University of Natural Resources and Life Sciences, Vienna

EGERTON UNIVERSITY

Microsatellite Cross-Species Amplification and Utility in Selected African Cichlids: A Valuable Tool for Tilapiine Fishery Management and Conservation

Thesis submitted for the award of
Master of Science
By:

John Mwaura Kariuki

Supervised by:

Harald Meimberg

Univ. Prof. Dipl. - Biol. Dr.rer.nat

Thesis submitted in partial fulfilment of the requirements for the award of
Joint academic degree of

Master of Science in Limnology and Wetland Management

University of Natural Resources and Life Sciences (BOKU), Vienna, Austria

April, 2018

University of Natural Resources and Life Sciences, Vienna

IHE DELFT

EGERTON UNIVERSITY

Microsatellite cross-species amplification and utility in Selected African cichlids: A valuable tool for tilapiine fishery management and conservation

Thesis submitted for the award of the title
Master of Science
By:

John Mwaura Kariuki

Supervised by:

Prof. Harald Meimberg

Co-supervised by:

Cardoso Curto, Manuel Antonio, Dr.

And Tibihika, Papius Dias

Thesis submitted in partial fulfilment of the requirements of the award of Joint academic degree of

Master of Science in Limnology and Wetland Management

Jointly awarded by
The University of Natural Resources and Life Sciences (BOKU), Vienna, Austria
The UNESCO-IHE Institute for Water Education, Delft, the Netherlands
Egerton University, Njoro, Kenya

University of Natural Resources and Life Sciences (BOKU), Vienna, Austria

Copyright ©

Declaration

This is my original work and has not been submitted to any other institution/University for the award of a degree.

Name: John Mwaura Kariuki

Registration number…….

Signature…….Date……

This thesis is submitted with the approval of main supervisor:

Name……..

Signature……..Date……..

Dedication

I dedicate this work to my family, especially my two boys, Lewis Kariuki and Marcus Kiraba, who stayed for a long time without their Daddy at home.

Acknowledgement

Firstly, I would like to thank God Almighty for the gift of life and health, to me and to my family, during the entire period of study. Secondly, I would like to pass my sincere gratitude to my supervisor, Professor Harald Meimberg, co-supervisors Dr. Manuel Antonio Curto and Papius Tibihika for their discerning ideas in addition to reading my thesis under time constraints, I will always be grateful to you.

Not to be forgotten is the Government of Netherlands for sponsoring my studies through the National Fellowship Program (NFP), may you continue to support students from needy countries for the advancement of the World. Special thanks also to Austrian Development Corporation (ADC) for sponsoring my fieldwork, I honestly appreciate. I would also like to thank the University of Nairobi for affording me the time to study abroad by offering me study leave, may you continue to be a fountain of knowledge.

Much appreciation also goes to the Department of Biological sciences, Egerton University, Kenya and the UNESCO-IHE Institute for Water Education, Delft, Netherlands for the semester period I studied there. They were such great institutions conducive for studies.

In this list also is University of Natural Resources and Life Sciences, Vienna for the three semesters I studied there, one for my coursework and two for my thesis. It was such a good place to study and do laboratory work. While at it, I wouldn't forget to thank Eva Dornstauder-Schrammel for her laboratory work assistance, life in the laboratory would have been unbearable without you.

I would also like to acknowledge the IPGL office for their good work in supporting me during my entire stay and studies in Vienna. In this list I would like to include Lisa Kargl, Susanne Kagl, Marie Theres Pfeiffer, Gerold Winkler and Nina Haslinger for their unwavering support. Special thanks to Nina for her exceptional character, she went out of her way to offer encouragement when I felt like giving up, receiving distress calls and replying to e-mails out of office hours just to keep me comfortable, that was special, may you be Blessed. And to all my LWM-2016 classmates, you were a special group, you will always be on my mind.

Finally, I would like to acknowledge my wife for playing the role of a Mum and Dad while I was away, you did a great job, keep it up. Many thanks also to my parents and parents in law for always believing in me, supporting and praying for me, I really appreciate.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Appendices

List of Abbreviations

µl	Microliter
µM	Micromolar
ACGT	Adenine, Cytosine, Guanine and Thymine (Nucleotide bases)
BLASTN	Basic Local Alignment Search Tools for Nucleotides
BP	Base Pair
CLUMPAK	Cluster Markov Packager Across K
DNA	Deoxynucleic acid
EDTA	Ethylene Diamine Tetra-Acetic acid
ePCR	Emulsion Polymerase Chain Reaction
EST	Expressed Sequence Tag
EtOH	Ethanol
GenAlEx	Genetic Analysis in Excel
GIFT	Genetically Improved Farmed Tilapia
GuHCl	Guanidine Hydrochloride
HCl	Hydrochloric acid
He	Expected heterozygosity
Ho	Observed heterozygosity
MCMC	Markov Chain Monte Carlo
MHC	Major Histocompatibility Complex
mM	Millimolar
NaCl	Sodium Chloride
NCBI	National Centre for Biotechnology Information
NGS	Next Generation Sequencing
PCoA	Principal Coordinate Analysis
PCR	Polymerase Chain Reaction
PIC	Polymorphic Information Content
PVP	Poly Vinyl Pyrrolidone
QTL	Quantitative Trait Loci

RAPD	Random Amplification of Polymorphic DNA
RFLP	Restriction Fragment Length Polymorphism
RPM	Revolutions Per Minute
SDS	Sodium Dodecyl Sulphate
SNP	Single Nucleotide Polymorphism
SRA	Sequence Read Archive
SSLP	Simple Sequence Length Polymorphism
SSR	Simple Sequence Repeat
SSTR	Simple Sequence Tandem Repeat
STMS	Sequence Tagged Microsatellites
STR	Short Tandem Repeat
TAE	Tris-Acetic acid EDTA
TL	Total Length (in fish measurement)
VNTR	Variable Number Tandem Repeats
WSI	Whole Sequence Information

Abstract

Oreochromis niloticus is one of the most farmed fish in the world, coming only second to carp. This has led to its massive relocation far from its native range mainly for aquaculture purposes. As a result, it has hybridized with other tilapiine species thus running the risk of losing pure breeds through introgression. This study aimed at developing microsatellite markers from *O*reochromis *niloticus*, designing primers for these markers and then testing their ability to cross-amplify DNA from 2 other tilapiine species i.e. *Sarotherodon galilaeus*, *Tilapia zillii* during a PCR reaction. The study also aimed at using the developed microsatellite markers to determine genetic structure of these tilapiine species. Consequently, it was possible to postulate the level of hybridization within these species. Forty-two microsatellite markers were developed consisting of the following repeats: 8 dinucleotides, 9 trinucleotides, 16 tetra nucleotides and 9 penta nucleotides. After running PCR reaction using the designed primers, amplicons were sequenced on a high throughput illumina MiSeq sequencing platform. Two methods were used in genotyping the sequence data and the results compared in terms of genetic structure and allelic numbers they produced. Whole sequence information (WSI) method, also called haplotype, gave the highest values in mean alleles per locus and polymorphic information content at 14.86 and 0.58, while sequence length genotyping method gave the least at 2.02 and 0.22 respectively. Thirty-two out of the 42 of the microsatellite primer pairs designed cross-amplified in *S. galilaeus* and 23 in *Coptodon zillii*. Possible hybridization and hence gene flow was observed especially in the two Nile tilapia populations from Lake Victoria and Lake Albert. Microsatellite markers developed in this study showed high level of cross-amplification and polymorphism and thus usable in monitoring hybridization in tilapiine species.

Keywords: Hybridization, Genetic structure, Cross-species amplification, Microsatellite markers, introgression

CHAPTER 1

1.0 Introduction

African cichlids are important in the study of evolutionary biology and applied genetics. This is mostly due to their species richness, with over 1600 species in around 220 genera officially described and documented as valid, but may count up to 3000 species (Dunz and Schliewen, 2013). Three lakes in Africa; Lakes Malawi, Tanganyika and Victoria have been termed as the centres for adaptive radiation of these cichlids, although comprehensive sampling is yet to been done, especially in Lake Victoria (Turner et al., 2001). In this regard, cichlids have been used in model studies for speciation, adaptation, behaviour and neurosciences (Kornfield and Smith, 2000). They have also been extensively studied as groups of importance for aquaculture and aquaponics, for strain selection, genetic improvement like the GIFT fish and stock assessment (Eknath and Hulata, 2009; Weyl et al., 2010).

The name Tilapia has been frequently used to refer to 3 genera i.e. *Oreochromis*, *Sarotherodon* and *Tilapia* (Fuller et al., 1998). The genus *Oreochromis* is the largest with over 40 species, followed by *Sarotherodon* with 13 described species. On the other hand, *Tilapia* has four species, although many more remain undescribed (Nagl et al., 2001). Their native ranges are in Africa, Latin America and Madagascar, and a few species in the Middle East. This may suggest that when the supercontinent of Gondwana started to split up around 120–160 million years ago, cichlids were already in existence (Turner et al., 2001).

Consequently, all the three genera have been utilized for aquaculture and in sport fishing. Globally, tilapia farming has expanded exponentially to satisfy the ever growing local and international markets (FAO, 2010). This expansion in production has raised tilapia status to the second most popular farmed fish after carps, with global production of at least 85 countries exceeding 2.5 million tonnes in 2007 (FAO, 2010).

There are several tilapia species that are cultured, with Nile tilapia, *Oreochromis niloticus* Linnaeus, 1758, remaining the most preferred in at least 50 countries, representing 80% of global production in the last 5 years (Eknath and Hulata, 2009). Most of these countries are outside Africa, although the natural distribution of tilapia genetic resource is found in Africa and some parts of the Middle East. These natural populations face severe threats which include irreversible loss or change due to factors such as pollution of natural habitats,

uncontrolled fishing, undocumented fry stocking, irresponsible and uncontrolled fish relocations among others (Eknath and Hulata, 2009).

Similarly, as tilapias have shown great tendency to hybridize, both inter and intra-generic hybridization in natural habitats have been reported (Rognon et al., 1996). This happens when originally allopatric populations are brought together through these translocations. Other intra-generic hybrids are obtained intentionally by fish-farmers who cross-breed to obtain an all-male population as they attain market size faster than their female counterparts (Rognon et al., 1996). These transfers and translocations promote hybridization, both intra and inter-generic. *Tilapia zillii* and *Sarotherodon galilaeus* are potential candidates for this inter-generic breeding and hence will be investigated in this study through their ability to be cross-amplified by SSR primers developed from *Oreochromis niloticus*.

Also, to be investigated in this study is the structure within the *Oreochromis niloticus* subpopulations. Genetic structure is a consequence of lack of exchange of genetic material through mating as a result of geographical separation (Colonna et al., 2009). This is important for a number of reasons among them, conservation, association mapping, detection of migration, adaptation studies as well as for describing habitats and barriers involved in separation (Fogelqvist et al., 2010; Guillot, 2008). Translocations may lead to the formation of distinct genetic structures due to the departure from panmixia population with the ultimate formation of subpopulations. Detecting genetic structure is a good way to signal hybridization, and therefore an important way to monitor hybridization within tilapiine species.

1.1 Research problem

The rampant hybridization in tilapia is likely to lead to loss of native genetic identity as well as introduction of hybrids that may not be well adapted to the local habitats (Nagl et al., 2001). Other forms of intra-generic hybridization produce an all-male population thereby reducing reproducing females (Rognon et al., 1996). Additionally, hybrids may introduce into farmed fish genotypes that are not well adapted, thereby reducing fish-farmers' profits. This may lead to collapse of fishery, with accompanying countless financial losses.

Hybridizations happen because of constant fish translocations, sometimes between continents, bringing together formally isolated genera. For example, the introduction of Nile tilapia to the Lake Victoria that led to the extinction of native species (Ogutu-Ohwayo, 1990).

Therefore, determination of genetic variation will provide valuable information on the genetic structure of the cultured stock allowing for sustainable and species-specific management of stocks.

Studies have been done to determine the genetic structure of Nile tilapia and *Tilapia zillii* but they have used molecular markers with low resolving powers such allozymes (Rognon et al., 1996), randomly amplified polymorphic DNA (RAPD), (Suresh et al., 2013) among others. These traditional methods however, are susceptible to artefacts and machine specific biases, making the results they produce neither reliable nor replicable (Hedrick, 2001). This study used microsatellites generated through a high throughput DNA sequencing technique, illumina MiSeq platform and used microsatellites markers which have higher resolving power, they are evenly distribution throughout the genome (Tautz, 1989; Weber and Wong, 1993). Most importantly, microsatellites are co-dominantly inherited in strict Mendelian fashion (Costa-Pierce, 2003), unless there has been high mutations. This property allows for investigation of hybrid populations in the succeeding generations (Costa-Pierce, 2003).

Obtained results were sequenced rather than measuring sequence lengths of amplified portions as was traditionally done. This enhanced the information obtained which may later lead to establishment of better marker genotyping systems in the future. In addition, sequencing reduced errors such as migration artefacts in gel electrophoresis based sequencing machines. Therefore, the markers developed here could become a standard practice in future for the detection of hybrids between these species.

1.2 Research questions

1. Can microsatellites, as genetic markers, be applied to detect genetic structure and hybridization in Tilapia?

2. (a) Has the high level of tilapia translocation led to introgression of Nile tilapia genepool by other tilapiine species?

(b). Is there a genetic threat, as a result of this introgression, which may necessitate call for conservation of the Nile tilapia?

1.3 Objectives

1. To design SSR primers for East African *Oreochromis niloticus* using Next Generation Sequencing (NGS) technology

2. To test the developed SSR primers for cross-amplification on *Sarotherodon galilaeus, Tilapia zillii*.

3. To determine the genetic structure of *Oreochromis niloticus*, *Sarotherodon galilaeus, Tilapia zillii* from various habitats in East and West Africa

4. To detect the level of hybridization within *Oreochromis niloticus*, *Tilapia zillii,* and *Sarotherodon galilaeus* in different habitats

5. To compare two allele calling methods i.e., whole sequence information and sequence length

CHAPTER 2

2.0 Literature review

2.1 *Oreochromis niloticus* and related genera

The genus *Oreochromis* is the largest with over 40 species, followed by *Sarotherodon* with 13 species, while *Tilapia* has four species (Nagl et al., 2001). Both *Oreochromis* and *Sarotherodon* are mouth-brooders, while *Tilapia* is a substrate spawner (Bezault et al., 2012). This reproductive system has led scientists to believe that *Oreochromis* and *Sarotherodon* diverged recently as compared to *Tilapia* which may have diverged earlier (Bezault et al., 2012). Included in these genera are *O. niloticus* in the genus *Oreochromis, Sarotherodon galilaeus* and *Tilapia zillii* in *Sarotherodon and Tilapia* genera respectively.

2.1.1 *Oreochromis niloticus*

Oreochromis niloticus is native to several Lakes in Africa and the Middle East. It is an omnivorous prolific breeder, has a long-life span (about nine years in the wild), attains table size rapidly, and therefore easily out-compete native species from feeding and breeding grounds (Goudswaard et al., 2002; Lowe-McConnell, 2000; Njiru et al., 2010). It is also an efficient algae feeder compared to native species in many water bodies further enhancing its competitiveness (Canonico et al., 2005). Moreover, *O. niloticus* is well-suited ecologically, genetically and physiologically for commercial fisheries as well as successful invader of ecosystems all over the tropical and subtropical regions of the world (Canonico et al., 2005).

It is thus frequently regarded as 'pioneer' species as it often migrates and reproduces within themselves and with native populations (Trewavas and Teugels, 1991). Females and males grow at different rates with males attaining market-size earlier than females. On the other hand, females reach sexual maturity earlier than males, easily filling the water body with small ones (Mair et al., 1997). Though environmental factors play a role, genetics is the main factor in sex determination in *O. niloticus* (Mair et al., 1997). Figure 1 shows an image of *O*. *niloticus,* as obtained from Wikipedia in the link (https://en.wikipedia.org/wiki/Tilapia).

Molecular markers such as microsatellites and amplified fragment length polymorphism (AFLP) have been used in Nile tilapia to generate genomic maps. These maps were later used to construct marker-assisted breeding programs for further detection of quantitative trait loci (QTL) involved in salinity and cold tolerance as well as carcass quality (Agresti et al., 2000; Kocher et al., 1998).

Figure 1. Showing *Oreochromis niloticus*. Adapted from Wikipedia

2.1.2 Sarotherodon galilaeus

Sarotherodon galilaeus (mango tilapia) is a species of fish from the cichlid family. The fish is also referred to as; Galilee St. Peter's fish, Galilean comb, Galilaea tilapia as well as St. Peter's fish (Trewavas and Teugels, 1991). The fish is native to Africa and Eurasia with distribution ranging from the Jordan system, the coastal rivers of Israel, the Nile system, to Lake Albert and Turkana (Fisheries, 2010).

