNA itself) and indirect selection (selection arising from disequilibrium with other maternally transmitted genes) (Ballard & Whitlock, 2004).

The COI DNA barcode is able to resolve most, but not all, of the species involved in this study. As a single, maternally inherited marker it is of limited use in analysing cases of hybridization/introgression. However, it is still likely to be useful in combination with multiple nuclear DNA markers. The *ADA* approach was partially successful, where four out of ten SNP markers derived from *ADA* sequences for *T. zillii* (Tzil_3_M170), *O. aureus* (Oaur_3_R122, Oaur_7_R626) and *O. mossambicus* (Omoss_10_Y879) could potentially be applied in identifying and discriminating among tilapia species.

In contrast, inferring phylogenetic trees based on unique shared markers between tilapia species gave data more representative of the whole genome, including coding and non-coding sequences. Constructing phylogeny from RADseq data from many loci throughout the genome is generally reliable with a wide range of clustering and filtering parameters (Rubin et al., 2012). However,

RADseq data gives promising results for reconstructing phylogenetic relationships only in younger clades in which sufficient numbers of orthologous restriction sites are retained across species. Studies have indicated that high-throughput sequencing of RAD tags is capable of resolving fine-scale genetic divergence among intraspecific populations that have been separated for less than 20,000y (Emerson et al. 2010). Molecular and geological evidence suggests that the *Haplochromis* species flock of Lake Victoria arose in the very recent geological past, (from 14- to 750,000 years ago) (Osborne & Benton 1996). So, it is obvious that tilapiine genera diverge from other cichlids quite recently. The divergence time of *O. niloticus* and *T. zillii* from African cichlid fish are less than 50 million years ago (Brawand et al., 2014).

6.3 Species-specific Markers

Coding sequence in *ADA* is highly polymorphic, but one single locus used to develop new SNP markers remains a weakness. In addition, finding SNPs using conventional sequencing is labour intensive and has a limited capacity for in-depth sequencing. In contrast, RAD sequencing, particularly ddRADseq, allows multiplexing of samples with reduced sampling error and reduces the effect of outlier loci by providing a much denser genome-wide sample of genotype data, thus providing a more precise estimate of actual phylogeographic relationships (Emerson et al., 2010).

DNA barcoding using COI can be a good tool for investigating maternal inheritance, but in this case did not resolve evolutionary history between, for example, *S. melanotheron* - *S. galilaeus* or *O. andersonii* - *O. marochir* because

only a single locus was used. Furthermore, to develop diagnostic SNP markers, more individuals must be sequenced at high depth. In RADseq, *Sbf*I restriction enzymes in the genome produce a large number of loci due to the infrequent cuts, but there is a restriction on the number of samples that can be included in the analysis and so the scope for detecting SNPs in multiple individuals may be limited. Double digest RADseq, however, produces fewer loci because the use of two restriction enzymes allows for a more specific fragment cut, but can involve more individuals. In this study, the number of samples increased from 50 in standard RADseq to 132 individuals in ddRADseq, while ddRADseq produced 1.5 times less the number of RAD markers (33,216) than standard RADseq (51,750).

Species-specific marker can be retrieved both from standard RADseq and double digested RADseq. In comparison to standard RAD, all of the species indicated a drop in the number of species-specific SNP markers in ddRAD-seq. In total, the numbers of species-specific markers decreased from 677 in the standard RAD to 38 in the ddRAD, in part because more species/sub-species and populations were added to the analysis. The reduction in specific restriction cut sites in the ddRADseq will also reduce the number of polymorphic loci and SNP for each species. Furthermore, the number of species-specific marker identified will also depend on how stringent the result will be. The number of species-specific markers in the standard RAD were retrieved from 75% of matching samples (7 species), which effected the number of SNP for each species, but some individual in the species might be lost or all individuals from one species could be excluded from the analysis. In contrast, there were no species-specific markers in

the ddRADseq for several *Oreochromis* spp when all 10 tilapia species were included with a minimum 75% loci present in each species and maximum 2 SNPs/locus, but we do find species-specific markers when a maximum of 5 SNPs/locus was allowed in every species studied. A sufficient number of SNP markers can be retrieved for a comparison of the four most economically important species, from 7 for *O. niloticus* to 13 for *O. aureus*. This improves when we are just interested in species pairs, for instance there are 22 SNPs at 20 loci between *O. niloticus* - *O. aureus* and 66 SNPs at 60 loci between *O. niloticus* - *O. mossambicus*. In culture, there is often a high occurrence of hybridization, for instance between *O. niloticus* x *O. aureus*, *O. niloticus* x *O. mossambicus*, and red hybrids involving multispecies and generation. While in the wild, there are many samples of hybridization and introgression, for example high degree of mixing between *O. mossambicus* and *O. niloticus* in Southern Sri Lanka. Species-specific SNP markers were also retrieved in the sub-species level in *O. niloticus*, where three SNP markers found between *O. n. niloticus* and *O. n. cancellatus*. The fact that diagnostic SNP markers can be retrieved between two subspecies proves that Nile tilapia is an evolving supra-species (Prunet & Bornancin, 1989), but is also evidence of its high diversity. The species-specific markers were also identified between two populations for *T. zillii* and *O. mossambicus*.

Mapping species-specific SNP markers for ten tilapia species to the reference genome of *O. niloticus* resulted in the loss of many loci/SNPs from the initial DBA. The density of the species-specific SNP markers in standard RADseq ranged from 0.47 SNPs/Mb in LG 3 to 1.53 SNPs/Mb in LG 9, while the ddRADseq ranged from 0 (LG 2,3,7, 9 and 20) to 0.19 SNPs/Mb in LG 23. The

species-specific markers were also distributed evenly across genome in the ddRADseq.

The current study identified the promise of multiple SNP as genetic marker for most of tilapia species discrimination across three genera. The SNP markers from RADseq have great potential to find out specific markers in tilapia species/sub species, where these will be the advanced molecular database for future studies. However, SNP validation needs to be conducted to optimize the number of functional SNP markers in genetic occurrence studies. The validation can be based on SNP markers retrieved across all 10 species or a subset of the detected SNPs across four commercial species using KASP genotyping technology, simple PCR continued with sequencing or SNP chip. To discriminate species *O. niloticus*, *O. mossambicus*, *O. andersonii* and *O. u. hornorum* among 10 species, we can use the SNPs up to five per locus. One of methods to apply these markers is designing the specific primer for particular locus analysed, PCR and sequencing them. So, the usefulness of this marker will depend on purpose or question to address. Due to economic value, most of tilapia cultured from one or a combination of four commercial species. Thus, the SNP markers derived from a subset will be more applicable applied to examine the genetic occurences from these species.

The species-specific markers consisting of 1-2 SNP can be used for simple and quick assay using KASP assay genotyping. When there are few SNPs with many samples or in the case of many SNPs but with less samples (KASP assay requires minimum 24 samples per assay), the KASP method is more flexible than multiplex methods, e.g SNP chip. Despite the cheaper cost

(approximately £10) per assay in comparison to £32, the KASP method also result in shorter turn around time, only 24 hours versus a week (Semagn, et al., 2014). Species-specific markers validation will be pivotal to test the opportunity in the larger set of species, individuals and wider area of populations in tilapia. In this case, further RADseq can also be implemented involving many more species/sub species and populations.

In the case of phylogenetic tree, the shared SNP markers based on ddRADseq can resolve some evolutionary relationships in tilapia species, e.g between *O. macrochir* and *O. andersonii*, while DNA barcode can not distinguish. Specimens for developing species-specific SNP markers in this study were selectively chosen as being pure across 10 species/sub species and their populations. Thus, these markers will be important to answer the question whether certain populations of tilapia species are still pure or already have been hybridized or introgessed both in the farm and the wild.

The strength of these markers is due to their representing in the genome as the results of specific fragments obtained based on restriction enzyme recognition sites. Furthermore, those fragments were sequenced in the high throughput Illumina sequencing, both nuclear and mitochondrial, involving coding and non-coding regions. Sufficient markers were identified across a minimum of 75% loci present in each population, thus allow the markers for distinguishing species, even between sub species *niloticus* and *cancellatus* for *O. niloticus*. Despite its strength, there are a few drawbacks of this study, e.g the low number of samples within species limit the ability to find a specific allele and application for investigating the multiple hybrids and introgression. However, population

diversity across a small number of samples for particular species increases the reliability of retrieving species-specific SNP markers. Further study is still needed to test the capability of species-specific markers to investigate the genetics of one of the most diverse freshwater cultured species, approximately across 140 countries, which are actually only native to Africa and the Middle East.

Future work will be pivotal to undertake some test cases investigating current status of Genetically Improved Farmed Tilapia (GIFT) broodstock in comparison to founder populations, Singaporean *mossambicus,* a hybrid occurrences for instance Molobicus (a hybrid between *O. niloticus* and *O. mossambicus*) (de Verdal et al., 2014), and also red hybrid (*O. niloticus* - *O. mossambicus*, *O. niloticus* - *O. aureus* or multiple species red hybrid). Another study can be addressed to look at hybridization level between an introduced species, *O. niloticus* into native species, *O. andersonii* and *O. macrochir* as reported by Deines et al. (2014) in the Kafue River, Zambia.

To summarise, SNP markers can be retrieved both from nuclear and mitochondrial DNA, but the efficient of discovery will depend on the method, biological divergence, and the depth of study. Some important results of this study are highlight as follows:

- The COI DNA barcode is able to resolve most phylogenetic relationship, but not all, of the tilapia species involved in this study.
- Phylogenetic trees constructed from multiple loci throughout the genome based on RADseq or ddRADseq data appears to be generally highly accurate in resolving evolutionary relationships between tilapia species, including population and subspecies divergence.

- The *ADA* approach was partially successful, where four out of ten SNP markers derived from *ADA* sequences could potentially be applied in identifying and discriminating among tilapia species.
- 38 species-specific SNP markers were identified using ddRADseq across the ten tilapia species (with a minimum 75% of loci present in each species), but most of them are for *T. zillii*.
- A sufficient number of species-specific SNP markers retrieved from a reduced set of the four economically important tilapia species, ranging from 7 in *O. niloticus*, 10 in *O. mossambicus*, 7 in *O. u. hornorum* to 13 in *O. aureus*, and many more identified between species pairs.
- The species-specific markers consisting of 1-2 SNP can be used for simple and quick assay using KASP assay genotyping.
- This study suggests that SNP discovery using ddRADseq with two enzymes restriction *Sbf*I and *Sph*I is very robust because it can be used with a high number of samples, allows production of size-specific fragments and is time efficient.

REFERENCES

Acosta, B. O., & Gupta, M. V. (2010). The genetic improvement of farmed tilapias project: Impact and lessons learned. In S. S. Silva & F. B. Davy (Eds.) (Ed.), *Success Stories in Asian Aquaculture* (pp. 149–171). Dordrecht: Springer, Netherlands.

Adépo-Gourène, B., Gourène, G., & Agnèse, J.-F. (2006). Genetic identification of hybrids between two autochthonous tilapia species, *Tilapia zillii* and *Tilapia guineensis*, in the man-made lake Ayamé. *Aquatic Living Resources, 19*(03), 239–245.

Agnése, J.-F., Adépo-Gourène, B., Abban, E. K., & Fermon, Y. (1997). Genetic differentiation among natural populations of the Nile tilapia *Oreochromis niloticus* (Teleostei, Cichlidae). *Heredity, 79*(1), 88–96.

Agnése, J.-F., Adépo-Gourène, B., Owino, J., Pouyaud, L., & Aman, R. (1999). Genetic characterization of a pure relict population of *Oreochromis esculentus*, an endangered tilapia. *Journal of Fish Biology, 54*, 1119–1123.

Agresti, J. J., Seki, S., Cnaani, A., Poompuang, S., Hallerman, E. M., Umiel, N., Hulata, G. and Gall, G. A. E., & May, B. (2000). Breeding new strains of tilapia: development of an artificial center of origin and linkage map based on AFLP and microsatellite loci. *Aquaculture, 185*(1-2), 43–56.

Aho, T., Rönn, J., Piironen, J., & Björklund, M. (2006). Impacts of effective population size on genetic diversity in hatchery reared Brown trout (*Salmo trutta* L.) populations. *Aquaculture, 253*, 244–248.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2008). *Molecular Biology of the cell*. UK: Garland Science Publishing.

Alexandrino, P. J., Sousa, C., Pereira, A., & Ferrand, N. (1993). Genetic polymorphism of adenosine deaminase (ADA; E.C. 3.5.4.4.) in allis shad, *Alosa alosa* and twaite shad, *Alosa fallax*. *Journal of Fish Biology, 43*, 951–953.

Allendorf, F. W., Luikart, G., & Aitken, S. N. (2013). *Conservation and the Genetics of Populations.* (2nd ed.). London, England: Wiley-Blackwell.

Allendorf, F. W., Mitchell, N., Ryman, N., & Ståhl, G. (1977). Isozyme loci in brown trout (*Salmo trutta* L.): detection and interpretation from population data. *Hereditas, 86*, 179–190.

Anderson, E. (1949). *Introgressive Hybridization*. New York: John Wiley & Sons, Inc.

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T., & Stern, D. L. (2011). Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research, 21*, 610–617.

Angienda, P. O., Lee, H. J., Elmer, K. R., Abila, R., Waindi, E. N., & Meyer, A. (2011). Genetic structure and gene flow in an endangered native tilapia fish (*Oreochromis esculentus*) compared to invasive Nile tilapia (*Oreochromis niloticus*) in Yala swamp, East Africa. *Conservation Genetics, 12* (1), 243-255.

Appleyard, S. A., & Mather, P. B. (2000). Investigation into the mode of inheritance of allozyme and random amplified polymorphic DNA markers in tilapia *Oreochromis mossambicus* (Peters). *Aquaculture Research, 31*, 435–445.

Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology, 22*(11), 3179–90.

Asgharian, H., Sahafi, H. H., Ardalan, A. A., Shekarriz, S., & Elahi, E. (2011). Cytochrome c oxidase subunit 1 barcode data of fish of the Nayband National Park in the Persian Gulf and analysis using meta-data flag several cryptic species. *Molecular Ecology Resources, 11*(3), 461–72.

B-Rao, C., & Majumdar, K. C. (1998). Multivariate map representation of phylogenetic relationships: application to tilapiine fish. *Journal of Fish Biology, 52*(6), 1199–1217.

Bailey, R. M., & Gosline, W. (1955). Variation and Systematic Significance of Vertebral Counts in the American Fishes of the Family Percidae. Miscellaneous publications, Museum of Zoology, University of Michigan, No. 93.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One 3*(10), e3376.

Ballard, J. W. O., & Whitlock, M. C. (2004). The incomplete natural history of mitochondria. *Molecular Ecology, 13*, 729–744.

Barbadilla, A., King, L. M., & Lewontin, R. C. (1996). What does electrophoretic variation tell us about protein variation? *Mol Biol Evol, 13*, 427–432.

Bardakci, F., & Skibinski, D. O. (1994). Application of the RAPD technique in tilapia fish: species and subspecies identification. *Heredity, 73* (Pt 2), 117–23.

Barriga-Sosa, I. D. L. A., Jiménez-Badillo, M. D. L., Ibánez, A. L., & Arredondo-Figueroa, J. L. (2004). Variability of tilapias (*Oreochromis* spp.) introduced in Mexico: Morphometric, meristic and genetic characters. *Journal of Applied Ichthyology, 20*, 7–14.

Bartley, D. M., Rana, K., & Immink, A. J. (2001). The use of inter-specific hybrids in aquaculture and fisheries. *Reviews in Fish Biology and Fisheries, 10*, 325–337.

Bekaert, M. (2014). Genetic-Mapper.

Bhassu, S., Yusoff, K., Panandam, J. M., Embong, W. K., Oyyan, S., & Tan, S. G. (2004). The genetic structure of *Oreochromis* spp.(Tilapia) populations in Malaysia as revealed by microsatellite DNA analysis. *Biochemical Genetics, 42*, 217–229.

Boore, J. L. (1999). Animal mitochondrial genomes. *Nucleic Acid Research*, 27, *1767-80*.

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., Johnson, J., Alcazar, R., Noh, H. J., Russell, P., Aken, B., Alföldi, J., Amemiya, C., Azzouzi, N., Baroiller, J-F., Barloy-Hubler, F., Berlin, A., Bloomquist, R., Carleton, K. L., Conte, M. A., D'Cotta, H., Eshel, O., Gaffney, L., Galibert, F., Gante, H. F., Gnerre, S., Greuter, L., Guyon, R., Haddad, N. S., Haerty, W., Harris, R. M., Hofmann, H. A., Hourlier, T., Hulata, G., Jaffe, D. B., Lara, M., Lee, A. P., MacCallum, I., Mwaiko, S., Nikaido, M., Nishihara, H., Ozouf-Costaz, C., Penman, D. J., Przybylski, D., Rakotomanga, M., Renn, S. C. P., Ribeiro, F. J., Ron, M., Salzburger, W., Sanchez-Pulido, L., Santos, M. E., Searle, S., Sharpe, T., Swofford, R., Tan, F. J., Williams, L., Young, S., Yin, S., Okada, N., Kocher, T. D., Miska, E. A., Lander, E. S., Venkatesh, B., Fernald, R. D., Meyer, A., Ponting, C. P., Streelman, J. T., Lindblad-Toh, K., Seehausen, O., Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature, 513*, 375-381.

Breed, M. & Sanchez, L. (2010). Both Environment and Genetic Makeup Influence Behavior. Nature Education Knowledge 3(10), 68.

Brinez, R. B., Caraballo, X., & Salazar, M. V. (2011). Genetic diversity of six populations of red hybrid tilapia, using microsatellites genetic markers. *Rev. MVZ Cordoba 16*(2), 2491–2498.

Brock, V. E. (1960). The introduction of aquatic animals into Hawaiian waters. *Internationale Revue Der Gesamten Hydrobiologie Und Hydrographie, 45*(4), 463–480.

Brown, W., George, M., & Wilson, A. (1979). Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America, 76*(4), 1967–71.

Brunelli, J. P., Thorgaard, G. H., Leary, R. F., & Dunnigan, J. L. (2008). Single-Nucleotide Polymorphisms Associated with Allozyme Differences between Inland and Coastal Rainbow Trout. *Transactions of the American Fisheries Society, 137*(5).

Cagigas, M. E., Vazquez, E., Blanco, G., & Sanchez, J. A. (1999). Genetic effects of introduced hatchery stocks on indigenous brown trout (*Salmo trutta* L.) populations in Spain. *Ecology of Freshwater Fish 8*, 141–150.

Campbell, N. R., Amish, S. J., Pritchard, V. L., McKelvey, K. S., Young, M. K., Schwartz, M. K., Garza, J. C., Luikart, G., & Narum, S. R. (2012). Development and evaluation of 200 novel SNP assays for population genetic studies of westslope cutthroat trout and genetic identification of related taxa. *Molecular Ecology Resources, 12*(5), 942–9.

Carleton, K. L., Streelman, J. T., Lee, B.-Y., Garnhart, N., Kidd, M., & Kocher, T. D. (2002). Rapid isolation of CA microsatellites from the tilapia genome. *Animal Genetics, 33*(2), 140–144.

Catchen, J., Hohenlohe, P., Bassham, S., Amores, A., & Cresko, W. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology, 22*(11), 3124–40.

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: building and genotyping loci de novo from short-read sequences. *G3 (Bethesda, Md.), 1*(3), 171–82.

Chat, J., Manicki, A., & Merchermek, N. (2008). Typing for brown trout LDH-C1 alleles together with microsatellites by automated sequencing. *Conservation Genetics, 9*(6), 1669–1671.

Chattopadhyay, B., Garg, K. M., & Ramakrishnan, U. (2014). Effect of diversity and missing data on genetic assignment with RAD-Seq markers. *BMC Research Notes, 7*, 841.

Chervinski, J. (1965). Sexual Dimorphism in Tilapia. *Nature, 208*(5011), 703.

Chervinski, J. (1971). Sexual dimorphism in *Tilapia aurea* and *Tilapia zillii* from Dor and from Lake Tiberias. *Bamidgeh, 23*(2), 56–59.

Chervinski, J. (1983). Brief Technical Note: Sexual Dimorphism in Tilapias. *Science*, *35*, 171–172.

Cnaani, A., Hallerman, E. M., Ron, M., Weller, J. I., Indelman, M., Kashi, Y., Gall, G. A. E., & Hulata, G. (2003). Detection of a chromosomal region with two quantitative trait loci, affecting cold tolerance and fish size, in an F2 tilapia hybrid. *Aquaculture*, *223*, 117–128.

Cnaani, A., Lee, B.-Y., Zilberman, N., Ozouf-Costaz, C., Hulata, G., Ron, M., D'Hont, A., Baroiller, J-F., D'Cotta, H., Penman, D. J., Tomasino, E., Coutanceau, J-P., Pepey, E., Shirak, A., & Kocher, T. D. (2008). Genetics of sex determination in tilapiine species. *Sexual Development : Genetics, Molecular Biology, Evolution, Endocrinology, Embryology, and Pathology of Sex Determination and Differentiation*, *2*(1), 43–54.

Costa-Pierce, B. A. (2003). Rapid evolution of an established feral tilapia (*Oreochromis* spp.): the need to incorporate invasion science into regulatory structures. *Biological Inv*. *5*, 71–84.

Cristalli, G., Costanzi, S., Lambertucci, C., Lupidi, G., Vittori, S., Volpini, R., & Camaioni, E. (2001). Adenosine deaminase: Functional implications and different classes of inhibitors, Adenosine deaminase: Functional implications and different classes of inhibitors. *Medicinal Research Reviews*, *21*, 105–128.

D'Amato, M. E., Esterhuyse, M. M., Waal, B. C. W., Brink, D., & Volckaert, F. A. M. (2007). Hybridization and phylogeography of the Mozambique tilapia *Oreochromis mossambicus* in Southern Africa evidenced by mitochondrial and microsatellite DNA genotyping. *Conservation Genetics*, *8*(2), 475–488.

Da Silva, A. S., Bellé, L. P., Bitencourt, P. E. R., Perez, H. A. G., Thomé, G. R., Costa, M. M., … Monteiro, S. G. (2011). *Trypanosoma evansi*: Adenosine deaminase activity in the brain of infected rats. *Experimental Parasitology*, *127*(1), 173–177.

Datta, N. C., & Roy, P. K. (1984). Urinogenital system of the exotic cichlid *Sarotherodon* and *Mossambica* (Peters). *International Journal of the Academy of Ichthyology*, *5*, 49–54.

Davey, J., & Blaxter, M. (2010). RADseq: Next-generation population genetics. *Briefings in Functional Genomics*, *9*(5), 416–423.

Davey, J., Hohenlohe, P., Etter, P., Boone, J., Catchen, J., & Blaxter, M. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499–510.

Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., & Blaxter, M. L. (2013). Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology, 22*(11), 3151–64.

De Donato, M., Peters, S. O., Mitchell, S. E., Hussain, T., & Imumorin, I. G. (2013). Genotyping-by-Sequencing (GBS): A Novel, Efficient and Cost-Effective Genotyping Method for Cattle Using Next-Generation Sequencing. *PLoS ONE*, *8*(5).

De Silva, C. D., & Ranasinghe, J. (1989). Biochemical evidence of hybrid gene introgression in some reservoir populations of tilapia in Southern Sri Lanka. *Aquaculture Research, 20*(3), 269–277.

De Verdal, H., Rosario, W., Vandeputte, M., Muyalde, N., Morissens, P., Baroiller, J.-F., & Chevassus, B. (2014). Response to selection for growth in an interspecific hybrid between *Oreochromis mossambicus* and *O. niloticus* in two distinct environments. *Aquaculture*, *430*, 159–165.

Deines, A., Bbole, I., Katongo, C., Feder, J., & Lodge, D. (2014). Hybridisation between native *Oreochromis* species and introduced Nile tilapia *O. niloticus* in the Kafue River, Zambia. *African Journal of Aquatic Science, 39* (1), 37–41

Dial, R. S. and Wainright, S. C. (1983). New distributional records for non-native fishes in Florida. Florida Scientist 46, 8-15.

Dinesh, K. R., Lim, T. M., Chan, W. K., & Phang, V. P. E. (1996). Genetic variation inferred from RAPD fingerprinting in three species of tilapia. *Aquaculture International, 4* (1), 19-30.

Dong, R. P., Kameoka, J., Hegen, M., Tanaka, T., Xu, Y., Schlossman, S. F., & Morimoto, C. (1996). Characterization of adenosine deaminase binding to human CD26 on T cells and its biologic role in immune response. *Journal of Immunology (Baltimore, Md. : 1950), 156*, 1349–1355.

Du, Y., Jiang, H., Chen, Y., Li, C., Zhao, M., Wu, J., Qiu, Y., Li, Q., & Zhang, X. (2012). Comprehensive evaluation of SNP identification with the Restriction Enzyme-based Reduced Representation Library (RRL) method. *BMC Genomics, 13*, 77.

Dunham, R. A. (2011). *Aquaculture and fisheries biotechnology: genetic approaches* (2nd ed., p. 504). Wallingford, UK: CABI.

Dunz, A. R., & Schliewen, U. K. (2013). Molecular phylogeny and revised classification of the haplotilapiine cichlid fishes formerly referred to as "*Tilapia*". *Molecular Phylogenetics and Evolution, 68*(1), 64–80.

Eaton, D. A. R. (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics (Oxford, England), 30*(13), 1844–9.

Edmunds, R. C., van Herwerden, L., Smith-Keune, C., & Jerry, D. R. (2009). Comparative characterization of a temperature responsive gene (lactate dehydrogenase-B, ldh-b) in two congeneric tropical fish, *Lates calcarifer* and *Lates niloticus*. *International Journal of Biological Sciences, 5*(6), 558–569.

El-Sayed, A. F. M. (2006). *Tilapia Culture*. CABI Publishing.

El-Serafy, S. S., Abdel-Hameid, N.-A. H., Awwad, M. H., & Azab, M. S. (2007). DNA riboprinting analysis of Tilapia species and their hybrids using restriction fragment length polymorphisms of the small subunit ribosomal DNA. *Aquaculture Research, 38*(3), 295–303.

El- Shazly, A. (1993). Biological studies on four cichlid fishes (*Tilapia nilotica*, *T. galilae*, *T. zillii*, *T. aurea*). MSc Thesis. Zagazig University, Egypt.

Emerson, K. J., Merz, C. R., Catchen, J. M., Hohenlohe, P. A., Cresko, W. A., Bradshaw, W. E., & Holzapfel, C. M. (2010). Resolving postglacial phylogeography using high-throughput sequencing. *Proc. Natl. Acad. Sci. USA, 107* (37).

Espinosa-Lemus, V., Arredondo-figueroa, J. L., & Barriga-sosa, I. D. L. A. (2009). Morphometric and genetic characterization of tilapia (Cichlidae : Tilapiini) stocks for effective fisheries management in two Mexican reservoirs. *Hidrobiologica, 19*(2), 95–107.

Etter, P. D., Preston, J. L., Bassham, S., Cresko, W. A, & Johnson, E. A. (2011). Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PloS One, 6*(4), e18561.

Everett, M. V., Grau, E. D., & Seeb, J. E. (2011). Short reads and nonmodel species: Exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources, 11*, 93–108.

Falter, U., & Dufayt, O. (1991). Behavioural isolation mechanisms in tilapia spp.: Courtship sequences in intra and interspecific encounters. *Annales - Musee Royal de l'Afrique Centrale. Sciences Zoologiques*, 262, 59–63.

FAO. (2014). *FAO year book 2012*.

Fishelson, L. (1996). Cichlidae of the genus Tilapia in Israel. *Bamidgeh, 18*, 67–80.

Fredholm, B. B., Chen, J. F., Cunha, R. A., Svenningsson, P., & Vaugeois, J. M. (2005). Adenosine and Brain Function. *International Review of Neurobiology*, *63*, 191–270.

Freeland, J. R., Kirk, H., & Peterson, S. D. (2011). *Molecular Ecology* (Second). Wiley-Blackwell.

Gaston, K. J. (2003). *The Structure and Dynamics of Geographic Ranges*. Oxford, UK: Oxford University Press.

Goodwin, N. B., Balshine-Earn, S., & Reynolds, J. D. (1998). Evolutionary transitions in parental care in cichlid Fish. *Proc. R. Soc. Lond. B*, *265*, 2265–2272.

Griffiths, A., Gelbart, W., Lewontin, R., & Miller, J. (2002). *Modern Genetic Analysis : Integrating Genes and Genomes.* New York: Freeman and Co.

Griffiths, A., Wessler, S., Lewontin, R., & Caroll, S. (2008). *Introduction to genetic Analysis* (9th ed.). New York: Freeman, W.H and Co.

Gu, W., Gurguis, C. I., Zhou, J., Zhu, Y., Ko, E.-A., Ko, J.-H., Wang, T., & Zhou, T. (2015). Functional and structural consequence of rare exonic single nucleotide polymorphisms: one story, two tales. *Genome Biology and Evolution*, *7*(10), 2929-2940.

Guyon, R., Rakotomanga, M., Azzouzi, N., Coutanceau, J. P., Bonillo, C., D'Cotta, H., Pepey, E., Soler, L., Rodier-Goud, M., D'Hont, A., Conte, M. A., Van Bers, N. E. M., Penman, D. J., Hitte C, Crooijmans, R. P. M. A, Kocher TD, Ozouf-Costaz C, Baroiller J. F., Galibert, F. (2012). A high-resolution map of the Nile tilapia genome: a resource for studying cichlids and other percomorphs. *BMC Genomics 13*, 222.

Hamblin, M. T., & Rabbi, I. Y. (2014). The Effects of Restriction-Enzyme Choice on Properties of Genotyping-by-Sequencing Libraries: A Study in Cassava (*Manihot esculenta*). *Crop Science, 54*, 2603-2608.

Hand, B., Hether, T., Kovach, R., Muhlfeld, C., Amish, S., Boyer, M., O'rourke, S. M., Miller, M. R., Lowe, W. H., Hohenlohe, P. A., & Luikart, G. (2015). Genomics and introgression : Discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. *Current Zoology*, *61*(1), 146–154.