Mating system varies from population to population, nonetheless many are bi-parental mouth brooders (Ros et al., 2003). The fish adapts well to different environmental conditions. Consequently, the fish can occasionally be found at temperatures of about 9°C, while a few can live in salty waters albeit for a short period. Additionally, it sometimes form schools, and is territorial (Ros et al., 2003). Usually, adults are often found in open waters whereas juveniles and breeding adults prefer inshore (Trewavas and Teugels, 1991). Mango tilapia feeds on algae and other fine organic debris. When sexually active, it is not easily noticeable since they do not have distinct sexual dichromatism as other fish though they form temporary pair bonds. The maximum size ever recorded was 41.0 cm TL (Bailey, 1994). Figure 2 shows an example of S. galilaeus, as obtained from Wikipedia at the link below.

(http://www.ag.auburn.edu/fish/image_gallery/details.php?image_id=1187&sessionid=eb4e832e58fada).

Figure 2. *Sarotherodon galilaeus* (mango tilapia). Adapted from the Wikipedia

2.1.3 *Coptodon zillii*

Coptodon zillii is also commonly referred to as redbelly tilapia or *Tilapia zillii*. It is also native to Africa and the Middle East although it has been introduced outside its native range mainly for aquaculture (Levêque et al., 2008). It can thrive in different environmental conditions ranging from fresh water to brackish water and in depths ranging from 1-7 meters, though it favours shallow vegetated areas (Lung'Ayia et al., 2000). The fry often lives in the marginal vegetation, while juveniles prefer seasonal floodplains. Occasionally, the fish form schools mainly during the day.

The fish primarily feeds on macrophytes (Levêque et al., 2008). Previously, Lake Victoria was dominated by diatoms and blue green algae but presently the lake is dominated by blue greens (Lung'Ayia et al., 2000). These types of algae may contain toxic groups thus offering unpleasant food for the fish. Subsequently, *Coptodon zillii* populations have dwindled considerably over the last decade (Ogutu-Ohwayo, 1990). This fish is implicated in the loss of two Lake Victoria native species, *Oreochromis variabilis* and *Oreochromis esculentus*, through direct competition for food and habitat or through genetic competition as they readily hybridize with other Tilapia species (Spataru, 1978). In addition to its commercial importance,

Tilapia zillii has been utilized ecologically for weed control as well as recreational fish in a number of countries (Mehanna, 2004). Figure 3 shows an example of redbelly tilapia, as obtained from Wikipedia in the link below.
(https://www.google.at/search?q=Redbelly+tilapia+or+Tilapia+zillii&dcr=0&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjBy7OL4p3aAhWDHpoKHdyKCCAQ_AUICigB&biw=1366&bih=611#imgrc=4q8Qwa1FVOEKyM :).

Figure 3. Redbelly tilapia or *Tilapia zillii*. Adapted from Wikipedia

2.2 Molecular markers

A number of molecular markers are used in aquaculture genetics research (Liu and Cordes, 2004). These markers are categorised into two main types, type I and type II markers. Type I markers are associated with known functional genes, while type II are those with unspecified genomic fragments (Liu and Cordes, 2004).

According to this classification, the following markers are recognized under type I markers; Restriction Fragment Length Polymorphism (RFLP), allozyme and Expressed Sequence Tag (EST) (O'brien, 1991). On the other hand, type II markers include Randomly Amplified Polymorphic DNA (RAPD), Amplified Fragment Length Polymorphism (AFLP), microsatellites (when not associated with genes of known function) as well as Single Nucleotide

Polymorphism (SNP), when they are not developed from expressed sequences for instance expression SNP and coding SNP, (O'brien, 1991). Type II markers such as AFLP and RAPD have been used successfully in aquaculture genetics for hybrid, species and strain identification, inbreeding studies and in the study of markers linked to Qualitative Trait Loci (QTL), although their resolving powers are low (Epifanio et al., 2003).

To determine the usability of such markers above, Polymorphic Information Content (PIC), which refers to the value of a marker for detecting polymorphism in a population, is used (Botstein et al., 1980). It measures the effectiveness of molecular markers and therefore their usability. It is equivalent to 1 minus the sum of square of all allele frequencies (Botstein et al., 1980). Additionally, PIC relies on the number of detectable alleles as well as their distribution frequencies. These molecular markers play a vital role in resolving numerous taxonomic challenges in threatened and endangered species as well as in designing breeding systems allowing genetic conservation (Caro and Laurenson, 1994).

2.2.1 Microsatellites

Microsatellites are simple, tandem repeat sequences of about one to six base pairs (BP). Microsatellites loci are also referred to as: Simple Sequence Repeats (SSR's), Short Tandem Repeats (STR's), Simple Sequence Tandem Repeats (SSTR), Variable Number Tandem Repeats (VNTR), Simple Sequence Length Polymorphisms (SSLP), Sequence Tagged Microsatellites (STMS) (Jarne and Lagoda, 1996; Liu et al., 1998). They are found anywhere in the genome, both in coding and non-coding regions of the DNA (Tóth et al., 2000). Dinucleotide are the most common while trinucleotide and hexanucleotide repeats usually appear in coding regions since they do not lead to frameshift mutations (Tóth et al., 2000). The molecular basis of these markers is the difference in number of repeats (Liu et al., 1998).

Simple microsatellites contain one type of repeat while compound microsatellites contain more than one type of repeat (Liu et al., 1998). Microsatellite are ideal molecular markers since they are highly polymorphic owing to their high rate of mutations (between 10^{-3} and 10^{-4} mutations per gamete per generation), even distribution throughout the genome, and are co-dominantly inherited (Tautz, 1989; Weber and Wong, 1993).

Studies estimate that approximately 1–4% of the genome consists of microsatellites, and at least one microsatellite appears about every 10 kilobases in fishes (Wright, 1993).

Microsatellites have therefore proven to be powerful molecular markers for quantifying genetic variations within and between population of species and in delineation of individual strains (Muneer et al., 2009).

2.2.2 Conventional marker development methods

Traditional methods of marker development involved fragmentation of high quality genomic DNA using either restriction enzymes or, less frequently by sonication (Zane et al., 2002). The desired average length of DNA fragments, the microsatellite repeat to be found and the type of ends (cohesive or blunt) of the restriction fragments were the factors used to determine the choice of restriction enzyme to be used (Zane et al., 2002). This was then followed by ligation of the DNA fragments into a plasmid vector either directly or after ligating them to specific adaptors. Thousands of recombinant clones were then generated following transformation of bacterial cells with the ligation products, which were then screened for presence of microsatellite sequences (Ostrander et al., 1992). However, this made identification and development of microsatellites costly and time-consuming hence the search for better alternative methods (Glenn and Schable, 2005).

2.2.3 Next generation sequencing (NGS) marker development method

The continued demand for microsatellite loci has made NGS a better alternative for marker development. NGS techniques serve as more efficient and effective tools as they avoid most of the laborious stages of microsatellite marker development process (Allentoft et al., 2009; Castoe et al., 2012; Mikheyev et al., 2010). This is because traditional marker development methods relied on construction of genomic libraries that were enriched for certain SSRs (Curto et al., 2013). This introduced biases related to the use of probes in addition to high financial costs and time required. This is in contrast to (NGS) where these earlier procedures are replaced by sequencing of fragmented genomic DNA at a fraction of the cost and much less time coupled with more informative data that is generated (Curto et al., 2013). Figure 4 shows a summary of comparison between convectional marker development methods and NGS.

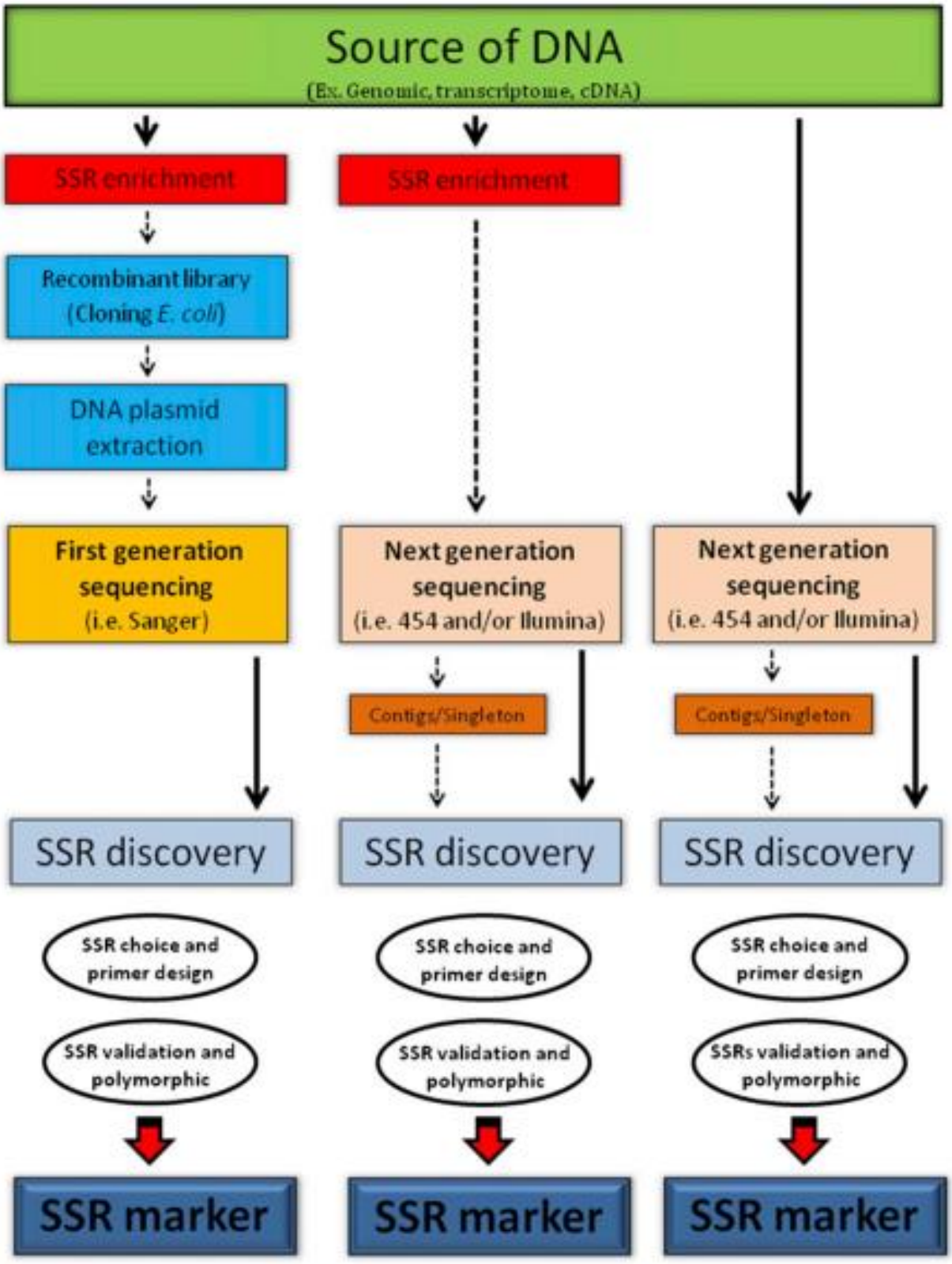


Figure 4. A flow chart summarizing the comparison between traditional marker development methods and NGS technology. Adapted from Zalapa et al. (2012).

Two main NGS technologies used in the discovery of SSRs are the 454 and Illumina sequencing. Roche's 454 pyrosequencing method utilizes fragmented nucleic acid template

of between 300 – 800 base pairs fitted with two different adaptor sequences at both ends. These are used as priming sites for emulsion Polymerase Chain Reaction (ePCR) and later sequencing reactions. However, Roche's 454 pyrosequencing method are really used these days (Margulies et al., 2005; Shendure and Ji, 2008).

Illumina technology on the other hand relies on bridge amplification of fragmented DNA in a PCR reaction. Four fluorescently labelled nucleotides are used to perform sequencing by synthesis whereby the four nucleotides are characterized by their reversible terminators which permit single base incorporation (Fedurco et al., 2006; Turcatti et al., 2008). The labelled nucleotides are identified by imaging in cyclic reactions (50 to 100 cycles) which results to 50 to 100 reads. Longer reads can be obtained although this increases the error rate (base substitution error) hence a limitation to this method (Shendure and Ji, 2008). Next generation sequencing technologies are revolutionary, reasonably priced, delivers fast, and gives precise genome information compared to conventional methods. Since they are PCR-based, they require only low quantities of template DNA (Nelson et al., 2010). Furthermore, NGS has advantage when it comes to genotyping. Conventional methods measured fragment lengths while NGS technology combines length and SNP information thereby giving results with high allelic information (De Barba et al., 2017). Additionally, NGS technologies are amenable to automation and reproducible, however, this technology is relatively new and not used by many people. (De Barba et al., 2017).

2.3 Cross amplification

Cross-species amplification, also referred to as transferability is the ability to use primer pairs developed from microsatellite loci of one species on other species of the same genus or different genera of the same family (Zucchi et al., 2002). As a result, cross-species amplification plays a key role in facilitating the application of microsatellites since it lowers the development costs especially when dealing with taxa possessing low microsatellite frequencies or in cases where microsatellite isolation is difficult (Primmer et al., 1996).

However, transferability depends on the existence of conserved microsatellite sequences (Barbara et al., 2007), which significantly reveal a high degree of conservation following many years of divergent evolution (Rico et al., 1996). Notably, the rate of success for cross-species amplification is directly correlated to the evolutionary distance between the species under

study (Barbara et al., 2007). This is principally because of conservation of flanking sequences across individuals of the same species and sometimes of different species (Selkoe and Toonen, 2006).

One major drawback that has been noted with cross-amplification is the high rate of mutation. Mutations occurring in primer regions means that single allele will be amplified in some individuals or will fail to amplify completely in others (Selkoe and Toonen, 2006).

2.4 Hybridization and introgression in Tilapia

Hybridization is the breeding of individuals from genetically discrete populations, their taxonomic status notwithstanding (Wasonga et al., 2017). “Hybridization” is more often used to refer to mating of individuals from different subspecies, and in some cases populations that are genetically different, although may not be differentiated taxonomically (Wasonga et al., 2017).

On the other hand, introgression refers to the gene flow within populations that have hybridized. It is fully realised when hybrids backcross to either of the parental populations. However, beyond F1 hybrids, the point at which an individual is no longer regarded as a hybrid but rather as a member of one of the parental populations that has undergone introgression is subjective (Wasonga et al., 2017).

Tilapia hybrids can be grouped into three categories: I) those created from species' introduction, in which case the introduced species hybridizes with a local species, or when two introduced species hybridize in a new environment (Pouyaud and Agnèse, 1995). II) Those created intentionally, such as in aquaculture settings to produce a determined male: female ratio 0or by disruption of the environment and III) those which are truly natural and occur mostly in the wild (Pouyaud and Agnèse, 1995).

2.5 Molecular markers and fish conservation

Several elusive aspects of mating systems can be elucidated using molecular markers as was demonstrated in the Gulf pipefish, *Syngnathus scovelli* (Jones and Avise, 1997). Moreover, markers such as microsatellites can be utilized to reveal Major Histocompatibility Complex (MHC) diversity in an attempt to determine genetic quality in animals (Charlesworth and Charlesworth, 1999). Apart from its major contribution in heterozygosity surveys, conservation genetics also assists in resolving disputed taxonomic groups, development of captive breeding schedules as well as in studying natural systems of breeding. Conservation genetics has also been shown to play an important role in assessing diversity in mating populations, gene flow management alongside determination of increased vigor contributors (Caro and Laurenson, 1994).

To this end, for conservation of species facing genetic threat such as the Nile tilapia, uncontrolled fish transfers, intentional cross breeding and other factors leading to erosion of native genes ought to be monitored. Subsequently, the use of multi-locus markers such as microsatellites as well as genetic conservation of the founder population need to be used to avert this loss of genetic diversity (Vrijenhoek, 1998).