Hartl, D., & Jones, E. (2008). *Genetics : Principles and Analysis* (4th ed.). Jone and Bartlett.

Hartl, D. L., & Clark, A. G. (1997). *Principles of Population genetics* (Third). Sinauer Associates, USA.

Harvey, S. C., Boonphakdee, C., Campos-Ramos, R., Ezaz, M. T., Griffin, D. K., Bromage, N. R., & Penman, P. (2003). Analysis of repetitive DNA sequences in the sex chromosomes of *Oreochromis niloticus*. *Cytogenetic and Genome Research, 101*(3-4), 314–319.

Hassanien, H., Elnady, M., Obeida, A., & Itriby, H. (2004). Genetic diversity of Nile tilapia populations revealed by randomly amplified polymorphic DNA (RAPD). *Aquaculture Research, 35*(6), 587–593.

He, A., Luo, Y., Yang, H., Liu, L., Li, S., & Wang, C. (2011). Complete mitochondrial DNA sequences of the Nile tilapia (*Oreochromis niloticus*) and Blue tilapia (*Oreochromis aureus*): genome characterization and phylogeny applications. *Molecular Biology Reports, 38*(3), 2015–21.

Hedrick, P. W. (2005). *Genetics of Populations* (Third, p. 737). Jones and Bartlett.

Hickling, C. F. (1960). The Malacca tilapia hybrids. *Journal of Genetics, 57*(1), 1–10.

Hirschhorn, R., Yang, D. R., & Israni, A. (1994). An Asp8Asn substitution results in the adenosine deaminase (ADA) genetic polymorphism (ADA 2 allozyme): occurrence on different chromosomal backgrounds and apparent intragenic crossover. *Annals of Human Genetics, 58*, 1–9.

Hohenlohe, P. A, Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A, & Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics, 6*(2), e1000862.

Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W., & Luikart, G. (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources, 11*, 117–122.

Hong-Mei, S., Jun-Jie, B., Ying-Chun, Q., & Sheng-Jie, L. (2009). Identification and structure analysis of three tilapia species using microsatellite markers. *Chinese Journal of Agricultural Biotechnology 6*(02), 119.

Houston, R. D., Davey, J. W., Bishop, S. C., Lowe, N. R., Mota-Velasco, J. C., Hamilton, A., Guy, D. R., Tinch, A. E., Thomson, M. L., Blaxter, M. L., Gharbi, K., Bron, J. E., & Taggart, J. B. (2012). Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. *BMC Genomics, 13*(1), 244.

Houston, R. D., Taggart, J. B., Cézard, T., Bekaert, M., Lowe, N. R., Downing, A., Talbot, R., Bishop, S. C., Archibald, A. L., Bron, J. E., Penman, D. J.,

Davassi, A., Brew, F., Tinch, A. E., Gharbi, K., & Hamilton, A. (2014). Development and validation of a high density SNP genotyping array for Atlantic salmon (*Salmo salar*). *BMC Genomics, 15*, 90.

Kaneijam, H., Tanaka, T., Nojima, Y., Schlossman, S., & C, M. (1993). Direct association of adenosine deaminase with a T cell activation antigen, CD26. *Science, 261*, 466–469.

Klett, V., & Meyer, A. (2002). What, if Anything, is a Tilapia ?— Mitochondrial ND2 Phylogeny of Tilapiines and the Evolution of Parental Care Systems in the African Cichlid Fishes. *Mol. Biol. Evol., 19* (6), 865–883.

Kucuktas, H., & Liu, Z. (2007). Allozyme and mitochondrial markers. In Z. Liu (Ed.), *Aquaculture Genome Technologies* (pp. 73–85). Blackwell publishing.

Lartillot, N., Lepage, T., & Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics (Oxford, England), 25*(17), 2286–8.

Lee, B.-Y., Lee, W.-J., Streelman, J. T., Carleton, K. L., Howe, A. E., Hulata, G., Slettan, A., Stern, J. E., Terai, Y., & Kocher, T. D. (2005). A second-generation genetic linkage map of tilapia (*Oreochromis* spp.). *Genetics*, *170*(1), 237–44.

Lee, W., & Kocher, T. (1996). Microsatellite DNA markers for genetic mapping in *Oreochromis niloticus*. *Journal of Fish Biology*, *49*, 169-171.

Lemmon, E. M., & Lemmon, A. R. (2013). High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 99–121.

Li, W. (1997). *Molecular evolution*. Sunderland, MA.: Sinauer Associates.

Liu, Z. J. (2007). *Aquaculture Genome Technologies* (First). Blackwell publishing.

Liu, Z. J. (2011). *Next Generation Sequencing and Whole Genome Selection in Aquaculture* (p. 232). Wiley-Blackwell.

Liu, Z. J., & Cordes, J. F. (2004). DNA marker technologies and their applications in aquaculture genetics. *Aquaculture*, *238*, 1–37.

Macaranas, J., Agustin, L., Ablan, M., Pante, M., Eknath, A., & Pullin, P. (1995). Genetic improvement of farmed tilapias: biochemical characterisation of strain differences in Nile tilapia. *Aquaculture International*, *3*, 43–54.

Majumdar, K. C., & McAndrew, B. J. (1986). Relative DNA content of somatic nuclei and chromosomal studies in three genera, *Tilapia*, *Sarotherodon*, and *Oreochromis* of the tribe Tilapiini (Pisces, Cichlidae). *Genetica*, *68*(3), 175–188.

Marzano, F. N., Corradi, N., Papa, R., Tagliavini, J., & Gandolfi, G. (2003). Molecular evidence for introgression and loss of genetic variability in *Salmo* (trutta) *macrostigma* as a result of massive restocking of Apennine populations (Northern and Central Italy). *Environmental Biology of Fishes*, *68*, 349–356.

Mazzuchelli, J., Kocher, T. D., Yang, F., & Martins, C. (2012). Integrating cytogenetics and genomics in comparative evolutionary studies of cichlid fish. *BMC Genomics*, *13*(1), 463.

Mcandrew, B. J. (1993). Sex control in tilapiines. In *Recent Advances in Aquaculture IV* (pp. 87–98).

McAndrew, B. J., & Majumdar, K. C. (1983). Tilapia stock identification using electrophoretic markers. *Aquaculture, 30*, 249-261.

McAndrew, B. J., Roubal, F. R., Roberts, R. J., Bullock, A. M., & McEwen, I. M. (1988). The genetics and histology of red, blond and associated colour variants in *Oreochromis niloticus*. *Genetica*, *76*(2), 127–137.

McCusker, M. R., Denti, D., Van Guelpen, L., Kenchington, E., & Bentzen, P. (2013). Barcoding Atlantic Canada's commonly encountered marine fishes. *Molecular Ecology Resources*, *13*(2), 177–88.

McKinna, E., Nandlal, S., Mather, P., & Hurwood, D. A. (2010). An investigation of the possible causes for the loss of productivity in genetically improved farmed tilapia strain in Fiji: inbreeding versus wild stock introgression. *Aquaculture Research*, *41*(11), e730–e742.

Messmer, A. M., Rondeau, E. B., Jantzen, S. G., Lubieniecki, K. P., Davidson, W. S., & Koop, B. F. (2011). Assessment of population structure in Pacific *Lepeophtheirus salmonis* (Krøyer) using single nucleotide polymorphism and microsatellite genetic markers. *Aquaculture*, *320*(3-4), 183–192.

Micha, J. C., Cuvelier, R., Tilquin, C., Muraille, B., Bourgois, M., & Falter, U. (1996). Comparative growth of hybrids (F sub(1), F sub(2) and F sub(3)) of *Oreochromis niloticus* (L.) and *O. macrochir* (Blgr.). In R. S. V Pullin, J. Lazzard, M. Legendre, J. B. A. Kothias, & D. Pauly (Eds.), *The third International Symposium on Tilapia in Aquaculture* (pp. 354–360). International Center for Living Aquatic Resources Management.

Mota-Velasco, J. C., Ferreira, I. A., Cioffi, M. B., Ocalewicz, K., Campos-Ramos, R., Shirak, A., Lee, B-Y., Martins, C., & Penman, D. J. (2010). Characterisation of the chromosome fusions in *Oreochromis karongae*. *Chromosome Research : An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *18*(5), 575–86.

Nagl, S., Tichy, H., Mayer, W. E., Samonte, I. E., McAndrew, B. J., & Klein, J. (2001). Classification and phylogenetic relationships of African tilapiine fishes inferred from mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, *20*(3), 361–74.

Ndiwa, T. C., Nyingi, D. W., & Agnese, J.-F. (2014). An important natural genetic resource of *Oreochromis niloticus* (Linnaeus, 1758) threatened by aquaculture activities in Loboi drainage, Kenya. *PloS One*, *9*(9), e106972.

Nei, M., & Kumar, S. (2000). *Molecular Evolution and Phylogenetics* (1st ed.). Oxford: Oxford University Press.

Nyingi, D., De Vos, L., Aman, R., & Agnèse, J.-F. (2009). Genetic characterization of an unknown and endangered native population of the Nile tilapia *Oreochromis niloticus* (Linnaeus, 1758) (Cichlidae; Teleostei) in the Loboi Swamp (Kenya). *Aquaculture*, *297*(1-4), 57–63.

Nyingi, D. W., & Agnèse, J.-F. (2007). Recent introgressive hybridization revealed by exclusive mtDNA transfer from *Oreochromis leucostictus* (Trewavas, 1933) to *Oreochromis niloticus* (Linnaeus, 1758) in Lake Baringo, Kenya. *Journal of Fish Biology*, *70* (Supplement A), 148–154.

O'Brien, S. J. (1991). Mammalian genome mapping: lessons and prospects. *Current Opinion in Genetics & Development*, *1*(1), 105–111.

Oliveira, C., Chew, J. S., Porto-Foresti, F., Dobson, M. J., & Wright, J. M. (1999). A LINE2 repetitive DNA sequence from the cichlid fish, *Oreochromis niloticus*: sequence analysis and chromosomal distribution. *Chromosoma*, *108*(7), 457–68.

Oliveira, C., & Wright, J. M. (1998). Molecular cytogenetic analysis of heterochromatin in the chromosomes of tilapia, *Oreochromis niloticus* (Teleostei: Cichlidae). *Chromosome Research : An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *6*(3), 205–11.

Oliveira, R. F., & Almada, V. C. (1995). Sexual dimorphism and allometry of external morphology in *Oreochromis mossambicus*. *Fish Biology*, *46*, 1055–1064.

Palaiokostas, C., Bekaert, M., Khan, M. G. Q., Taggart, J. B., Gharbi, K., McAndrew, B. J., & Penman, D. J. (2013). Mapping and validation of the major sex-determining region in Nile tilapia (*Oreochromis niloticus* L.) Using RAD sequencing. *PloS One*, *8*(7), e68389.

Palaiokostas, C., Bekaert, M., Khan, M. G., Taggart, J. B., Gharbi, K., McAndrew, B. J., & Penman, D. J. (2015). A novel sex-determining QTL in Nile tilapia (*Oreochromis niloticus*). *BMC Genomics*, *16*, 1–10.

Payne, A., & Collinson, R. (1983). A comparison of the biological characteristics of. *Aquaculture*, *30*, 335–351.

Penman, D. J., & McAndrew, B. J. (2000). Tilapias: Biology and Exploitation. In M. C. M. Beveridge & B. J. McAndrew (Eds.), *Tilapias : Biology and Exploitation* (pp. 227–266). Dordrecht: Springer Netherlands.

Peters, H., & Berns, S. (1982). Die Maulbrutpflege der Cichliden. Untersuchungen zur Evolution eines Verhaltensmusters. *Z. Zool. Syst. Evolutionforsch*, *20*, 18–52.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS One*, *7*(5), e37135.

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE*, *7*(2).

Ponzoni, R. W., Nguyen, N. H., Khaw, H. L., Hamzah, A., Bakar, K. R. A., & Yee, H. Y. (2011). Genetic improvement of Nile tilapia (*Oreochromis niloticus*) with special reference to the work conducted by the WorldFish Center with the GIFT strain. *Reviews in Aquaculture*, *3*(1), 27–41.

Popma, T. J., & Lovshin, L. L. (1995). *Worldwide Prospects for Commercial Production of Tilapia*. International Center for Aquaculture and Aquatic Environments. *Department of Fisheries and Allied Aquacultures, Auburn University, Alabama* (p. 42).

Pouyaud, L., & Agnese, J.-F. (1995). Phylogenetic relationships between 21 species of three tilapiine genera *Tilapia*, *Sarotherodon* and *Oreochromis* using allozyme data. *Journal of Fish Biology*, *47*(1), 26–38.

Primrose, S., & Twyman, R. (2006). *Principles of Gene Manipulation and Genomics*. (Blackwell Publishing, Ed.) (Seventh).

Prunet, P., & Bornancin, M. (1989). Physiology of salinity tolerance in tilapia: an update of basic and applied aspects. *Aquatic Living Resources*, *2*(2), 91–97.

Pullin, R. S. V. (1988). Tilapia genetic resources for aquaculture. In *The Second International Symposium on Tilapia in Aquaculture* (p. 108). ICLARM Conf. Proc. 16.

Reinhardt, J. A., Baltrus, D. A., Nishimura, M. T., Jeck, W. R., Jones, C. D., Dangl, J. L., Hill, C., & Carolina, N. (2009). De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. Genome Res., *19*(2), 294–305.

Robinson, P., & Holme, J. (2011). KASP version 4.0 SNP Genotyping Manual, 1–8.

Rognon, X., Andriamanga, M., & Mcandrew, B. (1996). Allozyme variation in natural and cultured populations in two tilapia species : *Oreochromis niloticus* and *Tilapia zillii*. *Heredity*, *76*, 640-650.

Rognon, X., & Guyomard, R. (1997). Mitochondrial DNA differentiation among East and West African Nile tilapia populations. *Journal of Fish Biology*, *51*(1), 204–7.

Rognon, X., & Guyomard, R. (2003). Large extent of mitochondrial DNA transfer from *Oreochromis aureus* to *O . niloticus* in West Africa. *Molecular Ecology*, *12*, 435–445.

Rosemberg, D. B., Rico, E. P., Guidoti, M. R., Dias, R. D., Souza, D. O., Bonan, C. D., & Bogo, M. R. (2007). Adenosine deaminase-related genes: Molecular identification, tissue expression pattern and truncated alternative splice isoform in adult zebrafish (*Danio rerio*). *Life Sciences*, *81*, 1526–1534.

Roskov, Y., Kunze, T., Orrell, T., Abucay, L., Culham, A., Bailly, N., Kirk P, Bourgoin, T., De Walt, R. E., Decock, W., De Wever, A. (2014). Species 2000 & ITIS Catalogue of Life.

Ross, L. G. (2000). Beveridge, M. C. M. and Mc. Andrew, B. J. In *Tilapias : Biology and Exploitation* (Fish and F). Kluwer Academic Publishers.

Rubin, B. E. R., Ree, R. H., & Moreau, C. S. (2012). Inferring phylogenies from RAD sequence data. *PloS One*, *7*(4), e33394.

Salini, J. P., Milton, D. A., Rahman, M. J., & Hussain, M. G. (2004). Allozyme and morphological variation throughout the geographic range of the tropical shad, hilsa *Tenualosa ilisha*. *Fisheries Research*, *66*, 53–69.

Sánchez, C. C., Smith, T. P. L., Wiedmann, R. T., Vallejo, R. L., Salem, M., Yao, J., & Rexroad, C. E. (2009). Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, *10*, 559.

Sarropoulou, E., Nousdili, D., Magoulas, A., & Kotoulas, G. (2008). Linking the genomes of nonmodel teleosts through comparative genomics. *Marine Biotechnology*, *10*, 227–233.

Seeb, J. E., Carvalho, G., Hauser, L., Naish, K., Roberts, S., & Seeb, L. W. (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, *11 Suppl 1*, 1–8.

Seitz, A. (1949). Vergleichende Verhaltensstudien an Buntbarschen: (Cichlidae). *Zeitschrift Für Tierpsychologie*, *6*, 202–235.

Semagn, K., Babu, R., Hearne, S., & Olsen, M. (2014). Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. *Molecular Breeding*, *33*(1), 1–14.

Shirak, A., Cohen-zinder, M., Barroso, R. M., Seroussi, E., Ron, M., & Hulata, G. (2009). DNA Barcoding of Israeli Indigenous and Introduced Cichlids. *Israeli Journal of Aquaculture*, *61*(2), 83–88.

Sodsuk, P. K., McAndrew, B. J., & Turner, G. F. (1995). Evolutionary relationships of the Lake Malawi *Oreochromis* species: Evidence from allozymes. *Journal of Fish Biology*, *47*, 321–333.

Sodsuk, P., & McAndrew, B. J. (1991). Molecular systematics of three tilapiine genera *Tilapia*, *Sarotherodon* and *Oreochromis* using allozyme data. *Journal of Fish Biology*, *39* (Supplement SA), 301–308.

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

Streelman, J. T. & Kocher, T. D. (2002). Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. *Physiol. Genomics*, 9, 1-4.

Tautz, D. (1989). Hypervariability of simple sequences as general source for polymorphic DNA markers. *Nucleic Acid Research*, *17* (16), 6463–6471.

Thodesen (Da-Yong Ma), J., Rye, M., Wang, Y. X., Li, S. J., Bentsen, H. B., & Gjedrem, T. (2013). Genetic improvement of tilapias in China: Genetic parameters and selection responses in growth, pond survival and cold-water

tolerance of blue tilapia (*Oreochromis aureus*) after four generations of multi-trait selection. *Aquaculture*, *396-399*, 32–42.

Thong, F. S. L., Lally, J. S. V., Dyck, D. J., Greer, F., Bonen, A., & Graham, T. E. (2007). Adenosine Receptor Increases Insulin-Stimulated Glucose Transport in Isolated Rat Soleus Muscle. *Applied Physiology, Nutrition, and Metabolism*, *32*, 701–710.

Toniato, J., Penman, D. J., & Martins, C. (2010). Discrimination of tilapia species of the genera *Oreochromis*, *Tilapia* and *Sarotherodon* by PCR-RFLP of 5S rDNA. *Aquaculture Research*, *41*(6), 934–938.

Trewavas, E. (1983). *Tilapiine Fishes of the genera Sarotherodon, Oreochromis and Danakilia.* (p. 583). London: British Museum (Natural History).

Van Bers, N., Crooijmans, Rpm., Groenen, M., Dibbits, B., & Komen, J. (2012). SNP marker detection and genotyping in tilapia. *Molecular Ecology Resources*, *12*(5), 932–41.

Vreven, E. J., Adepo-Gourene, B., Agnèse, J. F., & Teugels, G. G. (1998). Morphometric and allozyme variation in natural populations and cultured strains of the Nile tilapia *Oreochromis niloticus* (Teleostei, Cichlidae). *Belgian Journal of Zoology*, *128*, 23–34.

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R., & Hebert, P. D. N. (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360* (1462), 1847–57.

Webster, M. (2006). *Introduction to Geometric Morphometrics*. Department of the Geophysical Sciences, Univ. Chicago.

Welsh, J., & McClelland, M. (1991). Genomic fingerprints produced by PCR with consensus tRNA gene primers. *Nucleic Acids Research*, *19*(4), 861–866.

Wilkins, M. R., & Williams, K. L. (1997). Cross-Species Protein Identification using Amino Acid Composition, Peptide Mass Fingerprinting, Isoelectric Point and Molecular Mass: A Theoretical Evaluation. *Journal of Theoretical Biology*, *186*, 7–15.

Willing, E.-M., Hoffmann, M., Klein, J. D., Weigel, D., & Dreyer, C. (2011). Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics (Oxford, England)*, *27*(16), 2187–93.

Wirgin, I. I., & Waldman, J. R. (1994). What DNA can do for you. *Fisheries*, *19* (7).

Wohlfarth, G. W., & Hulata, G. (1983). *Applied genetics of Tilapias* (p. 26). International Center for Living Aquatic Resources Management.

Wu, L., & Yang, J. (2012). Identifications of captive and wild tilapia species existing in Hawaii by mitochondrial DNA control region sequence. *PloS One*, *7*(12), e51731.

Xia, J. H., Wan, Z. Y., Ng, Z. L., Wang, L., Fu, G. H., Lin, G., Liu, F., Yue, G. H. (2014). Genome-wide discovery and in silico mapping of gene-associated SNPs in Nile tilapia. *Aquaculture*, *432*, 67–73.

Xu, P., Wang, S., Liu, L., Peatman, E., Somridhivej, B. Thimmapuram, J., Gong, G., & Liu, Z. (2006). Channel catfish BAC-end sequences for marker development and assessment of syntenic conservation with other fish species. *Anim Genetics*, *37*(4), 321 – 326.

Yun, F., Zhong, J., Xie, X., Ye, W., Lin, B., Chen, H., & Zhang, H. (2008). Genetic diversity and specific AFLP bands in three cultured tilapia strains. *South China Fisheries Science*, *4*(6).

Zane, L., Bargelloni, L., & Patarnello, T. (2002). Strategies for microsatellite isolation: A review. *Molecular Ecology*, *11*, 1–16.

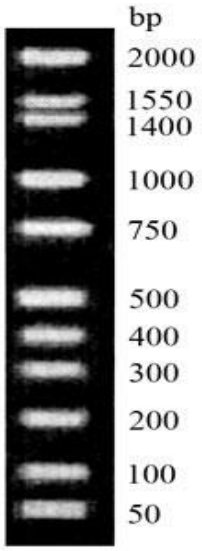
APPENDICES

Chapter II

Appendix II.1 Key consumables used in SNP markers development in tilapia

No	Consumable	Supplier	Quantity/Note
1.	1.7 mL microfuge tubes	Axygen	
2.	0.2 mL PCR tubes	Thermo	
3.	Filter tips (Box of 96)	Axygen	
4.	Preparative agarose gel	Bioline	
5.	Q5 Hot Start HF 2x master mix	NEB	500 reactions (50 µL)
6.	3 M NaAce pH 5.2	Sigma	100 mL
7.	Adhesive plate seals (100)	Thermo	
8.	96 well plates low profile	StarLabs	
9.	Qubit dsDNA BR assay kit	Invitrogen	100 assays
10.	*Sbf*I HF	NEB	2500U @20U/ µL
11.	*Sph*I HF	NEB	2500U @20U/ µL
12.	T4 DNA ligase	NEB	2500U @20U/ µL
13.	rATP	Promega	400 µL @ 100 mM
14	Minelute PCR clean up kit	Qiagen	50 preps
15.	Minelute gel extraction kit	Qiagen	50 preps
16.	AMPure XP magnetic beads	Beckman Coulter	5 mL

Appendix II.2 DNA/Hind III Digested DNA



1.5% agarose mini-gel
80V for 2 hours

Chapter III

Appendix III.1 The list of chemicals used in the RNA extraction and RT-PCR

RNA Extraction

No	Component	Source	Remark
1	Mini-beadbeater homogeniser	Biospec	
2	Centrifuge - SciQuip 4K15	Sigma	
3	Pipettes (P20, P200, P1000)	Gilson	
4	Nuclease-free aerosol barrier tips	Axygen	
5	TRI Reagent/Trizol	Ambion, Sigma, ABgene, Invitrogen	*@ 4ºC*
6	1-bromo-3-chloropropane [BCP]	Sigma (B9673)	*@ RT*
7	100% isopropanol (propan-2-ol, isopropyl Alcohol and 2-propanol)	Fluka (59304)	*@ 4ºC*
8	75% ethanol	Fisher Scientific (E/0650DF/21)	at 4ºC
9	Nuclease-free 1.5ml microfuge tubes	Axygen (MCT-175-C)	
10	Nuclease-free dH_2O		
11	RNA Precipitation Solution		at 4ºC
12	Dri-block (heat block)	Techne	Set @ 70ºC
13	Nanodrop	LabTech	

Reverse Transcription

a. High Capacity cDNA Reverse Transcription Kit :

10x RT Buffer

10x dNTP Mix (100mM)

MultiScribeTMReverse Transcriptase (50 U/µl)

b. Oligo dT

c. Nuclease-free water

Polymerase Chain Reaction (PCR)

a. 2x My Taq (My Taq buffer, dNTPs, $MgCl_2$, enhancers and stabilizers).
b. Primer-mix *ADA*-ON-1089 (10 µM)
c. Primer-mix *ADA*-ON-659 (10 µM)
d. Nuclease free-water
e. cDNA template
f. DNA genome (50 ng/µl

Appendix III.2 Number of species and origin used in *ADA* SNP assays genotyping.

No	Species/sub species	Population	Origin	n
1.	a) *O. niloticus niloticus*	a. Stirling	L. Manzala, Egypt	6
		b. Kpandu	Ghana	12
		c. Nyinuto	Ghana	12
		d. Pandu	Indonesia	3
		e. Kunti	Indonesia	3
		f. Larasati	Indonesia	3
	b)*O. niloticus cancellatus*	a. Hora	Ethiopia	13
		b. Koka	Ethiopia	12
		c. Metahara	Ethiopia	8
	Sub total 1			**63**
2.	*O. mossambicus*	a. Stirling	Zimbabwe	5
		b. Natal	South Africa	10
	Sub total 2			**15**
3.	*O. aureus*	a. Stirling	L. Manzala, Egypt	5
		b. Ain Faskha	Israel	10
	Sub total 3			**15**
4.	*O. karongae*	Stirling	L. Malawi, Tanzania	5
5.	*O. u. hornorum*	Israel	Israel	5
6.	*T. zillii*	a. Stirling	L. Manzala, Egypt	5
		b. Ghana	Ghana	5
	Sub total 6			**10**
7.	*S. galilaeus*	Israel	Israel	5
8.	*O. andersonii*	Itezhi-tezhi	Zambia	6
9.	*O. macrochir*	Itezhi-tezhi	Zambia	4
10	*S. melanotheron*	Ghana	Ghana	4
	Total samples			**148**

Appendix 3.3 cDNA quantity of tilapia samples after purification

No.	Species	Conc. ng/µl	OD 260/280	OD 260/230
1	*O. niloticus* 1	56.5	1.83	2.26
2	*O. niloticus* 2	19.3	1.8	1.59
3	*O. mossambicus* 1	17.5	1.54	1.02
4	*O. mossambicus* 2	24.5	1.73	0.77
5	*O. karongae* 1	25.0	1.83	1.8
6	*O. karongae* 2	13.1	1.8	1.59
7	*O. aureus* 1	17.6	1.71	0.48
8	*O. aureus* 2	20.9	1.54	0.76
9	*T. zillii* 1	39.6	2.12	1.3
10	*T. zillii* 2	11.9	2.17	1.89
11	*O. niloticus* 3	34.43	1.90	1.64
12	*O. niloticus* 4	51.13	1.93	1.50
13	*O. niloticus* 5	19.22	2.13	1.38
14	*O. niloticus* 6	12.03	1.73	1.25

Chapter IV

Appendix IV.1. Primer used for PCR and sequencing DNA barcode – COI gene

eurofins Genomics

Oligonucleotide Synthesis Report

Page 1/1

Ms. Jane Lewis
University of Stirling

Order ID: 3265888
Customer ID: 67260
Your Order ID (PO#): 6313526

Order Date: 17.04.2014
Lab No.: 4146
No. of Oligos: 4/4

Eurofins Genomics
Anzingerstraße 7a
D- 85560 Ebersberg

No.	Oligo Name	Sequence (5' -> 3')	Yield [OD]	Yield [µg]	Yield [nmol]	Concentration [pmol/µl]	Vol. for 100pmol/µl	Tm [°C]	MW [g/mol]	GC-Content	Synthesis Scale	Purification	Modification	Barcode IDO	QC Report
1	FishF1	TCAACCAACCACAAAGACATTGGCAC (26)	7.1	187	23.7	-	237	63.2	7886	46.2 %	0.01 µmol	HPSF	-	017830056	-
2	FishF2	TCGACTAATCATAAAGATATCGGCAC (26)	8.0	211	26.6	-	266	60.1	7947	38.5 %	0.01 µmol	HPSF	-	017830057	-
3	FishR1	TAGACTTCTGGGTGGCCAAAGAATCA (26)	6.1	164	20.5	-	205	63.2	8019	46.2 %	0.01 µmol	HPSF	-	017830058	-
4	FishR2	ACTTCAGGGTGACCGAAGAATCAGAA (26)	7.4	192	23.9	-	239	63.2	8037	46.2 %	0.01 µmol	HPSF	-	017830059	-