CHAPTER 3

3.0 Materials and Methods

3.1 Fish samples

A total of 187 fish samples were obtained from the institute of Integrative Biology and Biodiversity Research laboratory, University of Natural Resources and Life Sciences, Vienna. These samples were collected earlier from Uganda by Tibihika, PhD student, while others were provided by the Institute of Hydrobiology and Water Management in the same university (Table 1). The samples were collected in the areas shown by red spots on the map of Africa (Figure 5)

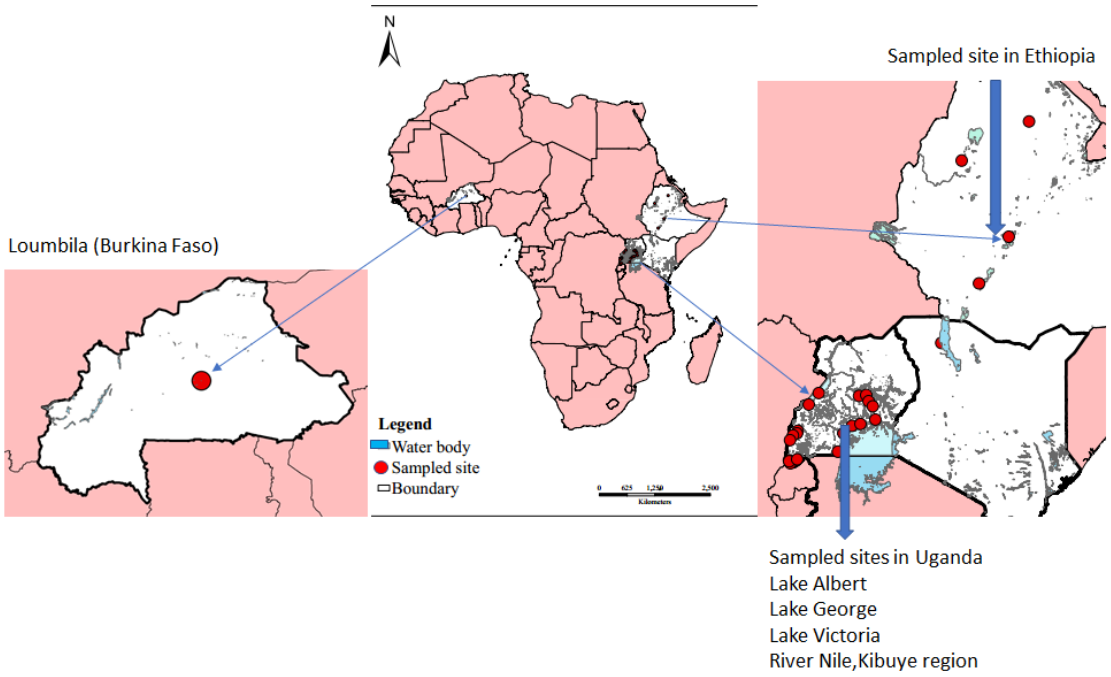


Figure 5. Map of Africa showing the different sampling sites in Uganda, Ethiopia and Burkina Faso. The red spots indicate the actual River or lake that was sampled. Though many water bodies were sampled, this study used samples from four of the sampling sites i.e., Lake Albert, Lake George, Lake Victoria and River Nile, Kibuye section in Uganda and Lake Ziway in Ethiopia.

Table 1. Shows source and number of the fish samples collected.

Species	Collected from	Number of samples collected
Oreochromis niloticus	Lake Albert (Uganda)	24
Oreochromis niloticus	Lake George (Uganda)	35
Oreochromis niloticus	River Nile (Uganda)	26
Oreochromis niloticus	Lake Victoria (Uganda)	25
Tilapia zillii	Lake Ziway (Ethiopia)	30
S. galilaeus	Loumbila (Burkina Faso)	30
Coptodon zillii	Loumbila (Burkina Faso)	17

3.2 DNA extraction

Genomic DNA was extracted from ethanol-preserved muscle tissues using magnetic beads method (MagSi-DNA beads). Extraction started with a single fish sampe from Nile tilapia obtained from Ethiopia which was used for SSR development and primer design. A small piece was cut using scalpel blade and forceps then placed on a 96 well microtiter plate, where 500 µl lysis buffer (2% SDS, 2% PVP 40, 250 mM NaCl, 200 mM Tris HCl, 5mM EDTA, pH 8.0) and an aliquot (20 µl) of proteinase K was added. The sample was then incubated overnight at 56 ^{0}C. The following day, 20 µl of RNase was added, followed by a further incubation for 30 minutes at 37 ^{0}C. A quick spin of 1 minute at 1000 rpm was then performed.

On a new microtiter plate, 15 µl of DNA magnetic beads was added followed by 500 µl of binding buffer (2 M GuHCl in 95% EtOH). Five hundred microliters of the sample was added to this new plate and properly mixed. The plate was then allowed to stand for 5 minutes after which it was placed over a magnetic separator (SL-MagSep96) for one minute. The supernatant was then removed and discarded. Washing was done twice using 600 µl wash buffer (10 mM Tris-HCl, pH 7.5 with 80% EtOH) and then dried for 10 minutes at room temperature to remove all traces of the wash buffer. Elution of DNA was done twice using 50 µl and 70 µl elution buffer(10 mM Tris-HCl, pH 7.5)to recover all the extracted DNA.

The quality of DNA was verified using 0.8% agarose gel (0.8 g agarose powder in 100 ml 1X TAE, pH 7.5). Five microliters of DNA sample was mixed with 15 µl of loading dye (Appendix 1) and loaded onto individual wells. The gel was run at 80 volts for 30 minutes and later visualized and documented using a trans illuminator system (Intas GEL IX IMAGER, Germany).

3.3 SSR development and primer design

Library preparation as well as illumina MiSeq sequencing of the extracted DNA was carried out at Genomics Service Unit, Ludwig-Maximilian-Universität München, Germany. Upon data receipt, FASTQC (Andrews, 2010) was used to determine the quality of the generated sequences while CUTADAPT version 0.11.1 (Martin, 2011) was used to trim adapter sequences and low quality regions i.e. those with less than 20 Phred value. Additionally, PEAR version 0.9.4 (Zhang et al., 2013) was applied in merging both the forward and reverse reads.

The cut off was set at minimum overlaps of 15 "BP" in length and those with a *p* value of less than 0.01 for the maximum observed or expected alignment scores. Next, SSR_pipeline (Miller et al., 2013) was used to screen for microsatellite motifs (from two to five repeat nucleotides) and only sequences possessing at least 6 repeats for 4 and 5mers, 8 repeats for 3mers and 10 repeats for 2mers were considered. Here, a total of 6,724 motif reads, made up of 4,629 2mers, 818 3mers 868 4mers and 409 5mers were obtained. For primer design, sequences equal to or greater than 350 "BP" in length as well as microsatellites flanked with regions longer than 30 bp were considered. The obtained raw reads were then submitted to Sequence Read Archive (SRA) database under the reference number SRX3398501.

The default Primer3 program (Untergasser et al., 2012) in Geneious software version 10.3 (Kearse et al., 2012) was used to design specific PCR primers. The conditions for manual Primer3 adjustments were: optimal primer melting temparature at 55 ^{0}C, GC content in the range of 20-50-80, optimal oligonucleotide length of between 18-20-23 bp and amplification product size of between 350-450 "BP". Consequently, a total of 47 primer pairs were initially designed.

Basic Local Alignment Search Tools for Nucleotides (BLASTN) algorithm was used for primer design using sequences previously aligned with the Nile tilapia genome available at the National Centre for Biotechnology Information (NCBI) database (Ref. MKQE01000000). BLASTN produces a list of pairwise alignment matches together with sequence hits above which a statistical threshold is presented (Xiong, 2006). In the present study, aligned sequences were considered based on the statistical power indicated by the E-value. The E-value is similar to probability since a lower E-value proposes less likelihood that the database matches arise from random chance but instead suggest that the database matches show a

significant similarity (Xiong, 2006). In this study, only primers obtained from sequences showing single matches were taken into consideration since they were more likely to represent regions with a single copy. Out of the 47 primer sequences previously mapped on the genome, only 42 pairs displayed satisfactory alignments.

3.4 SSR PCR single-plex for primer transferability and cross-amplification test

For transferablity test, DNA was extracted following the same procedure as for the *O. niloticus* above (section 3.2). One hundred and ten Nile tilapia samples from Uganda, 30 samples of *Tilapia zillii* from Ethiopia, 17 samples of *Coptodon zillii* from Burkina Faso and 30 samples of *S. galilaeus* also from Burkina Faso were assayed in an attempt to determine transferability and cross-amplification ability for the candidate loci following PCR amplification reactions.

These reactions were conducted in a 10 µl total volume. First, QIAGEN Multiplex PCR Master Mix kit (QIAGEN, CA, USA) was used to test all primer pairs. PCR reaction volume during DNA amplification consisted of 1 µl genomic DNA template, 5 µl master mix and 4 µl primer mix. Primer mix composed of: 1 µl of 100 µM reverse primer, 1 µl of 100 µM forward primer and 98 µl water. Thermocycler reaction conditions were performed as follows: initialization step at 95 ^{0}C for 15 minutes, followed by 35 cycles of denaturation at 95 ^{0}C for 30 seconds, annealing at 55 ^{0}C for 60 seconds, elongation at 72 ^{0}C for 60 seconds and final primer extension at 72 ^{0}C for 10 minutes.

The quality of amplified PCR products was then analysed by electrophoresis using 1.8% agarose gel (1.8 g agarose powder in 100 ml 1X TAE, pH 7.5). From this, primer pairs were successfully identified and later used for multiplex PCR approach on the test populations. Conversely, PCR-singleplex gel products after two replications were used to determine successful markers for cross-species amplification.

3.5 Multiplex PCR and SSR library preparation

For illumina sequencing, two illumina adapters P5 and P7 were used. Two PCR steps were initiated by first using marker-specific primers combined with a linker at the 5'end which served as a template for the second PCR step in the presence of primers containing all the necessary components for illumina sequencing. The marker –specific primers sequences were extended with part of the P5 and P7 Illumina adaptors (forward: TCT TTC CCT ACA CGA CGC

TCT TCC GAT CT and Reverse: CTG GAG TTC AGA CGT GTG CTC TTC CGA TCT). In the second PCR, for each sample, a novel combination of two different indices was added using the following primers: P5 (AAT GAT ACG GCG ACC ACC GAG ATC TAC AC [index] ACA CTC TTT CCC TAC ACG ACG and P7 (CAA GCA GAA GAC GGC ATA CGA GAT [index] GTG ACT GGA GTT CAG ACG TGT) allowing for pooling of large sample size for the downstream analysis. A multiplex resulting from a combination of all gel-analysed primers was set as the first PCR reaction. The reaction was performed in a 10 µl total volume consisting of 2 µl genomic DNA, 5 µl master mix, 2.5 µl water in addition to 0.5 µl primer mix ((1 µM final concentration). Thermal cycler conditions were similar to those previously described for single-plex PCR.

Agencourt AMPure XP PCR purification protocol (with slight modifications) was used for purification of PCR products using magnetic beads. An aliquot (4 µl) of PCR products was mixed with 2.86 µl of AMPure XP beads (Beckman Coulter, Inc, Bree, CA) using micropipette tips and the mixture incubated for 5 minutes at room temparature. An inverted magnetic bead extraction device, VP 407-AM-N (V&P SCIENTIFIC, INC) was used to capture bound DNA beads which were then washed twice using 200 µl of 80% ethanol for about 45 seconds. The beads were later dried and then eluted using 17 µl of elution buffer (10 mM Tris, pH 8) following drying for 5 minutes at room temperature.

The second PCR was carried out in a total reaction volume of 10 µl. This consisted of 1 µl template PCR products, 5 µl master mix and 2 µl each of index primer (1 µM) P5 and P7 . Thermal cycler profile for this reaction: heat lid to 110 °C, initial denaturation step at 95 °C for 15 minutes followed by 10 cycles of denaturation at 95 °C for 30 seconds, annealing at 58 °C for 1 minute, elongation at 72 °C for 1 minute, and final primer extension at 72 °C for 5 minutes. All the PCR products were then pooled prior to sending for paired-end 300 “BP” sequencing on an illumina MiSeq platform at the Genomics Service Unit at Ludwig Maximillian Universität, München, Germany.

3.6 Sequence analysis and SSR genotyping

Reads obtained from illumina sequencing were quality controlled and merged as earlier described for SSR discovery. Generally, the resultant sequences begin with forward primer and end with reverse primer sequences. De-multiplexing the sequences based on the primer content was applied leading to formation of a single file per sample per locus by an in-house

python script. Additionally, sequence length and whole sequence information (haplotype allele calling) methods were employed for allele calling. The two methods were then compared based on PIC, number of alleles per locus and the genetic structure each method produced.

To call alleles based on sequence length, awk commands in a bash script (Appendix 2) was used to calculate the length distribution per sample per locus. Automatic calling of the alleles was done when the length distribution dispalyed a clear pattern consistent with either the homozygote or heterozygote genotype using R script developed by Manuel and Silvia, from the Institute of Integrative Biology and Biodiversity Research laboratory, University of Natural Resources and Life Sciences, Vienna.

The threshold of relative length abundance of the read for calling heterozygote and homozygote was user defined with the default value set at 0.62/sample/marker. R script display histograms highlighting the called alleles in addition to a matrix with the related allele lengths (Appendix 2). However, manual control was applied for allele that displayed a more complex pattern.

The entire fragments with similar lengths were extracted and a 70% consensus sequence generated for Whole Sequence Information (WSI) allele calling method. Therefore, nucleotide positions that had the most common nucleotide with a frequency of less than 70% were considered potentially heterozygous for a given SNP. Often, the original sequences were recovered and the two most frequent nucleotides used to determine which nucleotides were present for these positions. If more than one SNP per marker was observed, an assessment of which nucleotides were linked was conducted.

Moreover, polymorphic positions where both nucleotides presented in homozygote state were considered. Nevertheless, this was only possible through aligning the consensus sequences. Due to the complexity of the repetitive motif alignment, this was done manually. An in-house python script was used to recall all alleles after calling ambiguous SNPs and this relied on the sequence content. Sequence analyses revealed 42 loci expressed in SSR matrices for additional statistical evaluation.

3.7 Statistical analysis

A number of programs were applied to determine descriptive statistical analyses for SSR loci. The Cervus software version 3.0.7 (Kalinowski et al., 2007) was used to assess observed heterozygosity (Ho), expected heterozygosity (He) as well as loci polymorphic information content. Fstat program version 2.9.3.3 (Goudet, 2001) was employed to determine the number of alleles per locus. Additionally, the genetic structure and Principal Coordinate Analysis (PCoA) were evaluated using STRUCTURE version 2.3.4 (Hubisz et al., 2009) and GenAlex version 6.5 (Peakall and Smouse, 2006), respectively. This was intended to assess the level of informativeness and thus usability of the developed primers.

In this study, STRUCTURE was set at 100, 000 burn-in period and Markov Chain Monte Carlo (MCMC) model run at 100,000 generations for 10 independent iterations. This was done from K = 1 to K= 10 The default STRUCTURE settings for admixture model and allele frequencies correlated were applied. STRUCTURE HARVESTER software obtained from http://taylor0.biology.ucla.edu/structureHarvester/ (Earl, 2012) was used to infer the best K that fits the data. In addition, the CLUMPAK clustering Markov package pipeline was run across the K values to summarise the several independent iterations in a single Q matrix (Kopelman et al., 2015).

CHAPTER 4

4.0 RESULTS

4.1 Assembly and characterization of microsatellite markers

From a total of 2,404,293 paired reads generated by illumina sequencing platform for microsatellite development, 6,724 contained SSR motifs and hence were assembled for further quality check. Di and tetra repeats were the majority at 4,629 and 868 followed by tri and Penta repeats at 818 and 409 respectively. From these reads, 47 markers were initially developed. Upon further tests, 5 failed and were eliminated (Appendix 3), leaving a total of 42 that were used for further analysis (Table 7).

4.2 Cross-amplification and transferability of the developed SSR primers

4.2.1 Transferability to Nile tilapia

A total of 624 alleles were identified in whole sequence information, here after referred to as haplotype, with a mean of 14.86 alleles per locus, while 378 alleles were identified in sequence length calling method, with a mean of 9 alleles per locus (Table 2). Observed (Ho) and expected heterozygosity (He) ranged from 0.02 to 0.97 (mean = 0.49) and 0.07 to 0.77 (mean = 0.60) respectively (Table 3). Mean polymorphic information content (PIC) was 0.58 in haplotype method and 0.48 in sequence length allele calling (Table 2). Allelic numbers ranged from 2 to 56 (Table 3).

All the 42 developed SSRs were transferable to the Ugandan Nile tilapia, with two of them, (TI39 and TI49) showing 100% transferability success (Appendix 4), i.e., amplifying in all the test species. Nine SSRs (TI12, TI17, TI27, TI50, TI51, TI52, TI57 TI59 and TI60) showed very strong transferability success. They amplified in 92% of the test samples (Appendix 4).