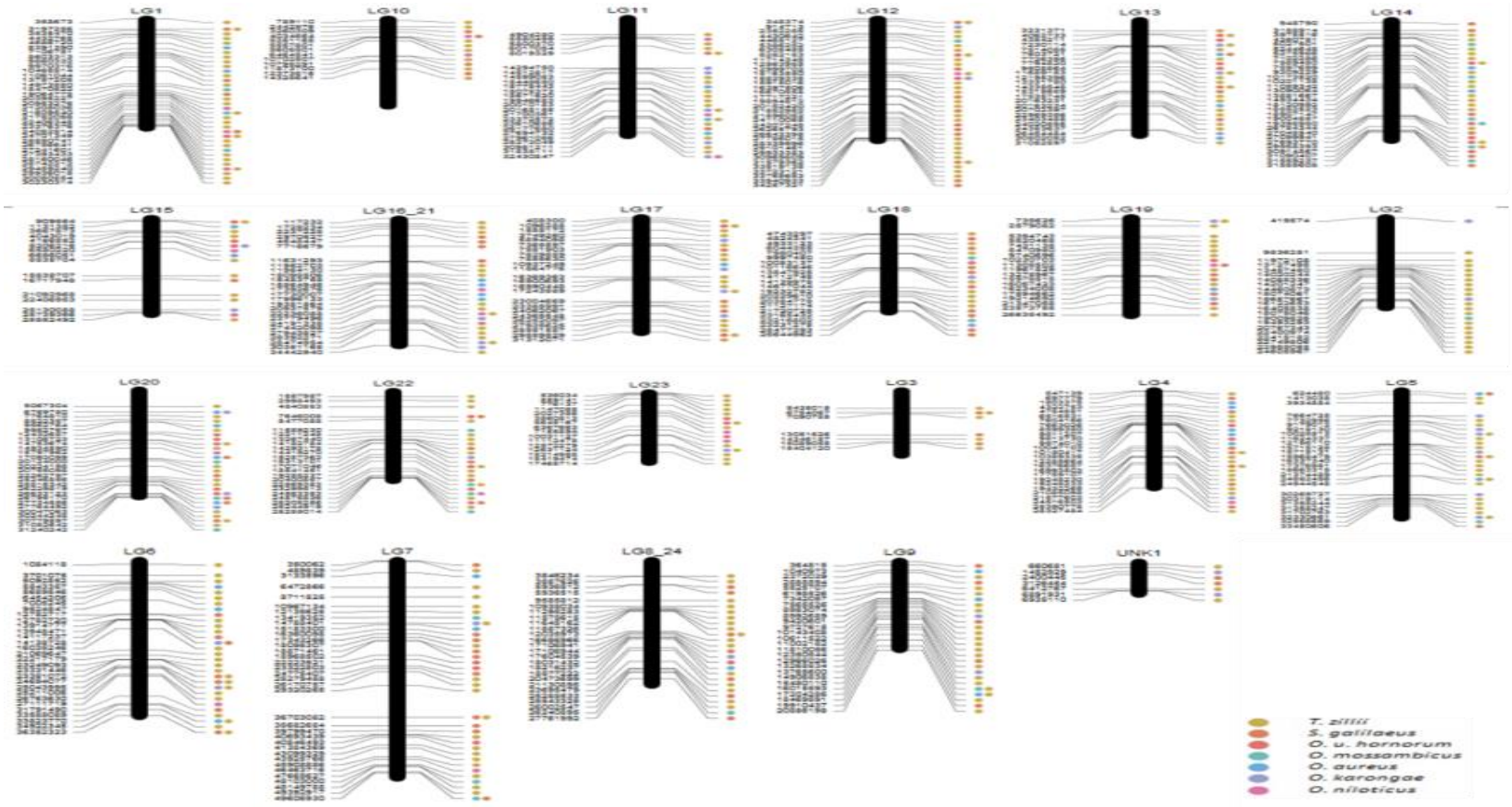
Appendix IV.2. Samples origin and barcode. Details each sample used: sample ID, species, barcode used, number of raw reads and number of RAD-tags.

Taxon_ID	Organism_name	Barcode1	Barcode2	Library	Run	Read number	Unique stacks	Polymorphic loci	
								DBA	RBA
8128	*O. niloticus*	CGATA	-	Lib 1	HiSeq (2x100nt)	8211632	65214	8245	5232
8128	*O. niloticus*	AACCC	-	Lib 3	HiSeq (2x100nt)	4443087	58484	7226	4827
8128	*O. niloticus*	CGTAT	-	Lib 2	HiSeq (2x100nt)	1461646	22645	1737	2216
8128	*O. niloticus*	GAGAT	-	Lib 4	HiSeq (2x100nt)	4137896	59970	6728	4818
8127	*O. mossambicus*	CTAGG	-	Lib 1	HiSeq (2x100nt)	6836092	63221	6150	2884
8127	*O. mossambicus*	CTGAA	-	Lib 1	HiSeq (2x100nt)	5569757	61109	5693	2810
8127	*O. mossambicus*	CTCTT	-	Lib 2	HiSeq (2x100nt)	1490811	24274	1528	1090
8127	*O. mossambicus*	CCTTG	-	Lib 2	HiSeq (2x100nt)	1373507	21335	1394	1158
8127	*O. mossambicus*	CGGCG	-	Lib 3	HiSeq (2x100nt)	4732619	59381	4208	2277
8127	*O. mossambicus*	GATCG	-	Lib 3	HiSeq (2x100nt)	1035656	8946	814	746
8127	*O. mossambicus*	GCCGG	-	Lib 4	HiSeq (2x100nt)	836851	3674	506	506
8127	*O. mossambicus*	GTGTG	-	Lib 4	HiSeq (2x100nt)	3273859	55594	5219	3017
47969	*O. aureus*	CGCGC	-	Lib 1	HiSeq (2x100nt)	1094747	12035	923	550
47969	*O. aureus*	GCTAA	-	Lib 1	HiSeq (2x100nt)	1194172	16138	1049	676
47969	*O. aureus*	GACTA	-	Lib 2	HiSeq (2x100nt)	1719528	31278	1614	925
47969	*O. aureus*	GCGCC	-	Lib 2	HiSeq (2x100nt)	5689694	61409	3573	1313
47969	*O. aureus*	GCATT	-	Lib 3	HiSeq (2x100nt)	6544527	62903	3745	1489
47969	*O. aureus*	GGTTC	-	Lib 3	HiSeq (2x100nt)	1069855	11439	887	469
47969	*O. aureus*	GTACA	-	Lib 4	HiSeq (2x100nt)	3987991	59446	4015	1711
47969	*O. aureus*	TCTCT	-	Lib 4	HiSeq (2x100nt)	645366	1555	339	243
167928	*O. karongae*	TAATG	-	Lib 1	HiSeq (2x100nt)	4955196	56862	4537	2311

167928	*O. karongae*	GTCAC	-	Lib 1	HiSeq (2x100nt)	1438581	22717	1504	1303
167928	*O. karongae*	GTTGT	-	Lib 2	HiSeq (2x100nt)	1944326	33285	2513	1902
167928	*O. karongae*	TACGT	-	Lib 2	HiSeq (2x100nt)	2604568	43653	3328	1836
167928	*O. karongae*	TATAC	-	Lib 3	HiSeq (2x100nt)	968772	7034	665	635
167928	*O. mossambicus*	TGCAA	-	Lib 4	HiSeq (2x100nt)	1435805	20643	1309	1539
167928	*O. karongae*	TTCCG	-	Lib 3	HiSeq (2x100nt)	1431661	21075	1447	1387
8130	*Tilapia zillii*	TAGCA	-	Lib 1	HiSeq (2x100nt)	4248041	57134	5752	934
8130	*Tilapia zillii*	TCGAG	-	Lib 2	HiSeq (2x100nt)	1421663	22084	1340	476
8130	*Tilapia zillii*	TCCTC	-	Lib 1	HiSeq (2x100nt)	5211705	57754	4792	821
8130	*Tilapia zillii*	AGCTG	-	Lib 3	HiSeq (2x100nt)	976000	4971	638	271
8130	*Tilapia zillii*	AGTCA	-	Lib 3	HiSeq (2x100nt)	1038157	8885	776	339
8130	*Tilapia zillii*	GAAGC	-	Lib 4	HiSeq (2x100nt)	6964887	62057	5062	817
40193	*O. urolepis*	TGTGG	-	Lib 1	HiSeq (2x100nt)	1723680	32097	1619	1045
40193	*O. urolepis*	ATTAG	-	Lib 1	HiSeq (2x100nt)	872242	6017	609	412
40193	*O. urolepis*	TGACC	-	Lib 2	HiSeq (2x100nt)	17171	4	2	N/A
40193	*O. urolepis*	TTTTA	-	Lib 2	HiSeq (2x100nt)	1192469	14942	1019	647
40193	*O. urolepis*	CTTCC	-	Lib 3	HiSeq (2x100nt)	1574115	28177	1533	993
40193	*O. urolepis*	CATGA	-	Lib 3	HiSeq (2x100nt)	912010	6275	592	461
40193	*O. urolepis*	GGAAG	-	Lib 4	HiSeq (2x100nt)	4436526	59627	3312	1337
40193	*O. urolepis*	GGGGA	-	Lib 4	HiSeq (2x100nt)	3519257	57295	3264	1422
8128	*O. niloticus*	AGAGT	-	Lib 1	HiSeq (2x100nt)	17896	7	3	N/A
8128	*O. niloticus*	TCAGA	-	Lib 2	HiSeq (2x100nt)	7953206	63596	7677	5024
8128	*O. niloticus*	CACAG	-	Lib 4	HiSeq (2x100nt)	786522	2768	402	598
8128	*O. niloticus*	ACTGC	-	Lib 4	HiSeq (2x100nt)	3185244	51212	6706	5140
8144	*S. galilaeus*	TCAGA	TAGCA & AGCTGA	Lib 5	MiSeq (2x150nt)	3403786	29409	1449	1740
8144	*S. galilaeus*	CGTATCA	TAGCA &	Lib 5	MiSeq	4381538	40209	1847	2879

			AGCTGA		(2x150nt)				
8144	*S. galilaeus*	AGAGT	TAGCA & AGCTGA	Lib 5	MiSeq (2x150nt)	3385800	28324	1398	1770
8144	*S. galilaeus*	GCTAACA	TAGCA & AGCTGA	Lib 5	MiSeq (2x150nt)	5849836	52715	1464	3352
8144	*S.galilaeus*	GATCG	TAGCA & AGCTGA	Lib 5	MiSeq (2x150nt)	2868718	20276	1168	1317
28827	*P. pulcher*	CTCTTCA	TAGCA & AGCTGA	Lib 5	MiSeq (2x150nt)	4964912	42343	4289	201
28827	*P. pulcher*	ATCGA	TAGCA & AGCTGA	Lib 5	MiSeq (2x150nt)	418494	79	32	N/A

Appendix IV.3. The complete mapping of species – specific SNP markers derived from standard RADseq



Chapter V

Appendix V.1 Samples origin and barcode. Details each sample used: sample ID, species, barcode used, number of raw reads and number of RAD-tags.

No	Sample ID	Taxon ID	Organism name	Barcode1	Barcode2	Library	Run	Read number	Unique Stacks
1	Md-08	8128	*O. niloticus*	TCAGA	CGATC	MiSeq	2x150nt	1046082	4631
2	Md-09	8128	*O. niloticus*	GATCG	CGATC	MISeq	2x150nt	1246958	4709
3	Md-10	8128	*O. niloticus*	CATGA	CGATC	MiSeq	2x150nt	1377062	4787
4	Md-11	8128	*O. niloticus*	ATCGA	CGATC	MiSeq	2x150nt	1109838	4593
5	Md-12	8127	*O. mossambicus*	TCGAG	CGATC	MiSeq	2x150nt	1612312	3528
6	Md-14	8127	*O. mossambicus*	GTCAC	CGATC	MiSeq	2x150nt	990964	3291
7	Md-15	8127	*O. mossambicus*	GCATT	CGATC	MISeq	2x150nt	1349408	3417
8	Md-16	8127	*O. mossambicus*	CGATA	CGATC	MiSeq	2x150nt	1017114	3244
9	Md-17	8127	*O. mossambicus*	TGCAAC	CGATC	MiSeq	2x150nt	848560	3249
10	Md-21	47969	*O. aureus*	CGTATC	CGATC	MiSeq	2x150nt	1521692	3671
11	Md-22	47969	*O. aureus*	CACAGA	CGATC	MiSeq	2x150nt	1310644	3645
12	Md-23	47969	*O. aureus*	ACTGCA	CGATC	MISeq	2x150nt	972380	3467
13	Md-26	47969	*O. aureus*	TCTCTCA	CGATC	MiSeq	2x150nt	2402842	3721
14	Md-29	167928	*O. karongae*	GTACAC	CGATC	MiSeq	2x150nt	1114616	3434
15	Md-30	167928	*O. karongae*	CTCTTCA	CGATC	MiSeq	2x150nt	875448	3348
16	Md-31	167928	*O. karongae*	CTAGGA	CGATC	MiSeq	2x150nt	1952048	3503
17	Md-32	167928	*O. karongae*	ACGTA	CGATC	MISeq	2x150nt	1077908	3415
18	Md-33	167928	*O. karongae*	AGAGT	CGATC	MiSeq	2x150nt	801202	3281
19	Md-35	8130	*T. zillii*	ATGCT	CGATC	MiSeq	2x150nt	856184	824
20	Md-37	8130	*T. zillii*	GACTA	CGATC	MiSeq	2x150nt	1509366	857
21	Md-38	8130	*T. zillii*	CAGTCA	CGATC	MiSeq	2x150nt	1012510	830
22	Md-39	8130	*T. zillii*	GCTAAC	CGATC	MISeq	2x150nt	1065834	841
23	Md-40	8130	*T. zillii*	ACACGA	CGATC	MiSeq	2x150nt	1453648	859
24	Md-66	47969	*O. aureus*	AGGACA	CGATC	MiSeq	2x150nt	836666	3440
25	M69	40193	*O. urolepis*	TCAGA	CATCTGT	MiSeq	2x150nt	570664	2951
26	M70	40193	*O. urolepis*	GATCG	CATCTGT	MiSeq	2x150nt	718066	3108
27	M75	40193	*O. urolepis*	CATGA	CATCTGT	MISeq	2x150nt	737788	3136
28	M77	40193	*O. urolepis*	ATCGA	CATCTGT	MiSeq	2x150nt	802414	3169
29	MM-78	40193	*O. urolepis*	TCGAG	CATCTGT	MiSeq	2x150nt	638506	3061
30	MM-294	8144	*S. galilaeus*	GTCAC	CATCTGT	MiSeq	2x150nt	673834	2226
31	MM-298	8144	*S. galilaeus*	GCATT	CATCTGT	MiSeq	2x150nt	529064	2118
32	MM-301b	8144	*S. galilaeus*	CGATA	CATCTGT	MISeq	2x150nt	915488	2387

33	MM-302	8144	*S. galilaeus*	TGCAAC CGTATC	CATCTGT	MiSeq	2x150nt	981768	2321
34	Md-309	8144	*S. galilaeus*	CACAGA	CATCTGT	MiSeq	2x150nt	660892	2220
35	Md-498	8128	*O. niloticus*	ACTGCA	CATCTGT	MiSeq	2x150nt	1172038	4607
36	Md-499	8128	*O. niloticus*		CATCTGT	MiSeq	2x150nt	1338556	4684
37	Md-91	158894	O. andersoni	CATGA	CTGGT	MISeq	2x150nt	232758	1366
38	Md-92	158894	O. andersoni	ATCGA	CTGGT	MiSeq	2x150nt	563288	2459
39	Md-93	158894	O. andersoni	TCGAG	CTGGT	MiSeq	2x150nt	675174	2676
40	Md-94	158894	O. andersoni	GTCAC	CTGGT	MiSeq	2x150nt	499538	2624
41	Md-98	158894	O. andersoni	GCATT	CTGGT	MiSeq	2x150nt	656560	2770
42	Md-100	158894	O. andersoni	CGATA	CTGGT	MISeq	2x150nt	77544	383
43	Md-113	158766	O. macrochir	TGCAAC	CTGGT	MiSeq	2x150nt	251368	1738
44	Md-115	158766	O. macrochir	CGTATC	CTGGT	MiSeq	2x150nt	931152	3149
45	Md-116	158766	O. macrochir	CACAGA	CTGGT	MiSeq	2x150nt	231874	1611
46	Md-117	158766	O. macrochir	ACTGCA	CTGGT	MiSeq	2x150nt	303394	1899
47	Md-134	8128	O. n. niloticus	CACAGA	GCATA	MISeq	2x150nt	1176840	4485
48	Md-136	8128	O. n. niloticus	ACTGCA	GCATA	MiSeq	2x150nt	777810	4323
49	Md-137	8128	O. n. niloticus	TCTCTCA	GCATA	MiSeq	2x150nt	316210	3590
50	Md-138	8128	O. n. niloticus	GTACAC	GCATA	MiSeq	2x150nt	497834	4066
51	Md-140	8128	O. n. niloticus	TCTCTCA	GTCAAGT	MiSeq	2x150nt	2031356	4502
52	Md-142	8128	O. n. niloticus	CTAGGA	GCATA	MISeq	2x150nt	449728	3969
53	Md-143	8128	O. n. niloticus	ACGTA	GCATA	MiSeq	2x150nt	543328	4162
54	Md-144	8128	O. n. niloticus	AGAGT	GCATA	MiSeq	2x150nt	398318	3908
55	Md-145	8128	O. n. niloticus	ATGCT	GCATA	MiSeq	2x150nt	413474	3948
56	Md-149	8128	O. n. niloticus	GACTA	GCATA	MiSeq	2x150nt	339772	3764
57	Md-150	8128	O. n. niloticus	CAGTCA	GCATA	MISeq	2x150nt	544544	4182
58	Md-152	8128	O. n. niloticus	GCTAAC	GCATA	MiSeq	2x150nt	507768	4052
59	Md-154	8128	O. n. niloticus	ACACGA	GCATA	MiSeq	2x150nt	543338	4086
60	Md-156	8128	O. n. niloticus	AGGACA	GCATA	MiSeq	2x150nt	432792	3952
61	Md-157	8128	O. n. niloticus	TCAGA	GAGATGT	MiSeq	2x150nt	225832	2918
62	Md-158	8128	O. n. niloticus	GATCG	GAGATGT	MISeq	2x150nt	245488	3097
63	Md-160	8128	O. n. niloticus	GTACAC	GTCAAGT	MiSeq	2x150nt	2114242	4213
64	Md-162	8128	O. n. niloticus	ATCGA	GAGATGT	MiSeq	2x150nt	312468	3428
65	Md-163	8128	O. n. niloticus	TCGAG	GAGATGT	MiSeq	2x150nt	309288	3398
66	Md-164	8128	O. n. niloticus	GTCAC	GAGATGT	MiSeq	2x150nt	200390	2579
67	Md-165	8128	O. n. niloticus	GCATT	GAGATGT	MISeq	2x150nt	229904	2448
68	Md-170	8128	O. n. niloticus	CGATA	GAGATGT	MiSeq	2x150nt	280092	3248
69	Md-172	8128	O. n. niloticus	TGCAAC	GAGATGT	MiSeq	2x150nt	525148	3828
70	Md-173	8128	O. n. niloticus	CGTATC	GAGATGT	MiSeq	2x150nt	291242	3357
71	Md-174	8128	O. n. cancellatus	CACAGA	GAGATGT	MiSeq	2x150nt	234006	2931

72	Md-175	8128	O. n. cancellatus	ACTGCA	GAGATGT	MISeq	2x150nt	287204	3146
73	Md-176	8128	O. n. cancellatus	TCTCTCA	GAGATGT	MiSeq	2x150nt	243358	3017
74	Md-177	8128	O. n. cancellatus	GTACAC	GAGATGT	MiSeq	2x150nt	429310	3495
75	Md-178	8128	O. n. cancellatus	CTCTTCA	GAGATGT	MiSeq	2x150nt	318984	3257
76	Md-179	8128	O. n. cancellatus	CTAGGA	GAGATGT	MiSeq	2x150nt	287244	3161
77	Md-180	8128	O. n. cancellatus	ACGTA	GAGATGT	MISeq	2x150nt	307572	3263
78	Md-182	8128	O. n. cancellatus	AGAGT	GAGATGT	MiSeq	2x150nt	266030	3140
79	Md-184	8128	O. n. cancellatus	ATGCT	GAGATGT	MiSeq	2x150nt	394908	3474
80	Md-186	8128	O. n. cancellatus	GACTA	GAGATGT	MiSeq	2x150nt	428914	3523
81	Md-188	8128	O. n. cancellatus	CAGTCA	GAGATGT	MiSeq	2x150nt	378424	3459
82	Md-189	8128	O. n. cancellatus	GCTAAC	GAGATGT	MISeq	2x150nt	854152	3823
83	Md-190	8128	O. n. cancellatus	ACACGA	GAGATGT	MiSeq	2x150nt	439406	3510
84	Md-191	8128	O. n. cancellatus	AGGACA	GAGATGT	MiSeq	2x150nt	552660	3669
85	Md-194	8128	O. n. cancellatus	TCAGA	CGATC	MiSeq	2x150nt	475644	3632
86	Md-195	8128	O. n. cancellatus	GATCG	CGATC	MiSeq	2x150nt	438622	3590
87	Md-196	8128	O. n. cancellatus	CATGA	CGATC	MISeq	2x150nt	442790	3589
88	Md-197	8128	O. n. cancellatus	ATCGA	CGATC	MiSeq	2x150nt	353382	3457
89	Md-198	8128	O. n. cancellatus	TCGAG	CGATC	MiSeq	2x150nt	292802	3237
90	Md-200	8128	O. n. cancellatus	GTCAC	CGATC	MiSeq	2x150nt	245206	3003
91	Md-201	8128	O. n. cancellatus	GCATT	CGATC	MiSeq	2x150nt	330730	3324
92	Md-202	8128	O. n. cancellatus	CGATA	CGATC	MISeq	2x150nt	467128	3648
93	Md-203	8128	O. n. cancellatus	TGCAAC	CGATC	MiSeq	2x150nt	328960	3393
94	Md-204	8128	O. n. cancellatus	CGTATC	CGATC	MiSeq	2x150nt	341820	3416
95	Md-210	8128	O. n. cancellatus	CACAGA	CGATC	MiSeq	2x150nt	329072	3378
96	Md-215	8128	O. n. cancellatus	CTCTTCA	GTCAAGT	MiSeq	2x150nt	1477568	3816
97	Md-221	8128	O. n. cancellatus	CTAGGA	GTCAAGT	MISeq	2x150nt	2620938	4093
98	Md-222	8128	O. n. cancellatus	ACGTA	GTCAAGT	MiSeq	2x150nt	1375230	3802
99	Md-223	8128	O. n. cancellatus	AGAGT	GTCAAGT	MiSeq	2x150nt	1272604	3764
100	Md-224	8128	O. n. cancellatus	ATGCT	GTCAAGT	MiSeq	2x150nt	877824	3498
101	Md-225	8128	O. n. cancellatus	GACTA	GTCAAGT	MiSeq	2x150nt	884208	3558
102	Md-226	8128	O. n. cancellatus	CAGTCA	GTCAAGT	MISeq	2x150nt	555372	3138
103	Md-228	8128	O. n. cancellatus	GCTAAC	GTCAAGT	MiSeq	2x150nt	783578	3438
104	Md-230	8128	S. melanotheron	GACTA	CGATC	MiSeq	2x150nt	233014	1502
105	Md-231	8128	S. melanotheron	CAGTCA	CGATC	MiSeq	2x150nt	240600	1516
106	Md-234	8128	S. melanotheron	ACACGA	GTCAAGT	MiSeq	2x150nt	3203984	2384
107	Md-235	8128	S. melanotheron	ACACGA	CGATC	MISeq	2x150nt	311392	1632
108	Md-237a	8130	T. zillii Ghana	AGGACA	GTCAAGT	MiSeq	2x150nt	1471698	752
109	Md-242	8130	T. zillii	TCAGA	CATCTGT	MiSeq	2x150nt	237680	406
110	Md-246a	8130	T. zillii	GATCG	CATCTGT	MiSeq	2x150nt	250400	460

111	Md-246b	8130	T. zillii	CATGA	CATCTGT	MiSeq	2x150nt	274682	470
112	MM-246c	8130	T. zillii	TCTCTCA	ATACGGT	MISeq	2x150nt	671644	698
113	MM-247	47969	O. aureus AFI	GTACAC	ATACGGT	MiSeq	2x150nt	1192988	3330
114	Md-249	47969	O. aureus	CTCTTCA	ATACGGT	MiSeq	2x150nt	1446800	3361
115	Md-256	47969	O. aureus	CTAGGA	ATACGGT	MiSeq	2x150nt	771570	3180
116	Md-257	47969	O. aureus	ACGTA	ATACGGT	MiSeq	2x150nt	1011382	3227
117	Md-259	47969	O. aureus	AGAGT	ATACGGT	MISeq	2x150nt	1459244	3379
118	Md-260	47969	O. aureus	ATGCT	ATACGGT	MiSeq	2x150nt	1431672	3400
119	Md-262	47969	O. aureus	GACTA	ATACGGT	MiSeq	2x150nt	1336868	3380
120	Md-263	47969	O. aureus	CAGTCA	ATACGGT	MiSeq	2x150nt	1718554	3395
121	Md-265	47969	O. aureus	GCTAAC	ATACGGT	MiSeq	2x150nt	1419794	3335
122	Md-266	47969	O. aureus	ACACGA	ATACGGT	MISeq	2x150nt	1194498	3340
123	Md-269	8127	O. mossambicus NSA	AGGACA	ATACGGT	MiSeq	2x150nt	829562	2966
124	Md-270	8127	O. mossambicus	TCAGA	GAAGC	MiSeq	2x150nt	540828	2879
125	Md-271	8127	O. mossambicus	GATCG	GAAGC	MiSeq	2x150nt	627674	2934
126	Md-274	8127	O. mossambicus	AGAGT	CATCTGT	MiSeq	2x150nt	317922	2501
127	Md-276	8127	O. mossambicus	CATGA	GAAGC	MISeq	2x150nt	772408	3009
128	Md-280	8127	O. mossambicus	GACTA	CATCTGT	MiSeq	2x150nt	266204	2402
129	Md-281	8127	O. mossambicus	CAGTCA	CATCTGT	MiSeq	2x150nt	356380	2668
130	Md-283	8127	O. mossambicus	GCTAAC	CATCTGT	MiSeq	2x150nt	460402	2684
131	Md-284	8127	O. mossambicus	ACACGA	CATCTGT	MiSeq	2x150nt	220052	1993
132	Md-288	8127	O. mossambicus	AGGACA	CATCTGT	MISeq	2x150nt	371932	2642

Appendix V.2. The list of species-specific SNP markers across 10 tilapia species derived from double digest RADseq.

Marker ID	Genotyping										Chr	Position	DNA Strand	Species
	Onil	Omos	Oaur	Okar	Tzil	Ohor	Sgal	Oan	Omac	Smel				
81_B	{GG}	{GG}	{GG}	{GG}	{AA}	{GG}	{GG}	{GG}	{GG}	{GG}	LG10	1494812	+	*T. zillii*
400_A	{CC}	{CC}	{CC}	{CC}	{TT}	{CC}	{CC}	{CC}	{CC}	{CC}	LG11	16907454	+	*T. zillii*
765_B	{CC}	{CC}	{CC}	{CC}	{CC}	{CC}	{GG}	{CC}	{CC}	{CC}	LG11	33107708	+	*S. galilaeus*
1109_A	{GG}	{GG}	{GG}	{GG}	{AA}	{GG}	{GG}	{GG}	{GG}	{GG}	LG12	22036688	+	*T. zillii*
1178_B	{CC}	{CC}	{CC}	{CC}	{GG}	{CC}	{CC}	{CC}	{CC}	{CC}	LG12	26523972	-	*T. zillii*
1669_A	{GG}	{GG}	{GG}	{GG}	{AA}	{GG}	{GG}	{GG}	{GG}	{GG}	LG13	18502360	+	*T. zillii*
1669_B	{AA}	{AA}	{AA}	{AA}	{GG}	{AA}	{AA}	{AA}	{AA}	{AA}	LG13	18502360	+	*T. zillii*
1670_A	{TT}	{TT}	{TT}	{TT}	{CC}	{TT}	{TT}	{TT}	{TT}	{TT}	LG13	18502495	-	*T. zillii*
1670_B	{CC}	{CC}	{CC}	{CC}	{TT}	{CC}	{CC}	{CC}	{CC}	{CC}	LG13	18502495	-	*T. zillii*
1756_B	{TT}	{TT}	{TT}	{TT}	{TT}	{TT}	{TT}	{TT}	{TT}	{CC}	LG13	23602908	+	*S. melanotheron*
1758_A	{AA}	{AA}	{AA}	{AA}	{AA}	{AA}	{AA}	{AA}	{AA}	{GG}	LG13	23603139	-	*S. melanotheron*
2068_A	{AA}	{AA}	{AA}	{AA}	{GG}	{AA}	{AA}	{AA}	{AA}	{AA}	LG14	1195460	+	*T. zillii*
2275_B	{AA}	{AA}	{AA}	{AA}	{CC}	{AA}	{AA}	{AA}	{AA}	{AA}	LG14	25734330	-	*T. zillii*
3053_A	{TT}	{TT}	{TT}	{TT}	{AA}	{TT}	{TT}	{TT}	{TT}	{TT}	LG15	8731974	+	*T. zillii*
3067_B	{AA}	{AA}	{AA}	{AA}	{GG}	{AA}	{AA}	{AA}	{AA}	{AA}	LG16_21	10110787	+	*T. zillii*
3602_A	{GG}	{GG}	{GG}	{GG}	{AA}	{GG}	{GG}	{GG}	{GG}	{GG}	LG16_21	6727366	+	*T. zillii*
3887_A	{AA}	{AA}	{AA}	{AA}	{GG}	{AA}	{AA}	{AA}	{AA}	{AA}	LG17	21927644	+	*T. zillii*
4246_A	{AA}	{AA}	{AA}	{AA}	{AA}	{AA}	{AA}	{AA}	{AA}	{GG}	LG18	10116279	+	*S. melanotheron*
4739_A	{AA}	{AA}	{AA}	{AA}	{GG}	{AA}	{AA}	{AA}	{AA}	{AA}	LG19	1328420	+	*T. zillii*
5164_A	{CC}	{CC}	{CC}	{TT}	{CC}	{CC}	{CC}	{CC}	{CC}	{CC}	LG19	7456400	+	*O. karongae*
5164_B	{GG}	{GG}	{GG}	{GG}	{GG}	{GG}	{GG}	{GG}	{CC}	{GG}	LG19	7456400	+	*O. macrochir*
5190_A	{TT}	{TT}	{TT}	{TT}	{TT}	{TT}	{TT}	{TT}	{TT}	{CC}	LG19	9517617	+	*S. melanotheron*
5373_A	{GG}	{GG}	{GG}	{GG}	{AA}	{GG}	{GG}	{GG}	{GG}	{GG}	LG1	23110020	-	*T. zillii*