Table 2. Showing a summary of comparison between the two allele calling methods: haplotype and sequence length. PIC = Polymorphic information content from haplotype allele calling, PIC* = Polymorphic information content from sequence length allele calling

	Haplotype		Sequence length	
Species	Number of alleles	Mean PIC	Number of Alleles	Mean PIC*
S. galilaeus	4.24	0.44	3.4	0.32
T. zillii	2.34	0.27	2.02	0.22
C. zillii	3.36	0.39	2.76	0.33
O. niloticus	14.86	0.58	9.0	0.48

Table 3. Showing transferability results for Oreochromis niloticus. Na = Number of alleles in haplotype allele calling, Na*= Number of alleles in sequence length allele calling, Ho = Observed heterozygosity, He = Expected heterozygosity, PIC = Polymorphic information

	Oreochromis niloticus					
Locus	**Na**	**Na***	**Ho**	**He**	**PIC**	**PIC***
TI49_TGT	14	5	0.33	0.32	0.31	0.16
TI39_ATGG	12	6	0.47	0.48	0.44	0.38
TI52_TAT	7	4	0.37	0.69	0.64	0.36
TI55_TCTA	13	13	0.80	0.87	0.86	0.86
TI15_TGC	6	4	0.07	0.10	0.10	0.07
TI8_AC	7	6	0.13	0.16	0.16	0.07
TI32_AAAAT	31	6	0.88	0.90	0.89	0.34
TI16_AAC	8	8	0.13	0.16	0.16	0.16
TI43_GAATA	22	13	0.34	0.45	0.45	0.39
TI34_TCTCT	26	12	0.84	0.91	0.90	0.86
TI27_TTTG	2	1	0.03	0.03	0.03	0.00
TI17_GAA	17	10	0.62	0.74	0.70	0.69
TI6_GA	25	16	0.97	0.78	0.75	0.68
TI14_TAA	6	3	0.43	0.58	0.54	0.35
TI61_TGGA	7	3	0.04	0.07	0.06	0.03
TI51_TGT	18	7	0.58	0.63	0.60	0.47
TI1_TG	8	5	0.48	0.52	0.46	0.44
TI2_CA	22	15	0.75	0.80	0.77	0.76
TI57_TCCA	5	2	0.02	0.07	0.07	0.02
TI29_TAAAA	12	9	0.20	0.60	0.52	0.19
TI7_AC	14	12	0.92	0.76	0.73	0.72
TI18_ATCT	33	16	0.89	0.93	0.92	0.88
TI44_GAAAA	10	7	0.63	0.77	0.73	0.64
TI41_AAAC	18	14	0.60	0.83	0.81	0.79
TI35_AAAAG	25	14	0.53	0.91	0.90	0.80
TI59_AGGA	8	5	0.68	0.69	0.63	0.59
TI12_TAC	10	5	0.31	0.44	0.39	0.06
TI60_ATCC	8	5	0.60	0.70	0.65	0.58
TI5_CA	9	7	0.50	0.56	0.52	0.25
TI56_TGTT	13	4	0.13	0.58	0.53	0.12
TI50_ATGG	8	7	0.50	0.60	0.55	0.54
TI13_ATG	5	5	0.27	0.40	0.38	0.38
TI28_ATTCA	2	1	0.26	0.30	0.26	0.00
TI4_GT	11	9	0.82	0.81	0.78	0.77
TI26_ACAA	10	10	0.56	0.68	0.64	0.64
TI33_TTCAA	9	8	0.75	0.85	0.82	0.82
TI24_TTAC	32	13	0.78	0.85	0.83	0.78
TI54_GGAT	24	10	0.80	0.90	0.88	0.83
TI22_CTAT	56	39	0.58	0.95	0.94	0.93
TI31_CTAAT	13	9	0.20	0.79	0.75	0.73
TI9_AC	8	4	0.27	0.32	0.30	0.13
TI53_ATAG	30	26	0.43	0.94	0.92	0.92
Mean	14.86	9	0.49	0.60	0.58	0.48

content in haplotype allele calling, PIC* = Polymorphic information content in sequence length allele calling

4.2.2 Cross-amplification

Cross-amplification was also successful in most of the developed SSRs primers. In *S. galilaeus*, twenty-seven showed over 60% amplification rate but in total 32 out of 42 amplified in all the test samples (Appendix 5). Mean observed (Ho) and expected (Ho) heterozygosity were 0.30 and 0.45 respectively (Table 4). Mean number of alleles and PIC values were 4.24 and 0.44 respectively in haplotype method (Table 4), while sequence length method gave a mean PIC value of 0.32 and mean number of alleles of 3.40 per locus (Table 4).

In *Tilapia* zillii, haplotype method identified a total of 102 alleles (mean 2.43), and mean PIC of 0.27 (Table 5) while sequence length method identified a total of 85 alleles, with a mean of 2.02 per locus, and mean PIC of 0.22 (Table 5). Additionally, 25 of the 42 developed primers cross-amplified while 17 failed (Appendix 6). Observed and expected heterozygosity were 0.32 and 0.36 respectively (Table 5).

Similarly, in *Coptodon zillii* 18 primers showed 80% amplification rate (Appendix 7) but in total 23 out of the 42 SSRs primers amplified in all the test samples. In haplotype method, mean allele and PIC values were 3.36 and 0.39 (Table 2), while mean observed and expected heterozygosity were 0.36 and 0.47 respectively (Table 6).

Table 4. Showing cross-amplification results in S. galilaeus. Na = Number of alleles in haplotype allele calling, Na*= Number of alleles in sequence length allele calling, Ho = Observed heterozygosity, He = Expected heterozygosity, PIC = Polymorphic information content in haplotype allele calling, PIC* = Polymorphic information content in sequence length allele calling

	Sarotherodon galilaeus					
Locus	**Na**	**Na***	**Ho**	**He**	**PIC**	**PIC***
TI49_TGT	3	2	0.04	0.20	0.18	0.17
TI39_ATGG	5	4	0.68	0.61	0.56	0.32
TI52_TAT	4	4	0.28	0.33	0.31	0.31
TI55_TCTA	18	13	0.75	0.87	0.84	0.77
TI15_TGC	4	2	0.46	0.44	0.41	0.20
TI8_AC	7	5	0.61	0.56	0.49	0.45
TI32_AAAAT	7	6	0.63	0.73	0.69	0.67
TI16_AAC	3	3	0.38	0.57	0.46	0.46
TI43_GAATA	6	6	0.52	0.59	0.54	0.54
TI34_TCTCT	8	6	0.70	0.68	0.61	0.52
TI27_TTTG	4	2	0.48	0.65	0.57	0.07
TI17_GAA	5	3	0.76	0.76	0.70	0.46
TI6_GA	4	4	0.21	0.55	0.49	0.49
TI14_TAA	5	3	0.48	0.73	0.66	0.41
TI61_TGGA	5	3	0.31	0.69	0.63	0.40
TI51_TGT	4	4	0.24	0.70	0.62	0.62
TI1_TG	2	2	0.04	0.04	0.04	0.04
TI2_CA	8	5	0.50	0.80	0.76	0.60
TI57_TCCA	3	1	0.52	0.56	0.45	0.00
TI29_TAAAA	4	4	0.40	0.63	0.55	0.55
TI7_AC	3	3	0.12	0.26	0.24	0.24
TI18_ATCT	6	5	0.18	0.68	0.62	0.62
TI44_GAAAA	9	9	0.25	0.87	0.81	0.81
TI41_AAAC	4	3	0.19	0.42	0.38	0.37
TI35_AAAAG	11	11	0.79	0.86	0.83	0.83
TI59_AGGA	4	4	0.11	0.24	0.22	0.22
TI12_TAC	6	3	0.50	0.64	0.59	0.44
TI60_ATCC	6	5	0.26	0.33	0.32	0.26
TI5_CA	4	4	0.11	0.18	0.17	0.17
TI56_TGTT	3	3	0.23	0.35	0.31	0.31
TI50_ATGG	3	2	0.28	0.45	0.39	0.32
TI13_ATG	3	3	0.54	0.60	0.52	0.52
TI28_ATTCA	2	1	0.00	0.67	0.38	0.00
TI4_GT	2	2	0.00	0.53	0.37	0.37
TI26_ACAA	1	1	0.00	0.00	0.00	0.00
TI33_TTCAA	0	0	0.00	0.00	0.00	0.00
TI24_TTAC	0	0	0.00	0.00	0.00	0.00
TI54_GGAT	0	0	0.00	0.00	0.00	0.00
TI22_CTAT	0	0	0.00	0.00	0.00	0.00
TI31_CTAAT	1	1	0.00	0.00	0.00	0.00
TI9_AC	0	0	0.00	0.00	0.00	0.00
TI53_ATAG	1	1	0.00	0.00	0.00	0.00
Mean	4.24	3.40	0.30	0.45	0.44	0.32

Table 5. Showing cross-amplification results in *T. zillii*. Na = Number of alleles in haplotype allele calling, Na*= Number of alleles in sequence length allele calling, Ho = Observed heterozygosity, He = Expected heterozygosity, PIC = Polymorphic information content in haplotype allele calling, PIC* = Polymorphic information content in sequence length allele calling.

	Tilapia zillii					
Locus	**Na**	**Na***	**Ho**	**He**	**PIC**	**PIC***
TI49_TGT	2	1	0.17	0.16	0.14	0.00
TI39_ATGG	2	2	0.33	0.28	0.24	0.24
TI52_TAT	2	2	0.50	0.38	0.31	0.31
TI55_TCTA	2	1	0.30	0.26	0.22	0.00
TI15_TGC	4	4	0.03	0.43	0.39	0.39
TI8_AC	5	4	1.00	0.77	0.71	0.57
TI32_AAAAT	5	5	1.00	0.73	0.67	0.67
TI16_AAC	2	2	0.03	0.03	0.03	0.03
TI43_GAATA	4	4	0.71	0.68	0.59	0.59
TI34_TCTCT	4	4	0.70	0.66	0.58	0.58
TI27_TTTG	3	3	0.48	0.46	0.37	0.37
TI17_GAA	2	2	0.00	0.23	0.20	0.20
TI6_GA	3	2	0.46	0.44	0.37	0.07
TI14_TAA	2	1	0.04	0.04	0.04	0.00
TI61_TGGA	3	2	0.21	0.19	0.18	0.17
TI51_TGT	2	2	0.04	0.04	0.04	0.04
TI1_TG	4	2	0.50	0.56	0.47	0.36
TI2_CA	2	2	0.04	0.04	0.04	0.04
TI57_TCCA	5	4	0.35	0.65	0.60	0.57
TI29_TAAAA	2	2	0.04	0.04	0.04	0.04
TI7_AC	2	1	0.07	0.13	0.12	0.00
TI18_ATCT	3	1	0.60	0.44	0.35	0.00
TI44_GAAAA	4	3	0.11	0.11	0.11	0.10
TI41_AAAC	2	2	0.17	0.16	0.15	0.15
TI35_AAAAG	2	2	0.04	0.19	0.17	0.17
TI59_AGGA	5	4	0.60	0.82	0.70	0.61
TI12_TAC	2	1	0.50	0.50	0.31	0.00
TI60_ATCC	3	3	0.50	0.83	0.56	0.56
TI5_CA	1	1	0.00	0.00	0.00	0.00
TI56_TGTT	2	1	1.00	1.00	0.38	0.00
TI50_ATGG	2	2	1.00	1.00	0.38	0.38
TI13_ATG	0	0	0.00	0.00	0.00	0.00
TI28_ATTCA	3	2	0.17	0.32	0.27	0.24
TI4_GT	2	2	0.00	0.43	0.31	0.31
TI26_ACAA	0	0	0.00	0.00	0.00	0.00
TI33_TTCAA	2	2	1.00	1.00	0.38	0.38
TI24_TTAC	0	0	0.00	0.00	0.00	0.00
TI54_GGAT	0	0	0.00	0.00	0.00	0.00
TI22_CTAT	3	3	0.67	0.80	0.59	0.59
TI31_CTAAT	3	3	0.25	0.46	0.37	0.37
TI9_AC	0	0	0.00	0.00	0.00	0.00
TI53_ATAG	1	1	0.00	0.00	0.00	0.00
Mean	2.43	2.02	0.32	0.36	0.27	0.22

Table 6. Showing cross-amplification results in *Coptodon zillii*. Na = Number of alleles in haplotype allele calling, Na*= Number of alleles in sequence length allele calling, Ho = Observed heterozygosity, He = Expected heterozygosity, PIC = Polymorphic information content in haplotype allele calling, PIC* = Polymorphic information content in sequence length allele calling

	Coptodon zillii					
Locus	**Na**	**Na***	**Ho**	**He**	**PIC**	**PIC***
TI49_TGT	4	4	0.40	0.53	0.48	0.48
TI39_ATGG	7	7	0.80	0.73	0.67	0.67
TI52_TAT	2	2	0.00	0.14	0.12	0.12
TI55_TCTA	2	1	0.21	0.20	0.17	0.00
TI15_TGC	4	3	0.21	0.55	0.49	0.44
TI8_AC	5	4	0.14	0.75	0.69	0.66
TI32_AAAAT	9	7	1.00	0.86	0.81	0.78
TI16_AAC	5	5	0.27	0.41	0.38	0.38
TI43_GAATA	9	5	0.43	0.87	0.82	0.57
TI34_TCTCT	8	8	0.80	0.85	0.80	0.80
TI27_TTTG	2	2	0.07	0.19	0.16	0.16
TI17_GAA	5	3	0.73	0.74	0.67	0.44
TI6_GA	2	2	0.07	0.30	0.25	0.25
TI14_TAA	6	3	0.43	0.76	0.70	0.46
TI61_TGGA	5	2	0.55	0.69	0.61	0.38
TI51_TGT	4	3	0.50	0.66	0.58	0.41
TI1_TG	4	4	0.71	0.56	0.48	0.48
TI2_CA	3	3	0.47	0.45	0.37	0.37
TI57_TCCA	2	2	0.00	0.53	0.37	0.37
TI29_TAAAA	5	4	0.46	0.74	0.66	0.63
TI7_AC	1	1	0.00	0.00	0.00	0.00
TI18_ATCT	1	1	0.00	0.00	0.00	0.00
TI44_GAAAA	5	4	0.29	0.79	0.70	0.64
TI41_AAAC	5	3	0.64	0.76	0.69	0.36
TI35_AAAAG	5	5	0.40	0.70	0.62	0.62
TI59_AGGA	2	2	0.33	0.33	0.24	0.24
TI12_TAC	1	1	0.00	0.00	0.00	0.00
TI60_ATCC	2	2	0.67	0.53	0.35	0.35
TI5_CA	2	2	0.00	0.43	0.31	0.31
TI56_TGTT	3	1	0.67	0.60	0.45	0.00
TI50_ATGG	1	1	0.00	0.00	0.00	0.00
TI13_ATG	1	1	0.00	0.00	0.00	0.00
TI28_ATTCA	2	2	0.00	0.57	0.38	0.38
TI4_GT	4	3	1.00	1.00	0.70	0.56
TI26_ACAA	1	1	0.00	0.00	0.00	0.00
TI33_TTCAA	4	4	1.00	0.80	0.62	0.62
TI24_TTAC	2	2	0.50	0.50	0.31	0.31
TI54_GGAT	1	1	0.00	0.00	0.00	0.00
TI22_CTAT	3	3	0.33	0.60	0.45	0.45
TI31_CTAAT	2	2	1.00	0.60	0.38	0.38
TI9_AC	0	0	0.00	0.00	0.00	0.00
TI53_ATAG	0	0	0.00	0.00	0.00	0.00
Mean	3.36	2.76	0.36	0.47	0.39	0.33

4.3 Characteristics of the developed microsatellite markers

Forty-two markers were developed in this study. They consisted of 8 dinucleotides, 9 trinucleotides, 16 tetra nucleotides and 9 Penta nucleotides (Table 7). The highest repeat units were found in dinucleotides at 23 repeats of GA, followed by 21 repeats of trinucleotides. On the other hand, the lowest repeats were found in tetra nucleotides and Penta nucleotides at 6 repeats each (Table 7).

Table 7. Showing sequences, repeat units and size lengths for the 42 polymorphic microsatellite loci developed for *Oreochromis niloticus*

Name	Primer sequence (5'- 3')	Repeat motif	Size (bp)
TI1_TG	F: TTATCACTGCTGAACGTCTT	$(TG)_{10}$	279-298
	R: GTTTTGGCTGCTACACATTC		
TI2_CA	F: TTCTGGGCTAACACACAAG	$(CA)_{18}$	220-255
	R: AAGGTGTCACACAGTTTAGG		
TI4_GT	F: TGTGCAGAATAGAATAGCCC	$(GT)_{18}$	108-143
	R: GAAAGGAAAAATGTTGGTGGT		
TI5_CA	F: AAGGAGGATGATCAGGACAC	$(CA)_{10}$	67-86
	R: AGACCTCCACTGTGATCTTA		
TI6_GA	F: CAGCTCTCATGAACACTTGA	$(GA)_{23}$	94-139
	R: ACCCATAAATCACACCAGTC		
TI7_AC	F: TCTTTGTGTCAGAACTGTGT	$(AC)_{17}$	296-329
	R: ACTCTGCTTTTAGCCAATCA		
TI8_AC	F: CTGAAGTCCTGCTGAGATTT	$(AC)_{15}$	180-209
	R: CATTGTTCTTGGCACCTCTA		
TI9_AC	F: CTCAGTGACGAAGCCAAA	$(AC)_{10}$	71-90
	R: CCTGGCAATCAAAAGAACAA		
TI12_TAC	F: GCCACCAAAATATTCGTGTT	$(TAC)_{12}$	104-139
	R: CCATGTTCTGTCTCCTTGAA		
TI13_ATG	F: AATCCGTTAGCTGCAGATAG	$(ATG)_{10}$	138-167
	R: GCTGATTAAACACAAAGTTGG		
TI14_TAA	F: TCCCTAAAATATGCCACCAA	$(TAA)_{19}$	283-339
	R: TAGTGCTTTAATGGCTCTGG		
TI15_TGC	F: GCTGTGATCATCTGGAGAAA	$(TGC)_{10}$	310-339
	R: AGGATCTAGAACCTCCAACC		
TI16_AAC	F: CAGACGTAGGCGATAAATCT	$(AAC)_{10}$	165-194
	R: GAACACATCCATTTCCACAC		
TI17_GAA	F: AACTGAAGAAGAAGCCTTGG	$(GAA)_{21}$	62-124
	R: ATCATCTTCCTCTACTGCCT		
TI18_ATCT	F: AGCAAGTGAGATAAGCACTG	$(ATCT)_{8}$	201-232
	R: TACATAGCAGTGCAGTTTGC		
TI22_CTAT	F: ACTGACCAAGTGCTTTGTAT	$(CTAT)_{20}$	81-160
	R: AACTGCTGTGTTGAACTTTG		
TI24_TTAC	F: ACTGACAACATAAAGACATATGC	$(TTAC)_{9}$	115-150

	R: CACAGTTTGAATCCACCATC		
TI26_ACAA	F: ATTGCTTCATCCCTTGAGTT	$(ACAA)_{10}$	114-153
	R: ACACGGAAAACCTAATGACA		
TI27_TTTG	F: CTGTCTTTCTTGATGTGGGA	$(TTTG)_6$	131-154
	R: ATGCACAAATTTTAAGGGCC		
TI28_ATTCA	F: TGTCTTGGGATTTGAGATCA	$(ATTCA)_8$	182-221
	R: CGGAGGTTTCTTCCTGTTAA		
TI29_TAAAA	F: AGGTCAAGATCAAGCAGTTT	$(TAAAA)_{10}$	18-,233
	R: CATCAACATAATTCAGTGTGGA		
TI31_CTAAT	F: GAAACTATCCACAGAAGCCA	$(CTAAT)_7$	328-362
	R: AGGCTTCTTACAGTTGGATG		
TI32_AAAAT	F: CAGGAAATGGCTCCAAAATG	$(AAAAT)_7$	202-236
	R: TTGTAGCTAGGAATCAGTGC		
TI33_TTCAA	F: GCTTATGGCTGTATGGAGTT	$(TTCAA)_6$	328-357
	R: CGACTTCTGTTGTGATTTGG		
TI34_TCTCT	F: GCTTACAGTACATTGTGTGC	$(TCTCT)_{10}$	296-345
	R: CTGATGAGAAAAACAGACGC		
TI35_AAAAG	F: TCAACCACAAACTCCTCTTT	$(AAAAG)_{14}$	116-185
	R: AAACTAAGTGCAGCTCATGA		
TI39_ATGG	F: TACCTGCCAGTCATGTGCTG	$(ATGG)_8$	329-360
	R: TGCTCAGACTGGTCCCTTCT		
TI141_AAAC	F: TCGCAGCTGCTCCTGTTTAA	$(AAAC)_{11}$	102-145
	R: TTGTGCACGTGGACATGTTG		
TI43_GAATA	F: ATTGCCATCACCAGGAACCA	$(GAATA)_6$	160-189
	R: TGCTAGCCCAGAGCATTTGA		
TI44_GAAAA	F: TGCTCCTGACTCAGCATCAC	$(GAAAA)_6$	201-230
	R: GCAGCACTCTGACATGAAGC		
TI49_TGT	F: TCGAAGTAGCGTGGAAAACCT	$(TGT)_8$	311-334
	R: ACAACAACAACAGGTCGGGA		
TI50_ATGG	F: CCTGTGACAGACTGGTGACC	$(ATGG)_7$	62-89
	R: ACACTGATGCGGTTTACGGT		
TI51_TGT	F: TGCTAAACGCCAGCTGATGA	$(TGT)_8$	72-95
	R: TTACCACACGATGTCGCAGG		
TI52_TAT	F: GAGAAACGTCCAGTGGCAGA	$(TAT)_8$	193-216
	R: TTTCGATCTGCTGCCCCTTT		
TI53_ATAG	F: ATGAGCCAGCGTTGAGTCAA	$(ATAG)_8$	271-302
	R: TTCCGAACACCTTGGTGTCC		
TI54_GGAT	F: TTTCTTGCCAGCAAAAACAGT	$(GGAT)_7$	288-315
	R: CAGATTCTTCCAGTGCTTGTGC		
TI55_TCTA	F: GAGCCCAGACAGCAGACAAT	$(TCTA)_7$	323-350
	R: AGGACCTTCTATGGCCCTGT		
TI56_TGTT	F: TGCAGTGAATTTGGCACCTG	$(TGTT)_6$	326-349)
	R: AGCCTGAGATACCTGTGCCT		
TI57_TCCA	F: CAGTGGGAGGAAGCTCCAAA	$(TCCA)_7$	136-163
	R: GCTGCATGGATCCAATAGGC		
TI59_AGGA	F: ATGGACTTAAGCTGCACCCC	$(AGGA)_6$	220-243
	R: TGAGCATTTGACCCCAGCAT		

Microsatellite Cross-species Amplification and Utility in African Cichlids:

TI60_ATCC	F: GAGCCGCCATAGTGTCACTT R: CCTGCTCTCACTCAAAGAGGG	$(ATCC)_7$	115-142
TI61_TGGA	F: GCTACACAGGAAAGCAGAGC R: ACTCAATGCTGGACGTGACC	$(TGGA)_6$	309-332

4.4 Diversity of the developed microsatellite markers

Diversity of the developed SSR primers was tested in terms of PIC, number of alleles per locus, and observed and expected heterozygosity. This was done for all the species and then for motif per species. Nile tilapia had the highest variation in PIC and in number of alleles as shown in (Figure 6a). There was an increase in PIC with increase in motif length (Figure 6b). Penta nucleotides (5 repeats) were more informative than other repeats in Nile tilapia and *Tilapia zillii* populations, the least informative being the trinucleotides, (3 repeats). This was not the case in *S. galilaeus* which had trinucleotide being the second most informative after Penta nucleotide.

Mean expected heterozygosity was higher than mean observed heterozygosity in all the populations except in *zillii* when both populations were combined (Cz_Zi), (Figure 6c). However, on splitting the two populations, both showed a higher mean expected heterozygosity than mean observed (Figure 6c). Nile tilapia had higher diversity than *S. galilaeus* and *Tilapia zillii*. River Nile populations had almost equal observed and expected heterozygosity.

a) Number of alleles and PIC vs Species

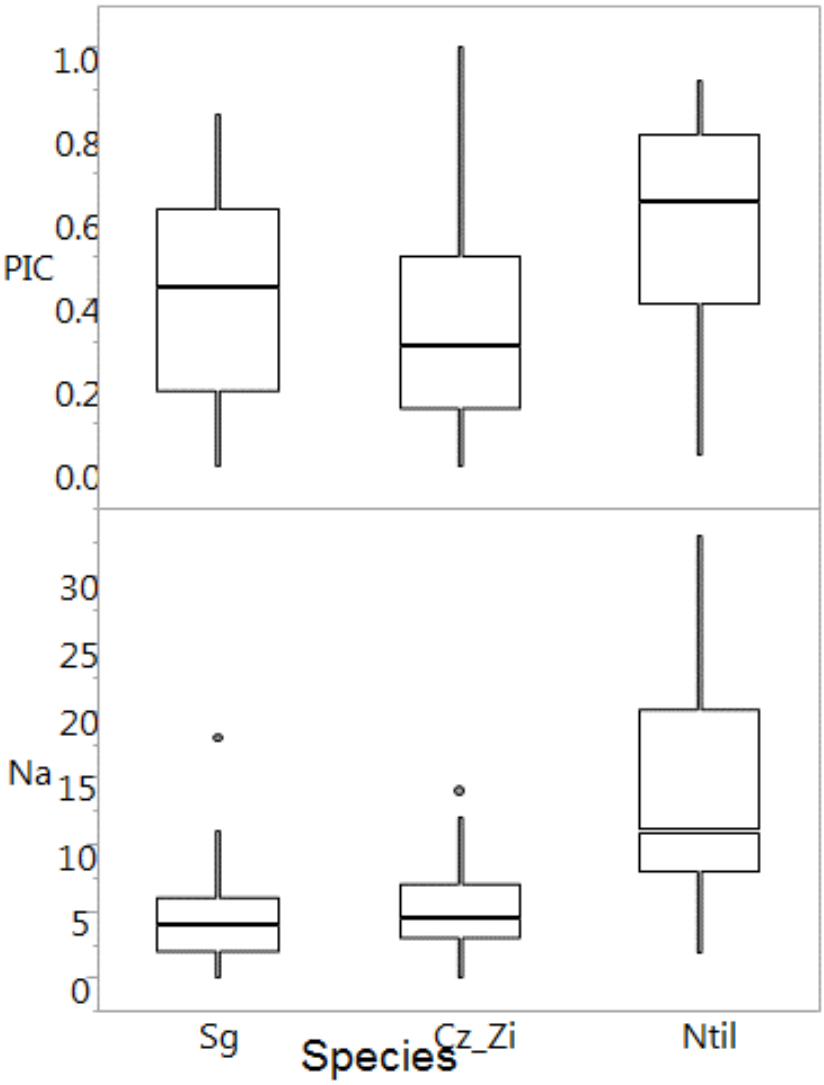


Figure 6a. Boxplots comparing number of alleles and PIC among the species understudy considering haplotype genotyping only. Nile tilapia had the highest number of alleles while *S. galilaeus* had the least. Marker (TI22_CTAT) with 56 alleles in Nile tilapia was excluded as an outlier. The greater the number of alleles, the greater the PIC. Na = Number of alleles, PIC = Polymorphic information content, Sg = *Sarotherodon galilaeus*, Cz_Zi = combined *Tilapia zillii*, Ntil = Nile tilapia

b) Motif vs PIC per species

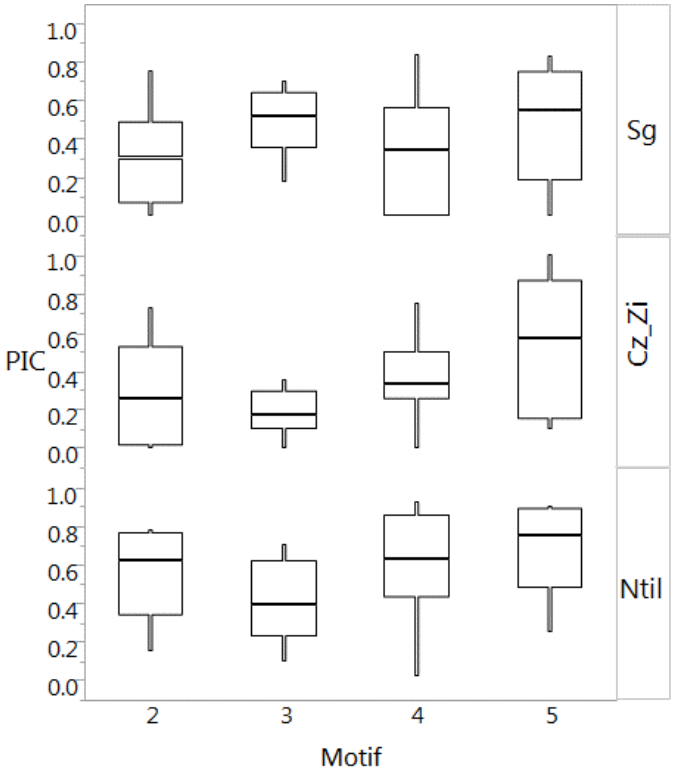


Figure 6b. Showing the different SSRs and their level of informativeness (PIC). Penta nucleotides were more informative than the other repeats Sg = Sarotherodon galilaeus, Cz_Zi = *Tilapia zillii* (composed of *Coptodon zillii* and *Tilapia zillii*), Ntil= Nile tilapia

c) Observed and Expected heterozygosity vs Population and Species

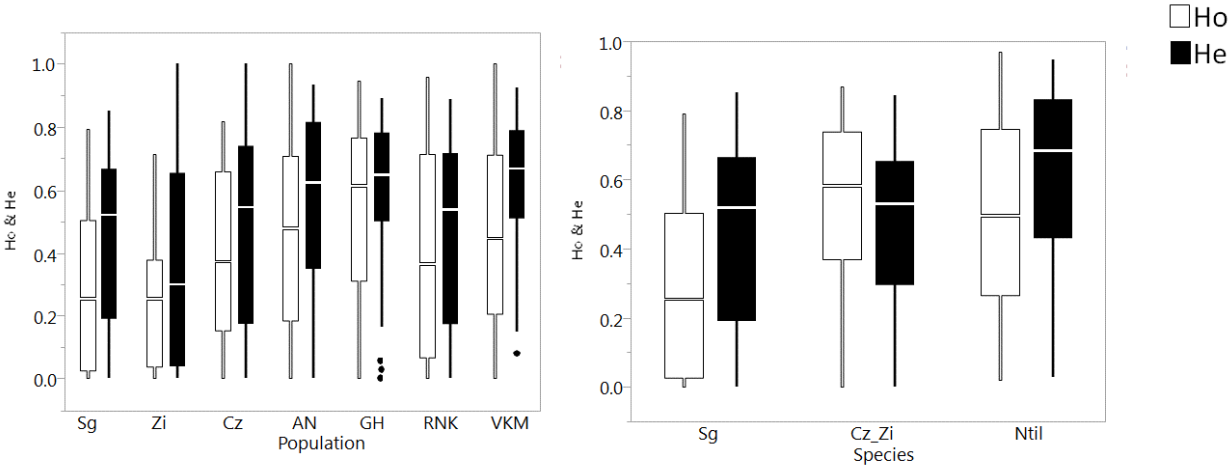


Figure 6c. Showing diversity of the developed markers in the different species. He = Expected heterozygosity, Ho = Observed heterozygosity. Sg = S. *galilaeus*, Zi = *Tilapia* zillii, Cz = *Coptodon zillii*, AN = Lake Albert Nile tilapia populations, GH = Lake George Nile tilapia populations, RNK = River Nile tilapia populations, VKM = Lake Victoria Nile tilapia populations.

4.5. Genetic structure as shown by:

4.5.1 Principal Coordinate Analysis (PCoA)

Figure 7, 8 and 9 shows PCoA analysis of haplotype and sequence length methods. Figure 7 shows analysis within Nile tilapia populations, Figure 8 shows analysis among all the populations used in this study while Figure 9 shows analysis between *Sarotherodon galilaeus* and *zillii.* This was done because *Sarotherodon galilaeus* and *Tilapia zillii* had shown some similarity to each other.

PCoA was chosen because it strongly differentiates populations into distinct clusters. The two allele calling methods showed almost equal results, with Lake Albert and Lake Victoria populations clustering together. Lake Albert and Lake Victoria have more populations that could be assigned to either populations than the others.