6473_A	{GG}	{GG}	{GG}	{GG}	{AA}	{GG}	{GG}	{GG}	{GG}	{GG}	LG22	24977115	-	*T. zillii*
6785_A	{CC}	{CC}	{CC}	{CC}	{GG}	{CC}	{CC}	{CC}	{CC}	{CC}	LG23	15047646	-	*T. zillii*
6785_B	{AA}	{AA}	{AA}	{AA}	{CC}	{AA}	{AA}	{AA}	{AA}	{AA}	LG23	15047646	-	*T. zillii*
6857_B	{CC}	{CC}	{CC}	{CC}	{AA}	{CC}	{CC}	{CC}	{CC}	{CC}	LG23	19006771	+	*T. zillii*
6972_A	{CC}	{CC}	{CC}	{CC}	{CC}	{CC}	{CC}	{CC}	{CC}	{TT}	LG23	6094625	+	*S. melanotheron*
8029_A	{AA}	{AA}	{GG}	{AA}	{AA}	{AA}	{AA}	{AA}	{AA}	{AA}	LG4	2024866	-	*O. aureus*
8512_B	{TT}	{TT}	{TT}	{TT}	{CC}	{TT}	{TT}	{TT}	{TT}	{TT}	LG5	19043569	-	*T. zillii*
8609_A	{AA}	{AA}	{AA}	{AA}	{GG}	{AA}	{AA}	{AA}	{AA}	{AA}	LG5	24554351	+	*T. zillii*
8609_B	{CC}	{CC}	{CC}	{CC}	{CC}	{CC}	{TT}	{CC}	{CC}	{CC}	LG5	24554351	+	*S. galilaeus*
8875_A	{GG}	{GG}	{GG}	{GG}	{AA}	{GG}	{GG}	{GG}	{GG}	{GG}	LG5	7472597	+	*T. zillii*
8875_B	{AA}	{AA}	{AA}	{AA}	{GG}	{AA}	{AA}	{AA}	{AA}	{AA}	LG5	7472597	+	*T. zillii*
9441_A	{CC}	{CC}	{CC}	{CC}	{TT}	{CC}	{CC}	{CC}	{CC}	{CC}	LG6	4326506	-	*T. zillii*
9441_B	{TT}	{TT}	{TT}	{TT}	{AA}	{TT}	{TT}	{TT}	{TT}	{TT}	LG6	4326506	-	*T. zillii*
10728_B	{GG}	{GG}	{GG}	{AA}	{GG}	{GG}	{GG}	{GG}	{GG}	{GG}	LG8_24	25515874	+	*O. karongae*
10758_B	{TT}	{TT}	{TT}	{TT}	{TT}	{TT}	{CC}	{TT}	{TT}	{TT}	LG8_24	26487018	+	*S. galilaeus*

Species-specific markers from a subset of four commercial tilapia species (*O. niloticus*, *O. mossambicus*, *O. aureus* and *O. u. hornorum*)

a. *O. niloticus*

Marker_ID	Genotyping		Chr	Position	DNA strand
	onil	other			
3057_B	{AA}	{TT}	LG15	909826	-
1276_B	{AA}	{GG}	LG12	3072836	-
2082_A	{AA}	{GG}	LG14	13140011	+
2082_B	{CC}	{AA}	LG14	13140011	+
2675_A	{GG}	{TT}	LG15	12494644	-
3531_A	{AA}	{GG}	LG16_21	4401532	+
5782_B	{AA}	{TT}	LG20	16532929	-

b. *O. mossambicus*

Marker_ID	Genotyping		Chr	Position	DNA strand
	other	omos			
1504_B	{CC}	{TT}	LG12	8753442	-
1125_A	{TT}	{AA}	LG12	22867680	+
2657_B	{GG}	{AA}	LG15	11282183	+
4742_A	{CC}	{TT}	LG19	13369097	+
5412_A	{GG}	{TT}	LG1	2700436	-
5760_A	{CC}	{GG}	LG20	15924279	+
5761_B	{GG}	{CC}	LG20	15924451	-
10120_A	{GG}	{AA}	LG7	40145502	-
10818_B	{GG}	{AA}	LG8_24	3390946	+
10819_A	{CC}	{TT}	LG8_24	3391091	-

c. *O. aureus*

Marker_ID	Genotyping		Chr	Position	DNA strand
	other	oaur			
8029_A	{AA}	{GG}	LG4	2024866	-
966_A	{CC}	{TT}	LG12	13862286	+
3236_B	{CC}	{TT}	LG16_21	18919216	+
3001_A	{GG}	{AA}	LG15	6075396	+
1736_A	{CC}	{TT}	LG13	21872800	-
1992_A	{CC}	{TT}	LG13	8028370	+
2890_B	{CC}	{TT}	LG15	24160607	-
2899_A	{CC}	{GG}	LG15	24719101	+
3531_B	{CC}	{TT}	LG16_21	4401532	+
3873_A	{TT}	{CC}	LG17	21279151	-

5416_A	{CC}	{TT}	LG1	27201771	+
5424_A	{CC}	{TT}	LG1	27706614	-
9418_A	{TT}	{CC}	LG6	36351716	+

d. *O. u. hornorum*

Marker_ID	Genotyping		Chr	Position	DNA strand
	other	ohor			
2680_B	{CC}	{TT}	LG15	12784552	+
8603_A	{AA}	{GG}	LG5	24413535	+
10199_B	{GG}	{AA}	LG7	45218647	+
2519_A	{TT}	{CC}	LG14	34149812	+
4270_A	{GG}	{AA}	LG18	1123631	-
10793_A	{AA}	{GG}	LG8_24	2824094	-
10951_A	{AA}	{TT}	LG8_24	8700535	-

Species-specific markers in sub-species level between *O. n. niloticus* and *O. n. cancellatus*

Marker_ID	Genotyping		Chr	Position	DNA strand
	niloticus	*cancellatus*			
5516_A	{AA}	{GG}	LG1	29403312	-
6674_B	{CC}	{TT}	LG23	10814820	+
7321_B	{GG}	{AA}	LG2	24612566	+

Species-specific markers for species pairs

a. Between *O. niloticus* - *O. mossambicus*

Marker ID	Genotyping		Chr	Position	DNA strand
	onil	omos			
765_A	{GG}	{AA}	LG11	33107708	+
1504_B	{CC}	{TT}	LG12	8753442	-
1756_A	{CC}	{TT}	LG13	23602908	+
4560_A	{GG}	{CC}	LG18	3512266	+
6674_A	{AA}	{GG}	LG23	10814820	+
8944_A	{TT}	{CC}	LG6	1059739	+
2007_B	{GG}	{AA}	LG13	9212591	-
7184_B	{CC}	{TT}	LG2	1791053	+
1594_A	{AA}	{GG}	LG13	13361049	-
3057_B	{AA}	{TT}	LG15	909826	-
3233_A	{TT}	{CC}	LG16_21	1885870	+
8001_B	{GG}	{AA}	LG4	17759107	+
8436_A	{GG}	{TT}	LG5	14437062	-
8436_B	{AA}	{GG}	LG5	14437062	-
11084_B	{GG}	{AA}	LG9	14493472	-
57_B	{TT}	{GG}	LG10	13769218	+
294_A	{TT}	{CC}	LG11	10698975	+
435_B	{GG}	{AA}	LG11	18262482	+
465_A	{CC}	{GG}	LG11	20373159	-
633_A	{TT}	{CC}	LG11	28450485	+
801_A	{TT}	{CC}	LG11	4784136	-
1125_A	{TT}	{AA}	LG12	22867680	+
1276_B	{AA}	{GG}	LG12	3072836	-
1288_A	{AA}	{GG}	LG12	3123541	+
1678_A	{CC}	{TT}	LG13	19237959	+
1777_B	{GG}	{CC}	LG13	24956987	+
2015_A	{TT}	{CC}	LG13	9915235	+
2082_A	{AA}	{GG}	LG14	13140011	+
2082_B	{CC}	{AA}	LG14	13140011	+
2451_A	{AA}	{GG}	LG14	31639274	-
2657_B	{GG}	{AA}	LG15	11282183	+
2675_A	{GG}	{TT}	LG15	12494644	-
3481_A	{TT}	{CC}	LG16_21	33845307	+
3531_A	{AA}	{GG}	LG16_21	4401532	+
3547_A	{GG}	{AA}	LG16_21	4862896	-
3582_B	{GG}	{AA}	LG16_21	563234	-
4092_A	{TT}	{CC}	LG17	3619802	+

4092_B	{AA}	{GG}	LG17	3619802	+
4742_A	{CC}	{TT}	LG19	13369097	+
5412_A	{GG}	{TT}	LG1	2700436	-
5760_A	{CC}	{GG}	LG20	15924279	+
5761_B	{GG}	{CC}	LG20	15924451	-
5782_B	{AA}	{TT}	LG20	16532929	-
6698_B	{AA}	{GG}	LG23	11566547	-
7156_A	{GG}	{AA}	LG2	16158433	+
7403_B	{AA}	{GG}	LG2	394398	-
7902_B	{AA}	{CC}	LG4	12341840	+
7956_A	{CC}	{TT}	LG4	14941897	-
7956_B	{AA}	{CC}	LG4	14941897	-
8042_A	{AA}	{GG}	LG4	21081739	-
8084_A	{AA}	{GG}	LG4	22474196	-
9160_A	{TT}	{CC}	LG6	23563557	-
9251_A	{CC}	{TT}	LG6	2869131	+
9497_B	{TT}	{CC}	LG6	7648628	-
10120_A	{GG}	{AA}	LG7	40145502	-
10261_B	{GG}	{AA}	LG7	478353	-
10312_B	{TT}	{CC}	LG7	50192416	+
10718_A	{AA}	{GG}	LG8_24	25465631	+
10818_A	{AA}	{GG}	LG8_24	3390946	+
10818_B	{GG}	{AA}	LG8_24	3390946	+
10819_A	{CC}	{TT}	LG8_24	3391091	-
10819_B	{TT}	{CC}	LG8_24	3391091	-
11177_B	{TT}	{GG}	LG9	1907642	-
11223_A	{TT}	{CC}	LG9	3211685	+
11224_A	{AA}	{GG}	LG9	3211826	-
11771_B	{AA}	{GG}	LG16_21	4599202	+

b. Between *O. niloticus* - *O. aureus*

Marker ID	Genotyping		Chr	Position	DNA strand
	onil	oaur			
3887_B	{AA}	{GG}	LG17	21927644	+
8029_A	{AA}	{GG}	LG4	2024866	-
966_A	{CC}	{TT}	LG12	13862286	+
3236_B	{CC}	{TT}	LG16_21	18919216	+
3001_A	{GG}	{AA}	LG15	6075396	+
3057_B	{AA}	{TT}	LG15	909826	-
1276_B	{AA}	{GG}	LG12	3072836	-
1736_A	{CC}	{TT}	LG13	21872800	-
1992_A	{CC}	{TT}	LG13	8028370	+
2082_A	{AA}	{GG}	LG14	13140011	+
2082_B	{CC}	{AA}	LG14	13140011	+
2675_A	{GG}	{TT}	LG15	12494644	-
2890_B	{CC}	{TT}	LG15	24160607	-
2899_A	{CC}	{GG}	LG15	24719101	+
3531_A	{AA}	{GG}	LG16_21	4401532	+
3531_B	{CC}	{TT}	LG16_21	4401532	+
3873_A	{TT}	{CC}	LG17	21279151	-
5416_A	{CC}	{TT}	LG1	27201771	+
5424_A	{CC}	{TT}	LG1	27706614	-
5782_B	{AA}	{TT}	LG20	16532929	-
9418_A	{TT}	{CC}	LG6	36351716	+
9497_B	{TT}	{CC}	LG6	7648628	-

c. Between *O. niloticus* and *O. u. hornorum*

Marker ID	Genotyping		Chr	Position	DNA strand
	onil	ohor			
765_A	{GG}	{AA}	LG11	33107708	+
1109_B	{TT}	{CC}	LG12	22036688	+
1527_B	{CC}	{TT}	LG13	10210435	-
1756_A	{CC}	{TT}	LG13	23602908	+
2680_B	{CC}	{TT}	LG15	12784552	+
3887_B	{AA}	{GG}	LG17	21927644	+
4560_A	{GG}	{CC}	LG18	3512266	+
6674_A	{AA}	{GG}	LG23	10814820	+
8944_A	{TT}	{CC}	LG6	1059739	+
11268_A	{AA}	{GG}	LG9	5526004	-
2007_B	{GG}	{AA}	LG13	9212591	-
5929_B	{GG}	{AA}	LG20	24797769	+
7184_B	{CC}	{TT}	LG2	1791053	+
1542_B	{CC}	{TT}	LG13	10617897	+
967_B	{GG}	{TT}	LG12	13862572	-
1594_A	{AA}	{GG}	LG13	13361049	-
2887_A	{CC}	{TT}	LG15	24004304	+
3057_B	{AA}	{TT}	LG15	909826	-
3233_A	{TT}	{CC}	LG16_21	1885870	+
5208_A	{CC}	{TT}	LG1	10198606	+
8001_B	{GG}	{AA}	LG4	17759107	+
8436_A	{GG}	{TT}	LG5	14437062	-
8436_B	{AA}	{GG}	LG5	14437062	-
8603_A	{AA}	{GG}	LG5	24413535	+
10199_B	{GG}	{AA}	LG7	45218647	+
11084_B	{GG}	{AA}	LG9	14493472	-
294_A	{TT}	{CC}	LG11	10698975	+
435_B	{GG}	{AA}	LG11	18262482	+
465_A	{CC}	{GG}	LG11	20373159	-
482_A	{GG}	{AA}	LG11	21524281	-
581_A	{CC}	{TT}	LG11	26210952	-
633_A	{TT}	{CC}	LG11	28450485	+
801_A	{TT}	{CC}	LG11	4784136	-
925_B	{GG}	{AA}	LG12	1089338	+
1125_A	{TT}	{NN}	LG12	22867680	+
1276_B	{AA}	{GG}	LG12	3072836	-
1288_A	{AA}	{GG}	LG12	3123541	+
1678_A	{CC}	{TT}	LG13	19237959	+
1777_B	{GG}	{CC}	LG13	24956987	+

1916_A	{AA}	{CC}	LG13	4934070	-
1992_B	{AA}	{TT}	LG13	8028370	+
1993_A	{TT}	{AA}	LG13	8028573	-
2015_A	{TT}	{CC}	LG13	9915235	+
2082_A	{AA}	{GG}	LG14	13140011	+
2082_B	{CC}	{AA}	LG14	13140011	+
2451_A	{AA}	{GG}	LG14	31639274	-
2519_A	{TT}	{CC}	LG14	34149812	+
2675_A	{GG}	{TT}	LG15	12494644	-
3257_B	{GG}	{TT}	LG16_21	20086287	+
3258_A	{CC}	{AA}	LG16_21	20086434	-
3450_A	{GG}	{TT}	LG16_21	31395057	+
3481_A	{TT}	{CC}	LG16_21	33845307	+
3531_A	{AA}	{GG}	LG16_21	4401532	+
3547_A	{GG}	{AA}	LG16_21	4862896	-
3582_B	{GG}	{AA}	LG16_21	563234	-
3765_A	{GG}	{AA}	LG17	15770891	-
4092_A	{TT}	{CC}	LG17	3619802	+
4092_B	{AA}	{GG}	LG17	3619802	+
4270_A	{GG}	{AA}	LG18	1123631	-
5782_B	{AA}	{TT}	LG20	16532929	-
5920_B	{AA}	{GG}	LG20	24477975	-
6199_A	{GG}	{CC}	LG20	9394436	+
6611_B	{GG}	{AA}	LG22	7522007	-
6698_B	{AA}	{GG}	LG23	11566547	-
7156_A	{GG}	{AA}	LG2	16158433	+
7403_B	{AA}	{GG}	LG2	394398	-
7902_B	{AA}	{CC}	LG4	12341840	+
7956_A	{CC}	{TT}	LG4	14941897	-
7956_B	{AA}	{CC}	LG4	14941897	-
8042_A	{AA}	{GG}	LG4	21081739	-
9160_A	{TT}	{CC}	LG6	23563557	-
9251_A	{CC}	{TT}	LG6	2869131	+
9412_A	{AA}	{GG}	LG6	36061135	+
9659_A	{GG}	{AA}	LG7	17150157	-
10261_B	{GG}	{AA}	LG7	478353	-
10312_B	{TT}	{CC}	LG7	50192416	+
10718_A	{AA}	{GG}	LG8_24	25465631	+
10793_A	{AA}	{GG}	LG8_24	2824094	-
10818_A	{AA}	{GG}	LG8_24	3390946	+
10819_B	{TT}	{CC}	LG8_24	3391091	-
10951_A	{AA}	{TT}	LG8_24	8700535	-
11177_B	{TT}	{GG}	LG9	1907642	-