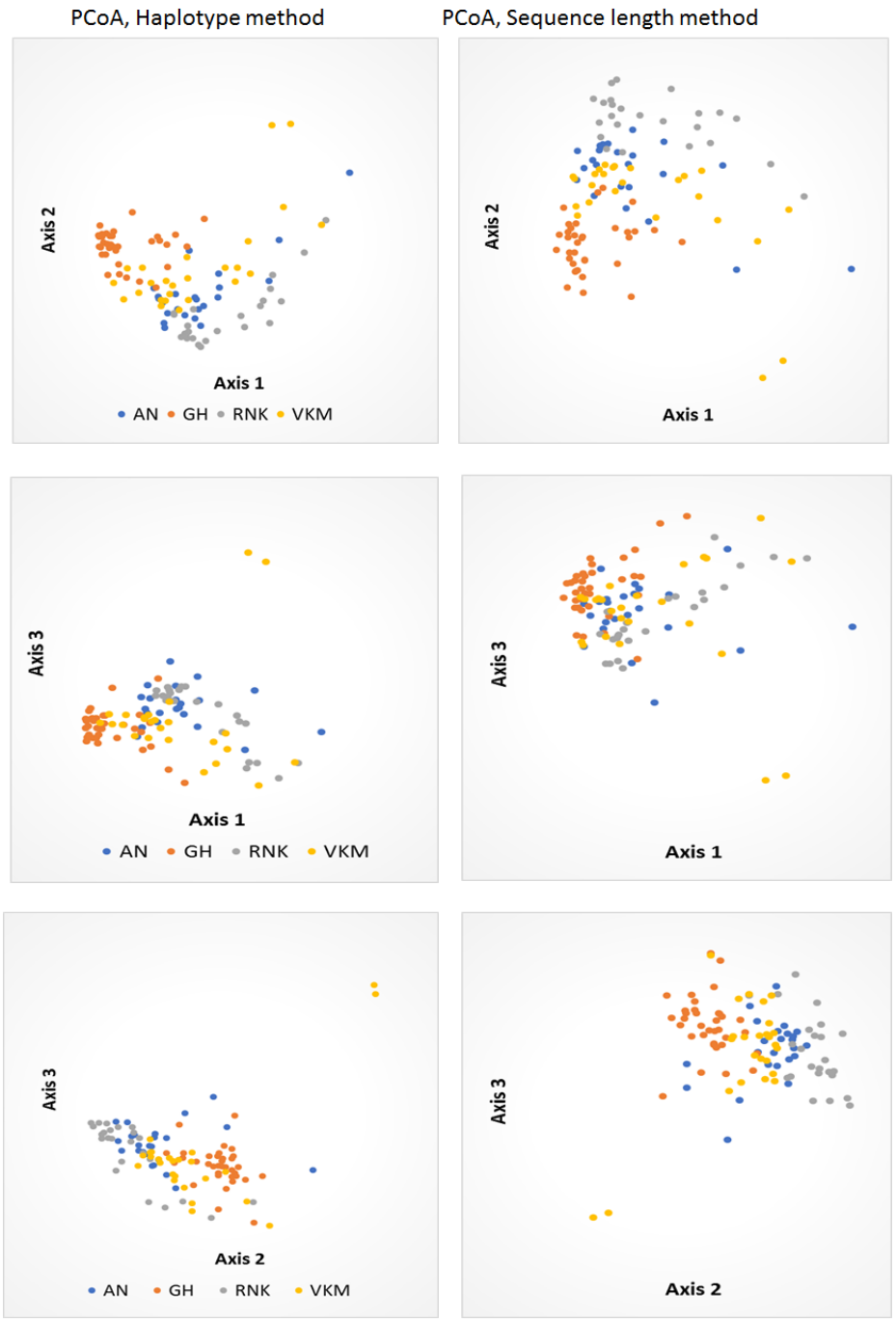


Figure 7. PCoA plots comparing haplotype and sequence length allele calling methods in Nile tilapia populations from Uganda. Axis are shown next to each other for both allele calling methods for ease of comparison. Lake Victoria populations show some admixture with Lake Albert populations. Lake Albert (AN) populations are shown in blue, Lake George (GH) populations in orange, River Nile (RNK) populations in grey and Lake Victoria (VKM) populations in yellow.

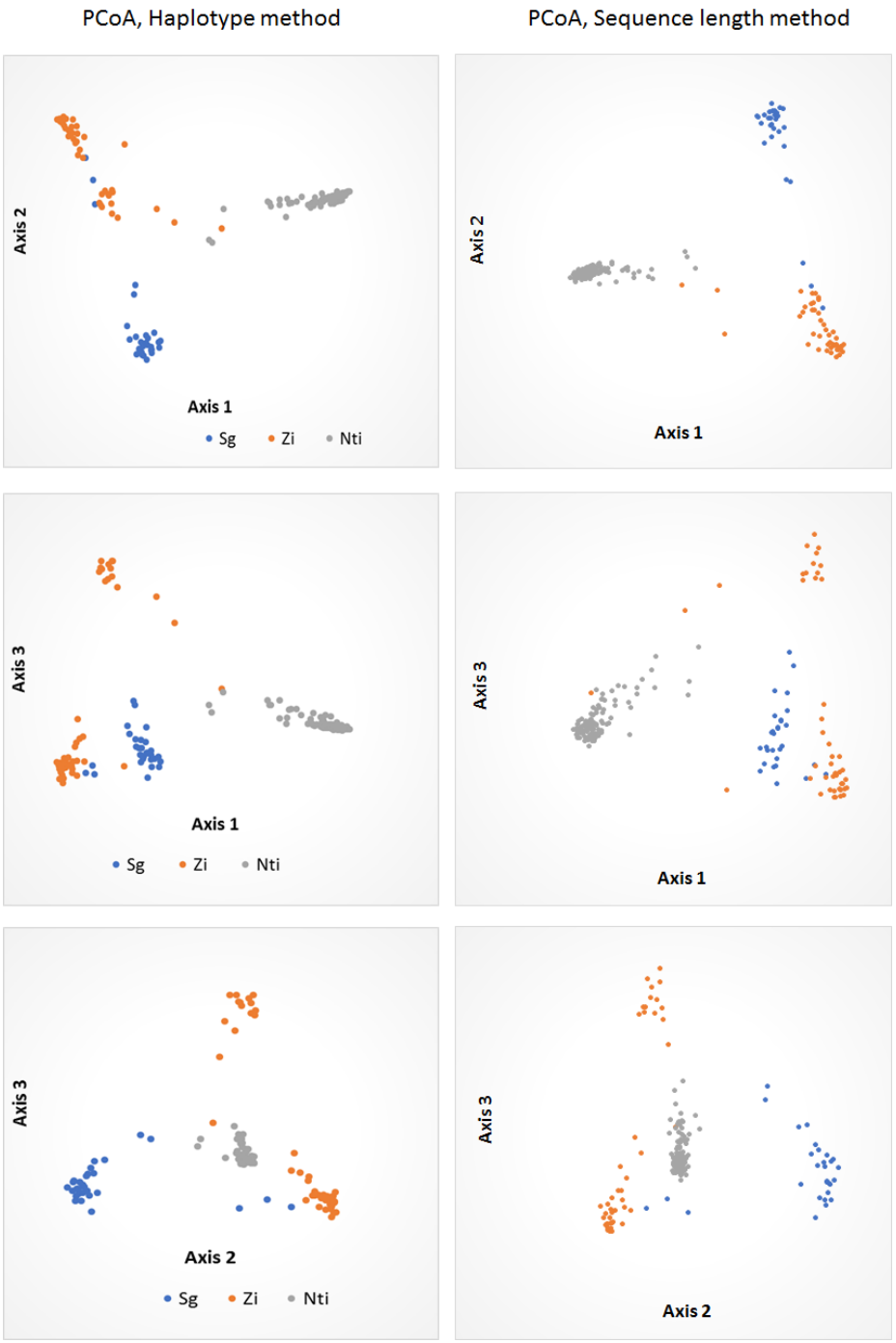


Figure 8. PCoA plots comparing haplotype and sequence length allele calling methods in all species. Axis are shown next to each other for both methods for ease of comparison. Sg = *Sarotherodon galilaeus* shown in blue, Zi = *Tilapia zillii* shown in orange, Nti = Nile tilapia shown in grey. The two orange Zillii clusters observed are two species of zillii, upper = *Tilapia zillii* (from Ethiopia) and lower = *Coptodon zillii* (from Burkina Faso).

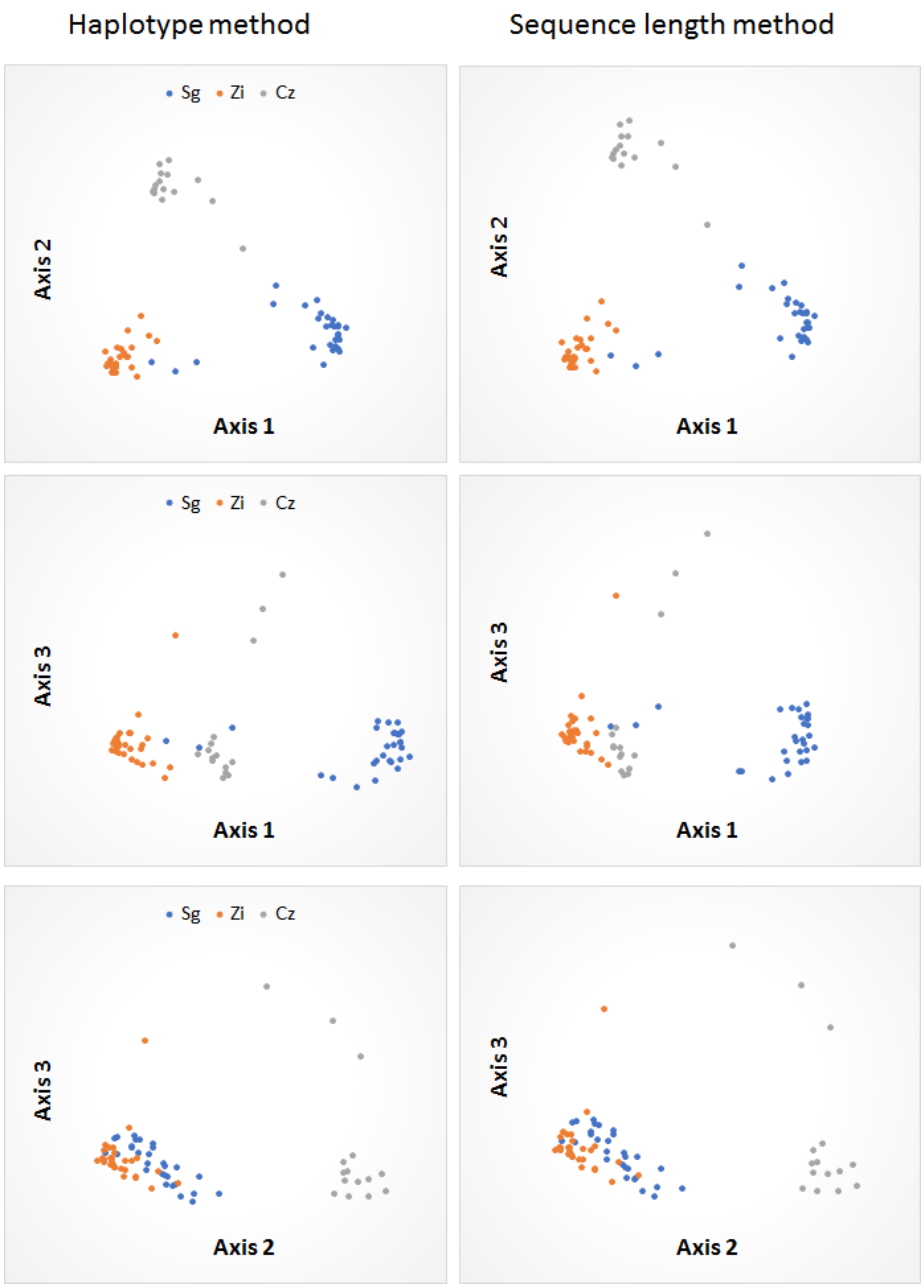


Figure 9. PCoA plots comparing haplotype and sequence length allele calling methods in 2 species, *Sarotherodon* galilaeus and *zillii*. Axis are shown next to each other for both methods. Sg = Sarotherodon galilaeus shown in blue, Zi = *Tilapia zillii* shown in orange.

4.5.2 Genetic structure as shown by STUCTURE 2.3.4

STRUCTURE analysis following haplotype and sequence length allele calling methods are shown below. These plots were obtained after running STRUCTURE HARVERSTER, an online program that validates multiple K values for maximum detection. The modal value of the distribution is the optimum K or the uppermost level of the structure. In the case of, *S. galilaeus* the best K is 2 in both methods (Figure 10a and b), while in *Tilapia zillii*, best K is 2 in both methods (Figure 10c and d). On the other hand, in Nile tilapia haplotype method, best K = 3 but in sequence length best K = 2 (Figure 11e and f) while in all populations, best K is 9 in haplotype but 7 in sequence length (Figure 11g and h).

In the structure plots (Figure 12a, b, c and d,) all species form their own clusters. The strongest split is between Nile tilapia and the other species (Figure 12c). Within species, two populations in *Tilapia zillii* were observed while three were observed in Nile tilapia: Lake George population was different from Albert and River Nile populations.

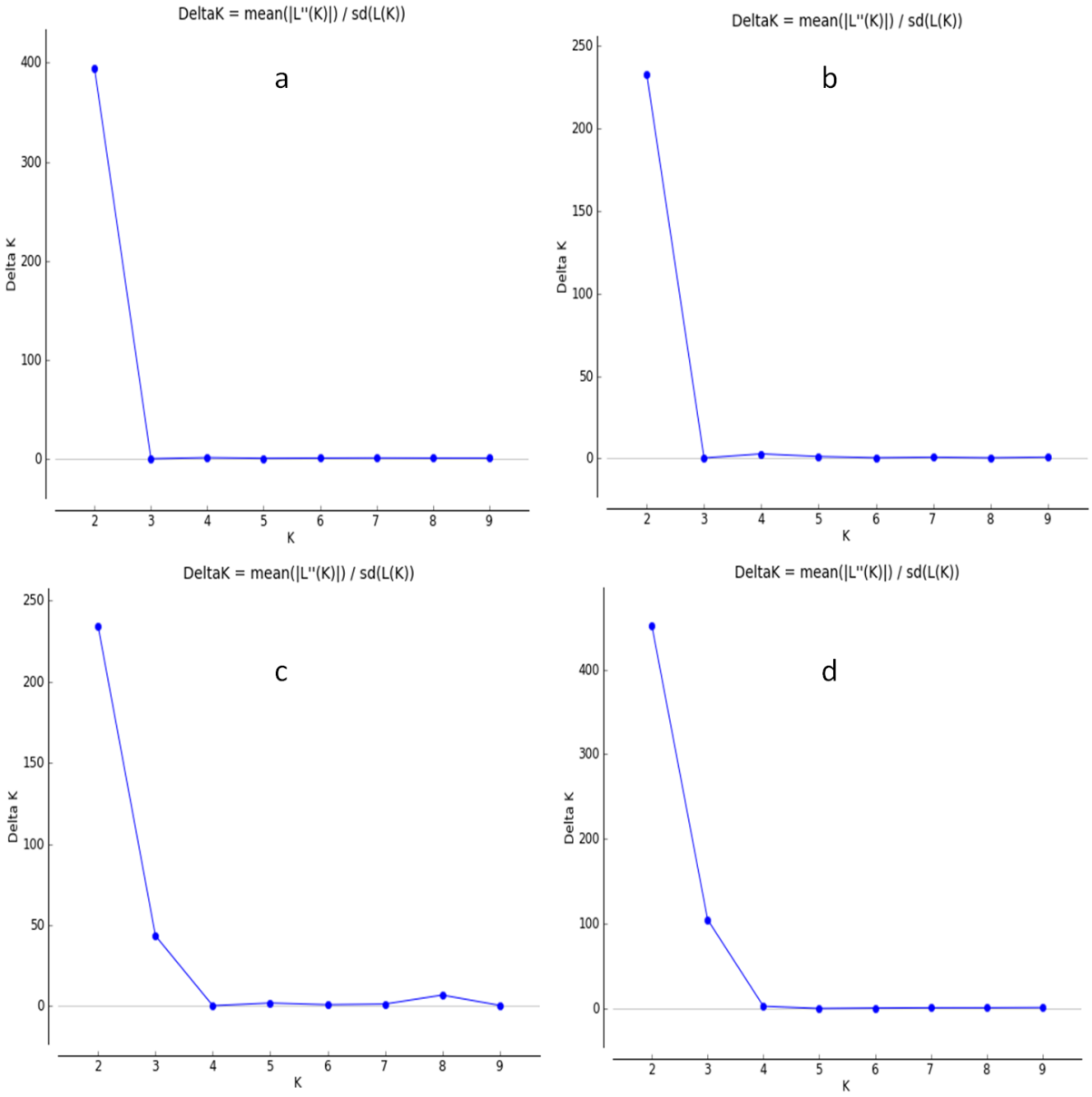


Figure 10. STRUCTURE HARVESTER plot showing K against Delta K. (a) *S. galilaeus*, haplotype method (b) *S. galilaeus*, sequence length method, (c) *Tilapia zillii*, haplotype method, (d) *Tilapia zillii*, sequence length method

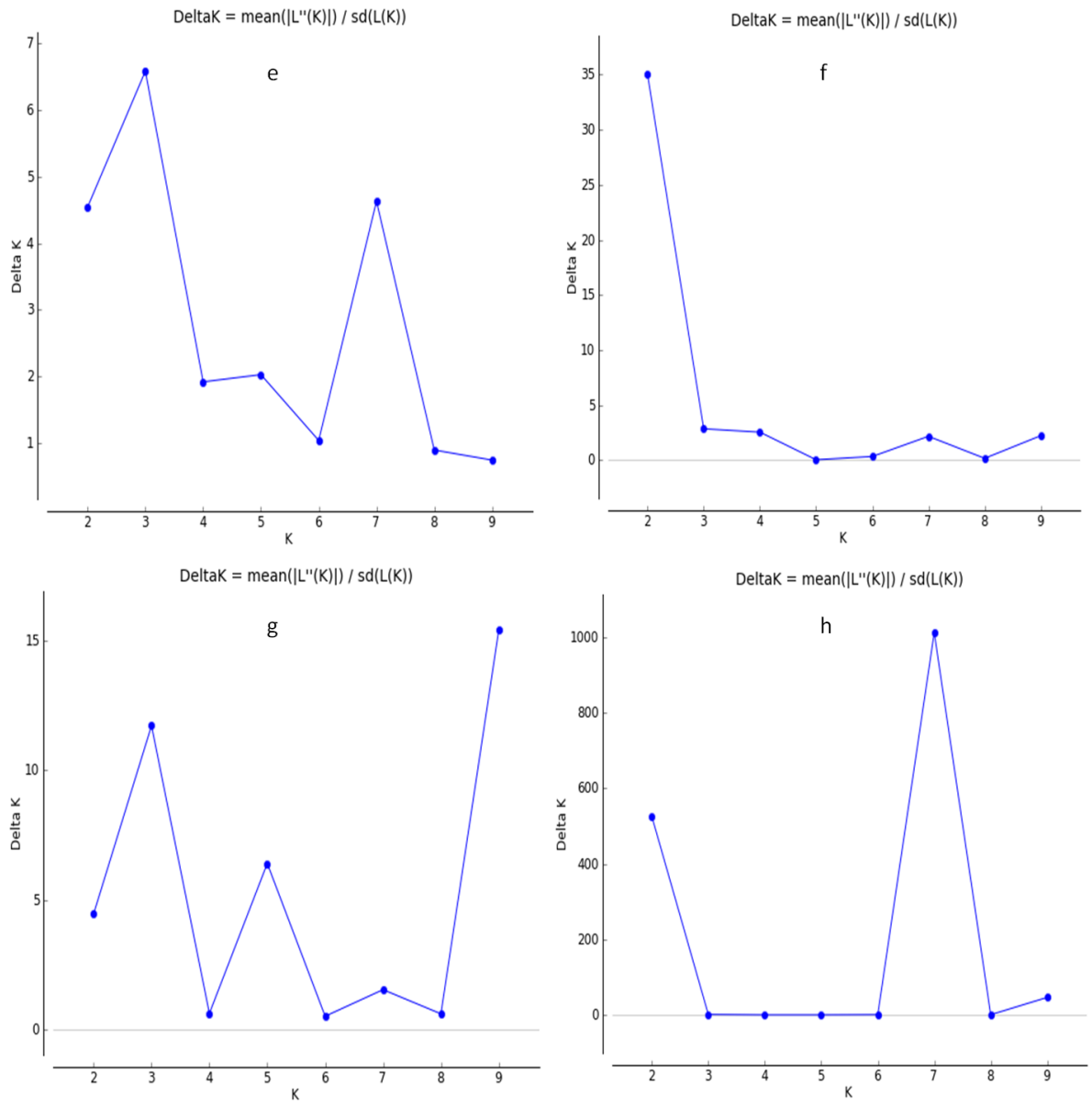
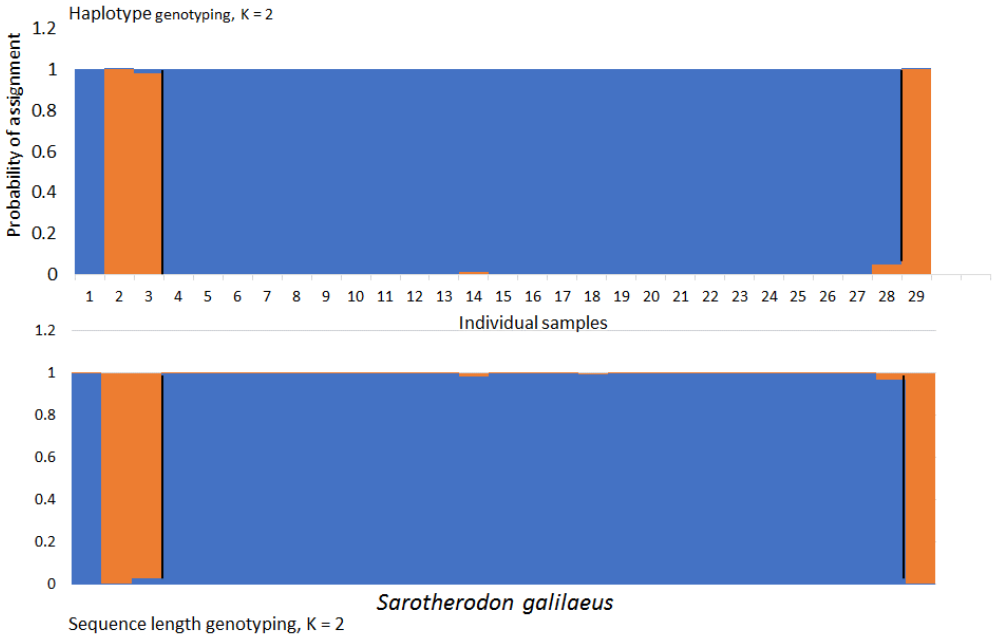


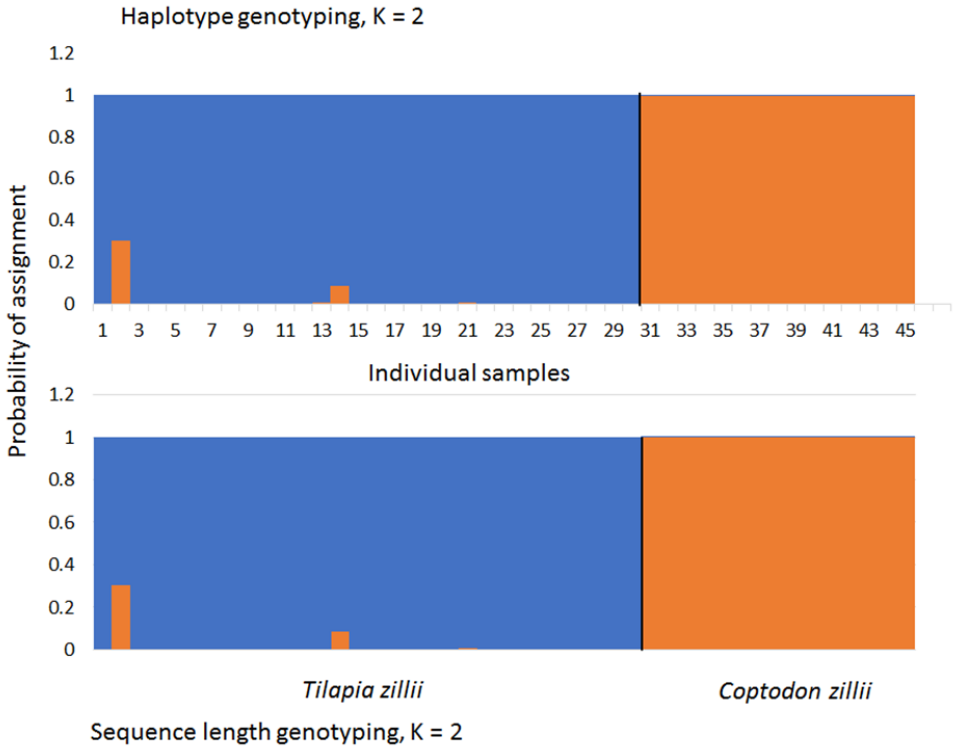
Figure 11. STRUCTURE HARVESTER plot showing K against Delta K in, (e) Nile tilapia population, haplotype method, (f) Nile tilapia population sequence length method, (g) all populations, haplotype method, (h) all populations, sequence length method. Optimum K varies only in two populations, Nile tilapia, K = 3 in Haplotype method, while K = 2 in Sequence length, and in all populations, K = 9 in Haplotype method, while K = 7 in Sequence length method

Microsatellite Cross-species Amplification and Utility in African Cichlids:

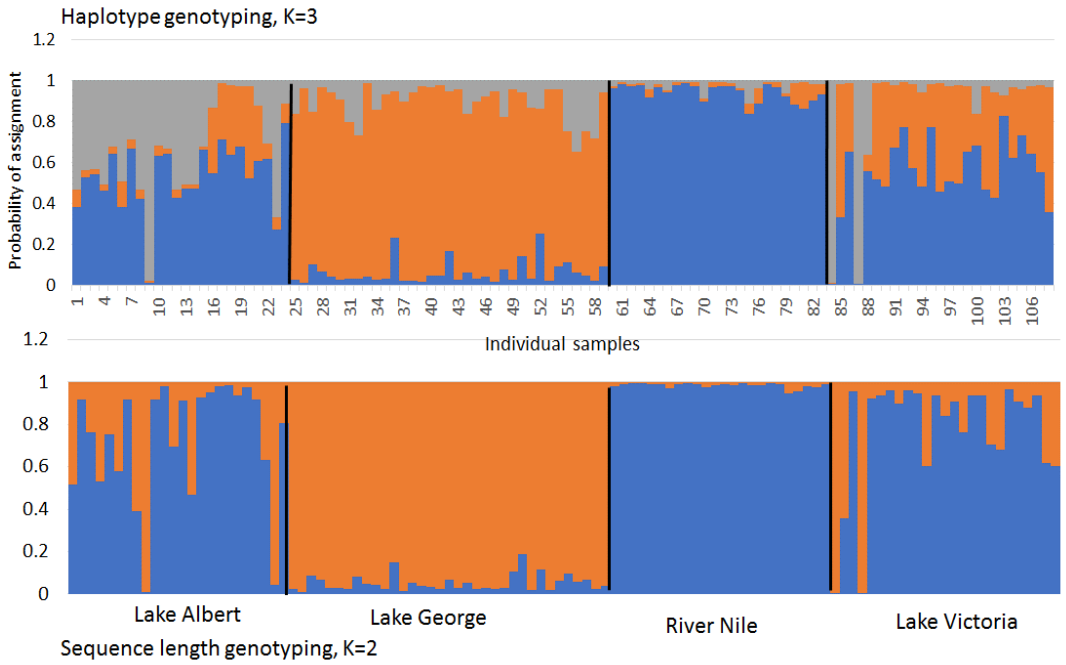
(a)



(b)



(c)



(d)

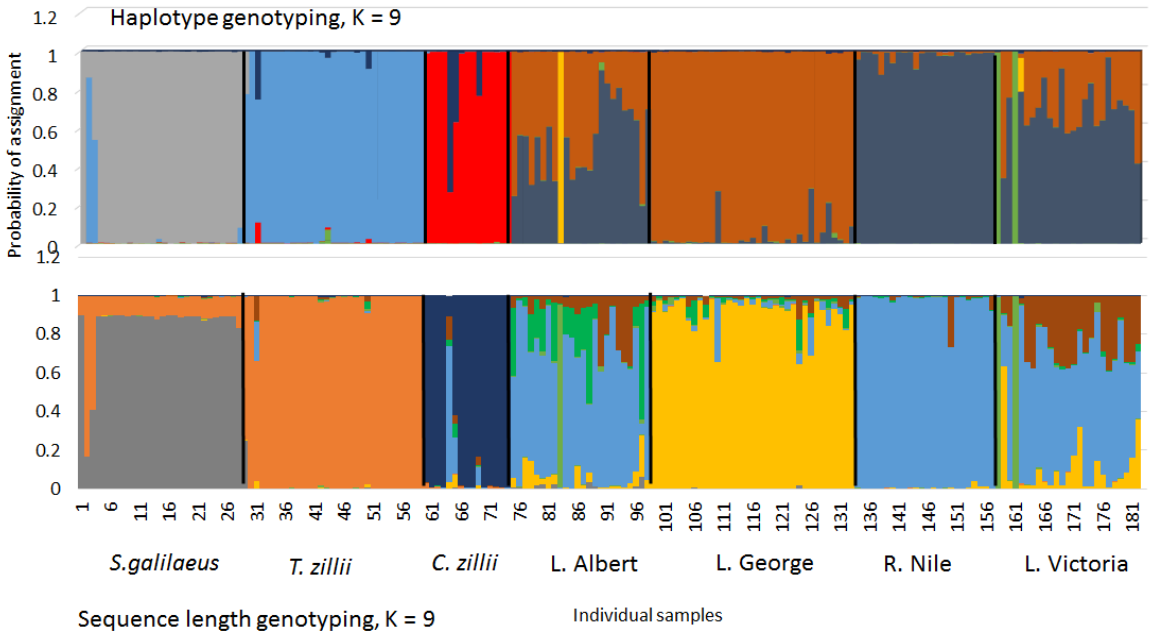


Figure 12. Bar plots showing probability of assignment of the populations to the genetic clusters inferred by clustering analysis (STRUCTURE). Each vertical bar represents individual sample and its probability of assignment to that cluster. The number of genetic clusters (K) which was best supported by the procedure proposed by Evanno et al. (ΔK) was represented. (a) *S. galilaeus*, (b) *Tilapia zillii* (c) Nile tilapia (d) all populations. In all populations graph (d) K = 9 was presented for both methods, although K = 7 was the optimum in sequence length method.

CHAPTER 5

5.0 DISCUSSION

5.1 Next generation sequencing (Illumina MiSeq)

As a tool for SSR discovery, next generation sequencing has proven to be excellent. This is mostly due to its ability to generate large amounts of DNA sequences many of which contain SSRs (Zalapa et al., 2012). Although this new technique has great capacity for *de novo* marker development, there remains a challenge in SSR genotyping and allele calling. Furthermore, there are other issues to be put into consideration during marker development. Such issues include the high rates of polymorphism which are more often associated with shorter motifs and higher number of repeats. This is as a result of higher rates of mutation in comparison to longer motifs with lower number of repeats (Zalapa et al., 2012). This study specifically used illumina MiSeq sequencing for SSR development, a NGS technique. The main advantage that was considered in choosing NGS is that there is no need to create enriched DNA libraries to retrieve all types of motifs. This is because by enriching DNA with probes, it makes it difficult to retrieve all types of motifs for example "AT" rich motifs (Curto et al., 2013). All these factors were put into consideration while developing markers in this study. Therefore, they have the capacity to be used for further population genetics studies.

5.2 Comparison of allele calling methods

During this study, two allele calling methods were used. The two methods differed in the number of alleles and in PIC in almost all the loci. This is not unusual since unlike in sequence length allele calling method, haplotype calling method systematically surveys the whole sequence thereby summarising all the variability from the repetitive motifs to SNPs in the flanking regions. Additionally, whereas the traditional method measures fragment length and relates the variation using capillary electrophoresis, NGS genotypes SSR markers through high throughput sequencing platforms such as illumina (De Barba et al., 2017), that was used in this study. This allows the combination of length and SNP information with the resultant high allelic information. This is in addition to other advantages such as possibility to automate, ease of replicability as well as low cost (De Barba et al., 2017).

5.3 Transferability of the developed primers

A few of the developed SSRs had PIC values less than 0.25 but 29 of them had values higher than 0.5 with an overall mean of 0.58 (Table 2), considering the haplotype allele calling.

Therefore, 29 of the SSRs were quite informative, and could therefore be fully utilized for future large-scale population genetics studies in Nile tilapia. PIC values agreed with the expected heterozygosity and allelic number among the studied populations thus further enhancing their usability.

According to Botstein et al. (1980) classification, developed SSRs with PIC values greater than 0.25 are regarded as reasonably informative and could therefore still be used. This study had 8 SSRs with PIC values greater than 0.25 but less than 0.5, therefore increasing usable SSRs developed here.