11223_A	{TT}	{CC}	LG9	3211685	+
11224_A	{AA}	{GG}	LG9	3211826	-
11771_B	{AA}	{GG}	LG16_21	4599202	+

d. Between *O. mossambicus* - *O. aureus*

Marker ID	Genotyping		Chr	Position	DNA strand
	omos	oaur			
765_A	{AA}	{GG}	LG11	33107708	+
1504_B	{TT}	{CC}	LG12	8753442	-
1756_A	{TT}	{CC}	LG13	23602908	+
4411_B	{CC}	{TT}	LG18	20048688	-
4560_A	{CC}	{GG}	LG18	3512266	+
6674_A	{GG}	{AA}	LG23	10814820	+
8029_A	{AA}	{GG}	LG4	2024866	-
8944_A	{CC}	{TT}	LG6	1059739	+
1748_A	{TT}	{CC}	LG13	22809186	+
2007_B	{AA}	{GG}	LG13	9212591	-
7184_B	{TT}	{CC}	LG2	1791053	+
966_A	{CC}	{TT}	LG12	13862286	+
3236_B	{CC}	{TT}	LG16_21	18919216	+
3001_A	{GG}	{AA}	LG15	6075396	+
3233_A	{CC}	{TT}	LG16_21	1885870	+
5585_B	{AA}	{TT}	LG1	5654752	-
8001_B	{AA}	{GG}	LG4	17759107	+
8436_A	{TT}	{GG}	LG5	14437062	-
8436_B	{GG}	{AA}	LG5	14437062	-
11084_B	{AA}	{GG}	LG9	14493472	-
57_B	{GG}	{TT}	LG10	13769218	+
294_A	{CC}	{TT}	LG11	10698975	+
435_B	{AA}	{GG}	LG11	18262482	+
465_A	{GG}	{CC}	LG11	20373159	-
633_A	{CC}	{TT}	LG11	28450485	+
801_A	{CC}	{TT}	LG11	4784136	-
801_B	{CC}	{TT}	LG11	4784136	-
1125_A	{AA}	{TT}	LG12	22867680	+
1288_A	{GG}	{AA}	LG12	3123541	+
1678_A	{TT}	{CC}	LG13	19237959	+
1736_A	{CC}	{TT}	LG13	21872800	-
1777_B	{CC}	{GG}	LG13	24956987	+
1992_A	{CC}	{TT}	LG13	8028370	+
2451_A	{GG}	{AA}	LG14	31639274	-
2657_B	{AA}	{GG}	LG15	11282183	+

2890_B	{CC}	{TT}	LG15	24160607	-
2899_A	{CC}	{GG}	LG15	24719101	+
3435_A	{TT}	{GG}	LG16_21	30668793	+
3481_A	{CC}	{TT}	LG16_21	33845307	+
3531_B	{CC}	{TT}	LG16_21	4401532	+
3547_A	{AA}	{GG}	LG16_21	4862896	-
3547_B	{TT}	{GG}	LG16_21	4862896	-
3582_B	{AA}	{GG}	LG16_21	563234	-
3873_A	{TT}	{CC}	LG17	21279151	-
4079_A	{GG}	{CC}	LG17	3491547	+
4092_A	{CC}	{TT}	LG17	3619802	+
4092_B	{GG}	{AA}	LG17	3619802	+
4267_A	{TT}	{CC}	LG18	1123143	+
4368_B	{CC}	{TT}	LG18	18054452	-
4555_A	{GG}	{TT}	LG18	3292409	+
4684_A	{AA}	{GG}	LG18	9534063	-
4742_A	{TT}	{CC}	LG19	13369097	+
4935_A	{GG}	{AA}	LG19	21031988	+
5084_A	{CC}	{AA}	LG19	318532	+
5412_A	{TT}	{GG}	LG1	2700436	-
5416_A	{CC}	{TT}	LG1	27201771	+
5424_A	{CC}	{TT}	LG1	27706614	-
5760_A	{GG}	{CC}	LG20	15924279	+
5761_B	{CC}	{GG}	LG20	15924451	-
6698_B	{GG}	{AA}	LG23	11566547	-
7015_B	{GG}	{AA}	LG23	8019952	-
7156_A	{AA}	{GG}	LG2	16158433	+
7403_B	{GG}	{AA}	LG2	394398	-
7956_A	{TT}	{CC}	LG4	14941897	-
7956_B	{CC}	{AA}	LG4	14941897	-
8084_A	{GG}	{AA}	LG4	22474196	-
8236_A	{AA}	{GG}	LG4	3340101	+
8238_B	{TT}	{CC}	LG4	3340249	-
9160_A	{CC}	{TT}	LG6	23563557	-
9251_A	{TT}	{CC}	LG6	2869131	+
9418_A	{TT}	{CC}	LG6	36351716	+
9432_A	{CC}	{TT}	LG6	3920256	-
10120_A	{AA}	{GG}	LG7	40145502	-
10261_B	{AA}	{GG}	LG7	478353	-
10312_B	{CC}	{TT}	LG7	50192416	+
10563_B	{CC}	{GG}	LG8_24	18990874	+
10718_A	{GG}	{AA}	LG8_24	25465631	+
10818_A	{GG}	{AA}	LG8_24	3390946	+

10818_B	{AA}	{GG}	LG8_24	3390946	+
10819_A	{TT}	{CC}	LG8_24	3391091	-
10819_B	{CC}	{TT}	LG8_24	3391091	-
10956_A	{GG}	{AA}	LG8_24	8727508	+
11177_B	{GG}	{TT}	LG9	1907642	-
11223_A	{CC}	{TT}	LG9	3211685	+
11224_A	{GG}	{AA}	LG9	3211826	-
11771_B	{GG}	{AA}	LG16_21	4599202	+

e. Between *O. mossambicus* - *O. u. hornorum*

Marker ID	Genotyping		Chr	Position	DNA strand
	omos	ohor			
1504_B	{TT}	{CC}	LG12	8753442	-
2680_B	{CC}	{TT}	LG15	12784552	+
8603_A	{AA}	{GG}	LG5	24413535	+
10199_B	{GG}	{AA}	LG7	45218647	+
1125_A	{AA}	{NN}	LG12	22867680	+
2519_A	{TT}	{CC}	LG14	34149812	+
2657_B	{AA}	{GG}	LG15	11282183	+
4270_A	{GG}	{AA}	LG18	1123631	-
4742_A	{TT}	{CC}	LG19	13369097	+
5412_A	{TT}	{GG}	LG1	2700436	-
5760_A	{GG}	{CC}	LG20	15924279	+
5761_B	{CC}	{GG}	LG20	15924451	-
9497_B	{CC}	{TT}	LG6	7648628	-
10120_A	{AA}	{GG}	LG7	40145502	-
10793_A	{AA}	{GG}	LG8_24	2824094	-
10818_B	{AA}	{GG}	LG8_24	3390946	+
10819_A	{TT}	{CC}	LG8_24	3391091	-
10951_A	{AA}	{TT}	LG8_24	8700535	-

f. Between *O. aureus* - *O. u. hornorum*

Marker ID	Genotyping		Chr	Position	DNA strand
	oaur	ohor			
765_A	{GG}	{AA}	LG11	33107708	+
1109_B	{TT}	{CC}	LG12	22036688	+
1527_B	{CC}	{TT}	LG13	10210435	-
1756_A	{CC}	{TT}	LG13	23602908	+
2680_B	{CC}	{TT}	LG15	12784552	+
4411_B	{TT}	{CC}	LG18	20048688	-
4560_A	{GG}	{CC}	LG18	3512266	+
6674_A	{AA}	{GG}	LG23	10814820	+
7321_B	{GG}	{AA}	LG2	24612566	+
8029_A	{GG}	{AA}	LG4	2024866	-
8944_A	{TT}	{CC}	LG6	1059739	+
11268_A	{AA}	{GG}	LG9	5526004	-
1748_A	{CC}	{TT}	LG13	22809186	+
2007_B	{GG}	{AA}	LG13	9212591	-
5929_B	{GG}	{AA}	LG20	24797769	+
7184_B	{CC}	{TT}	LG2	1791053	+
966_A	{TT}	{CC}	LG12	13862286	+
1542_B	{CC}	{TT}	LG13	10617897	+
3236_B	{TT}	{CC}	LG16_21	18919216	+
967_B	{GG}	{TT}	LG12	13862572	-
2887_A	{CC}	{TT}	LG15	24004304	+
3001_A	{AA}	{GG}	LG15	6075396	+
3233_A	{TT}	{CC}	LG16_21	1885870	+
5208_A	{CC}	{TT}	LG1	10198606	+
5585_B	{TT}	{AA}	LG1	5654752	-
8001_B	{GG}	{AA}	LG4	17759107	+
8436_A	{GG}	{TT}	LG5	14437062	-
8436_B	{AA}	{GG}	LG5	14437062	-
8603_A	{AA}	{GG}	LG5	24413535	+
10199_B	{GG}	{AA}	LG7	45218647	+
11084_B	{GG}	{AA}	LG9	14493472	-
294_A	{TT}	{CC}	LG11	10698975	+
435_B	{GG}	{AA}	LG11	18262482	+
465_A	{CC}	{GG}	LG11	20373159	-
482_A	{GG}	{AA}	LG11	21524281	-
581_A	{CC}	{TT}	LG11	26210952	-
633_A	{TT}	{CC}	LG11	28450485	+
801_A	{TT}	{CC}	LG11	4784136	-
801_B	{TT}	{CC}	LG11	4784136	-

925_B	{GG}	{AA}	LG12	1089338	+
1125_A	{TT}	{NN}	LG12	22867680	+
1288_A	{AA}	{GG}	LG12	3123541	+
1678_A	{CC}	{TT}	LG13	19237959	+
1736_A	{TT}	{CC}	LG13	21872800	-
1777_B	{GG}	{CC}	LG13	24956987	+
1916_A	{AA}	{CC}	LG13	4934070	-
1992_A	{TT}	{CC}	LG13	8028370	+
1992_B	{AA}	{TT}	LG13	8028370	+
1993_A	{TT}	{AA}	LG13	8028573	-
2451_A	{AA}	{GG}	LG14	31639274	-
2519_A	{TT}	{CC}	LG14	34149812	+
2890_B	{TT}	{CC}	LG15	24160607	-
2899_A	{GG}	{CC}	LG15	24719101	+
3257_B	{GG}	{TT}	LG16_21	20086287	+
3258_A	{CC}	{AA}	LG16_21	20086434	-
3435_A	{GG}	{TT}	LG16_21	30668793	+
3450_A	{GG}	{TT}	LG16_21	31395057	+
3481_A	{TT}	{CC}	LG16_21	33845307	+
3531_B	{TT}	{CC}	LG16_21	4401532	+
3547_A	{GG}	{AA}	LG16_21	4862896	-
3547_B	{GG}	{TT}	LG16_21	4862896	-
3582_B	{GG}	{AA}	LG16_21	563234	-
3697_B	{AA}	{CC}	LG17	12228708	-
3873_A	{CC}	{TT}	LG17	21279151	-
4079_A	{CC}	{GG}	LG17	3491547	+
4092_A	{TT}	{CC}	LG17	3619802	+
4092_B	{AA}	{GG}	LG17	3619802	+
4267_A	{CC}	{TT}	LG18	1123143	+
4270_A	{GG}	{AA}	LG18	1123631	-
4368_B	{TT}	{CC}	LG18	18054452	-
4555_A	{TT}	{GG}	LG18	3292409	+
4684_A	{GG}	{AA}	LG18	9534063	-
4935_A	{AA}	{GG}	LG19	21031988	+
4935_B	{TT}	{CC}	LG19	21031988	+
5084_A	{AA}	{CC}	LG19	318532	+
5416_A	{TT}	{CC}	LG1	27201771	+
5424_A	{TT}	{CC}	LG1	27706614	-
5920_B	{AA}	{GG}	LG20	24477975	-
6199_A	{GG}	{CC}	LG20	9394436	+
6611_B	{GG}	{AA}	LG22	7522007	-
6698_B	{AA}	{GG}	LG23	11566547	-
7015_B	{AA}	{GG}	LG23	8019952	-

7156_A	{GG}	{AA}	LG2	16158433	+
7403_B	{AA}	{GG}	LG2	394398	-
7956_A	{CC}	{TT}	LG4	14941897	-
7956_B	{AA}	{CC}	LG4	14941897	-
8236_A	{GG}	{AA}	LG4	3340101	+
8238_B	{CC}	{TT}	LG4	3340249	-
8543_B	{CC}	{AA}	LG5	20953232	-
9160_A	{TT}	{CC}	LG6	23563557	-
9251_A	{CC}	{TT}	LG6	2869131	+
9412_A	{AA}	{GG}	LG6	36061135	+
9418_A	{CC}	{TT}	LG6	36351716	+
9432_A	{TT}	{CC}	LG6	3920256	-
9497_B	{CC}	{TT}	LG6	7648628	-
9659_A	{GG}	{AA}	LG7	17150157	-
10261_B	{GG}	{AA}	LG7	478353	-
10312_B	{TT}	{CC}	LG7	50192416	+
10563_B	{GG}	{CC}	LG8_24	18990874	+
10718_A	{AA}	{GG}	LG8_24	25465631	+
10793_A	{AA}	{GG}	LG8_24	2824094	-
10818_A	{AA}	{GG}	LG8_24	3390946	+
10819_B	{TT}	{CC}	LG8_24	3391091	-
10951_A	{AA}	{TT}	LG8_24	8700535	-
10956_A	{AA}	{GG}	LG8_24	8727508	+
11177_B	{TT}	{GG}	LG9	1907642	-
11223_A	{TT}	{CC}	LG9	3211685	+
11224_A	{AA}	{GG}	LG9	3211826	-
11771_B	{AA}	{GG}	LG16_21	4599202	+

Appendix V.3. The complete mapping of species – specific SNP markers derived from double digest RADseq