5.4 Cross amplification

Even though there are a lot of advances in current marker development procedures, isolation of new markers is still cumbersome (Zane et al., 2002). To solve this problem, cross-amplification, which is the use of primer pairs developed from one species on another, is a convenient alternative (Gen-Hua et al., 2010). However, some unexpected results were observed here as some locus that failed to amplify in Nile tilapia, the source species, were strongly amplified in *Tilapia zillii*, the target species. These loci are TI38, TI45 and TI46 (results not shown). Though these markers were not considered, they raised some questions.

It was however suspected that this may have happened due to allele drop out because of mutations in the primer binding sites. Additionally, this may also imply that the ability of markers to cross-amplify might be locus dependent (Gen-Hua et al., 2010), even though phylogenetic relationships or homoplasy could have played a role. Other reasons that could have explained this result is outbreeding and hybridization leading to admixture, which is rampant in tilapia, (O'connell and Wright, 1997). The markers were developed from Nile tilapia from Ethiopia and tested in Uganda. These populations could have been genetically different among each other. Additionally, the samples from *Tilapia zillii* had come from Ethiopia and therefore may have hybridized with the Nile tilapia therefore making transferability higher.

Most of the other loci however showed expected results, strongly amplifying in Nile tilapia and a few failings in *Sarotherodon* and *Tilapia* genera. This was congruent with earlier findings that the rate of success for cross-species amplification is directly correlated to the evolutionary distance between the species under study (Barbara et al., 2007). For example,

markers had higher cross-amplification rate (76%) in *Sarotherodon galilaeus* than in *T. zillii* (59%) and *C. zillii* (55%). Both *Oreochromis* and *Sarotherodon* genera are made up of mouth-brooders species, and constitute a monophyletic clade of tilapiine cichlids which is thought to have diverged about 12-21 million years ago (Bezault et al., 2012). This is in contrast to *Tilapia* genus, a substrate spawner, that is estimated to have diverged about 30-40 million years ago (Bezault et al., 2012).

5.5 Population structure

PCoA results indicated a possible gene flow between the populations understudy. This is possibly due to translocation of tilapiine species between these lakes thereby breaking the geographical isolation barrier. Consequently, populations that are closer to each other geographically were likely to show a closer genetic relationship (Karn and Jasieniuk, 2017). This is evident in Lake George and Lake Victoria as well as in Lake Albert populations. It is also remarkable that *Sarotherodon galilaeus* and *Coptodon zillii,* both from the same water body in Burkina Faso, were clustering closer together than other populations, though they belong to different genera (Figure 8).

STRUCTURE analysis also showed a potential presence of gene flow in *Oreochromis species* as shown by the many admixed Nile tilapia species especially in Lakes Albert and Victoria that had many individuals that could partially be assigned to either population (Figure 12). This is consistent with the generally believed fact that Nile tilapia populations in Lake Victoria were introduced from Lake Albert in the 1950s (Balirwa, 1992; Ogutu-Ohwayo, 1990) which lead to the systematic decline of native species such as *Oreochromis esculentus* and *Oreochromis variabilis*. As a result, it possible that Nile tilapia genepool may have been mixed with genes from other tilapiine species through hybridization and hence the appearance of unexpected results in the cross-amplification test as stated above.

5.6 Population genetic diversity

Populations that are not genetically diverse are likely to be incapable of adapting to the changing environment. Hybridization lowers the levels of genetic diversity either in farmed or wild populations, as is the case with tilapia (Wasonga et al., 2017). In addition to lowering genetic diversity, hybridization in tilapias has a direct effect on the purity of strains and species. Brood stocks from impure strains used in aquaculture are likely to cause reduction in productivity as many individuals may fail to survive changing environment while the surviving

ones may record lower growth rates (Wasonga et al., 2017). For example, in Asian countries of Thailand and The Philippines, the use of impure stocks led to serious reduction in yields. This led to concerted efforts by international organizations to remedy this situation. Finally, this culminated to the production of genetically improved farmed tilapia (GIFT) which started in April 1988 (Macaranas et al., 1995).

It is therefore important to constantly evaluate the level of hybridization and hence genetic diversity in conservation and population genetics (Frankham, 2005). Results of this study demonstrated an ongoing hybridization process that if unchecked may lead to the loss of pure breeds of *Oreochromis niloticus*.

5.7 Diversity of the developed SSRs

Forty-two SSR markers were discovered in this study, and all were used in primer design. Dinucleotides were the least with 8 discovered, followed by trinucleotides and Penta nucleotides. Even though dinucleotides are the most common type of repeats found in the genome (Tóth et al., 2000), this study discovered more tri, tetra and Penta repeats at 9, 16 and 9 respectively. Probably this was due to the fact that dinucleotides are harder to score than the other repeats.

Another thing to note is that this study avoided the use of mononucleotides for primer design mainly because of their high association with slippage errors as compared to other repeats (Curto et al., 2013). Furthermore, mononucleotide repeats are less dependable because of difficulties with amplification (Selkoe and Toonen, 2006).

The higher mean expected heterozygosity observed on splitting the two *zillii* populations, (Figure 6c), *Coptodon zillii* and *Tilapia zillii* was probably caused by hybridization in *Coptodon zillii*. In addition, the higher PIC (Figure 6b) observed in Nile tilapia was probably due to the higher number of samples used in this species. This is because PIC takes into consideration the sample size used, in which case Nile tilapia had the highest at 110 samples in comparison to 47 for *T. zillii* and 30 for S. *galilaeus*. River Nile populations had almost equal observed and expected heterozygosity (Figure 6c).

6.0 Conclusion and Prospects

6.1 Conclusion

Often, hybridization, with or without introgression, is a threat to native populations in various plant and animal taxa. Though some cases are reported, many others go unreported and therefore those reported do not portray the real situation on the ground. This then calls for the need to increase the use of microsatellite markers, that are more informative than morphological features and other markers with lower resolving power for detecting hybridization. Consequently, hybrids should be identified earlier, and effective measures undertaken.

However, it should be noted here that not all forms of hybridizations are bad. Some forms introduce to the native breeds combinations of desirable traits which may lead to increase in production. All that is required is monitoring to separate desirable from undesirable traits, in addition to separating pure stocks from hybrid stocks. Accordingly, genetic purity of cultured tilapiine will be maintained thus guarding against loss of diversity.

In this study, rather possible of hybridization was observed in Lake Albert and Lake Victoria populations, while appreciable levels were noticed in Lake George (Figure 12c). However, complete sampling is required to reveal further details as this study did not do it. In this regard, microsatellites have proven to be markers of choice in species/hybrid delineation. They can therefore be used by scientists to monitor hybridization and later give their findings to relevant authorities for implementation.

In addition, next generation sequencing technologies such as illumina sequencing are valuable tools in the development of such polymorphic markers, and therefore when utilized together with microsatellites, hybrid monitoring programs should be able to yield reproducible results in short amount of time and at lower costs. This study utilized both of these (microsatellites and illumina sequencing) and therefore the markers developed here are quite useful for further population and conservation genetics studies, especially in other Nile tilapia species.

As far as genotyping is concerned, haplotype method reduced bias and artefacts, thus giving more information than sequence length genotyping. This means that haplotype allele calling is a superior method and therefore recommended by this study. In addition, cross

amplification results were quite encouraging, further supporting the usability of the SSRs developed here.

6.2 Prospects

This study noted hybridization in Nile tilapia species. Going to the future, proactive measures of preventing both intentional and unintentional hybridization ought to be considered. Such measures include blocking of conduits that lead to unintentional introduction of alien species. On the other hand, intentional hybridizations or introduction of alien species for improvement of native tilapia breeds should only be done after proper risk assessment have been done, and where desirable traits are conferred to native species.

Additionally, timely detection and swift response procedures that takes into account all dynamics such as climate change should be put in place. Finally, monitoring programs that embrace scientific methods, such as use of microsatellites and other genetic markers for species/hybrid identification should be used. As a final resort, elimination of invasive species, where applicable, could be used to curb hybridization when not required.

7.0 REFERENCES

Agresti, J.J. et al., 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.

Allentoft, M.E. et al., 2009. Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. Biotechniques, 46(3): 195-200.

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data.

Bailey, R., 1994. Guide to the fishes of the River Nile in the Republic of the Sudan. Journal of Natural History, 28(4): 937-970.

Balirwa, J., 1992. The evolution of the fishery of Oreochromis niloticus (Pisces: Cichlidae) in Lake Victoria. Hydrobiologia, 232(1): 85-89.

Barbara, T. et al., 2007. Cross-species transfer of nuclear microsatellite markers: potential and limitations. Molecular Ecology, 16(18): 3759-3767.

Bezault, E., Rognon, X., Gharbi, K., Baroiller, J.-F., Chevassus, B., 2012. Microsatellites cross-species amplification across some African cichlids. International Journal of Evolutionary Biology, 2012.

Botstein, D., White, R.L., Skolnick, M., Davis, R.W., 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American journal of human genetics, 32(3): 314.

Canonico, G.C., Arthington, A., McCrary, J.K., Thieme, M.L., 2005. The effects of introduced tilapias on native biodiversity. Aquatic Conservation: Marine and Freshwater Ecosystems, 15(5): 463-483.

Caro, T., Laurenson, M.K., 1994. Ecological and genetic factors in conservation: a cautionary tale. Science-AAAS-Weekly Paper Edition-including Guide to Scientific Information, 263(5146): 485-486.

Castoe, T.A. et al., 2012. Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. PloS one, 7(2): e30953.

Charlesworth, B., Charlesworth, D., 1999. The genetic basis of inbreeding depression. Genetics Research, 74(3): 329-340.

Colonna, V. et al., 2009. Comparing population structure as inferred from genealogical versus genetic information. European Journal of Human Genetics, 17(12): 1635.

Costa-Pierce, B.A., 2003. Rapid evolution of an established feral tilapia (Oreochromis spp.): the need to incorporate invasion science into regulatory structures. Biological Invasions, 5(1-2): 71-84.

Curto, M.A. et al., 2013. Evaluation of microsatellites of Catha edulis (qat; Celastraceae) identified using pyrosequencing. Biochemical systematics and ecology, 49: 1-9.

De Barba, M. et al., 2017. High-throughput microsatellite genotyping in ecology: improved accuracy, efficiency, standardization and success with low-quantity and degraded DNA. Molecular ecology resources, 17(3): 492-507.

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.

Earl, D.A., 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation genetics resources, 4(2): 359-361.

Eknath, A.E., Hulata, G., 2009. Use and exchange of genetic resources of Nile tilapia (Oreochromis niloticus). Reviews in Aquaculture, 1(3-4): 197-213.

Epifanio, J.M., Johnson, J., Kassler, T.W., 2003. Genetic analysis of Tennessee sport fisheries (2000-2003): largemouth bass, smallmouth bass, rainbow trout, and brown trout, INHS Center for Aquatic Ecology.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., Turcatti, G., 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic acids research, 34(3): e22-e22.

Fisheries, F., 2010. Aquaculture Department (2009) The state of world fisheries and aquaculture 2008. Food and agriculture organization of the United Nations, Rome.

Fogelqvist, J., Niittyvuopio, A., Ågren, J., Savolainen, O., Lascoux, M., 2010. Cryptic population genetic structure: the number of inferred clusters depends on sample size. Molecular ecology resources, 10(2): 314-323.

Frankham, R., 2005. Genetics and extinction. Biological conservation, 126(2): 131-140.

Fuller, R. et al., 1998. The integration of field survey and remote sensing for biodiversity assessment: a case study in the tropical forests and wetlands of Sango Bay, Uganda. Biological conservation, 86(3): 379-391.

Gen-Hua, Y., Kovacs, B., Orban, L., 2010. A new problem with cross-species amplification of microsatellites: generation of non-homologous products. Zoological Research, 31(2): 131-140.

Glenn, T.C., Schable, N.A., 2005. Isolating microsatellite DNA loci. Methods in enzymology, 395: 202-222.

Goudet, J., 2001. FSTAT, a program to estimate and test gene diversity and fixation indices (version 2.9. 3). http://www2. unil. ch/popgen/softwares/fstat. htm.

Goudswaard, K.P., Witte, F., Chapman, L.J., 2002. Decline of the African lungfish (Protopterus aethiopicus) in Lake Victoria (East Africa). African Journal of Ecology, 40(1): 42-52.

Guillot, G., 2008. Inference of structure in subdivided populations at low levels of genetic differentiation—the correlated allele frequencies model revisited. Bioinformatics, 24(19): 2222-2228.

Hedrick, P.W., 2001. Conservation genetics: where are we now? Trends in Ecology & Evolution, 16(11): 629-636.

Hubisz, M.J., Falush, D., Stephens, M., Pritchard, J.K., 2009. Inferring weak population structure with the assistance of sample group information. Molecular ecology resources, 9(5): 1322-1332.

Jarne, P., Lagoda, P.J., 1996. Microsatellites, from molecules to populations and back. Trends in ecology & evolution, 11(10): 424-429.

Jones, A., Avise, J., 1997. Microsatellite analysis of maternity and the mating system in the Gulf pipefish Syngnathus scovelli, a species with male pregnancy and sex-role reversal. Molecular Ecology, 6(3): 203-213.

Kalinowski, S.T., Taper, M.L., Marshall, T.C., 2007. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Molecular ecology, 16(5): 1099-1106.

Karn, E., Jasieniuk, M., 2017. Genetic diversity and structure of Lolium perenne ssp. multiflorum in California vineyards and orchards indicate potential for spread of herbicide resistance via gene flow. Evolutionary applications, 10(6): 616-629.

Kearse, M. et al., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 28(12): 1647-1649.

Kocher, T.D., Lee, W.-J., Sobolewska, H., Penman, D., McAndrew, B., 1998. A genetic linkage map of a cichlid fish, the tilapia (Oreochromis niloticus). Genetics, 148(3): 1225-1232.

Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., Mayrose, I., 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Molecular ecology resources, 15(5): 1179-1191.

Kornfield, I., Smith, P.F., 2000. African cichlid fishes: model systems for evolutionary biology. Annual Review of Ecology and Systematics, 31(1): 163-196.

Levêque, C., Oberdorff, T., Paugy, D., Stiassny, M., Tedesco, P., 2008. Global diversity of fish (Pisces) in freshwater. Hydrobiologia, 595(1): 545-567.

Liu, Z., Li, P., Argue, B., Dunham, R., 1998. Inheritance of RAPD markers in channel catfish (Ictalurus punctatus), blue catfish (I. furcatus), and their F1, F2 and backcross hybrids. Animal Genetics, 29(1): 58-62.

Liu, Z.J., Cordes, J., 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 238(1): 1-37.

Lowe-McConnell, R., 2000. The roles of tilapias in ecosystems, Tilapias: Biology and exploitation. Springer, pp. 129-162.

Lung'Ayia, H.B., M'harzi, A., Tackx, M., Gichuki, J., Symoens, J., 2000. Phytoplankton community structure and environment in the Kenyan waters of Lake Victoria. Freshwater Biology, 43(4): 529-543.

Macaranas, J.M. et al., 1995. Genetic improvement of farmed tilapias: biochemical characterization of strain differences in Nile tilapia. Aquaculture International, 3(1): 43-54.

Mair, G., Abucay, J., Abella, T., Beardmore, J., Skibinski, D., 1997. Genetic manipulation of sex ratio for the large-scale production of all-male tilapia Oreochromis niloticus. Canadian Journal of Fisheries and Aquatic Sciences, 54(2): 396-404.

Margulies, M. et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437(7057): 376-380.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal, North America. 17.

Mehanna, R.-A., 2004. Poverty and economic development: not as direct as it may seem. The Journal of Socio-Economics, 33(2): 217-228.

Mikheyev, A.S., Vo, T., Wee, B., Singer, M.C., Parmesan, C., 2010. Rapid microsatellite isolation from a butterfly by de novo transcriptome sequencing: performance and a comparison with AFLP-derived distances. PLoS One, 5(6): e11212.

Miller, M.P., Knaus, B.J., Mullins, T.D., Haig, S.M., 2013. SSR_pipeline: A bioinformatic infrastructure for identifying microsatellites from paired-end Illumina high-throughput DNA sequencing data. Journal of Heredity, 104(6): 881-885.

Muneer, P.A. et al., 2009. Genetic variation and population structure of endemic yellow catfish, Horabagrusbrachysoma (Bagridae) among three populations of Western Ghat region using RAPD and microsatellite markers. Molecular biology reports, 36(7): 1779-1791.

Nagl, S. et al., 2001. Classification and phylogenetic relationships of African tilapiine fishes inferred from mitochondrial DNA sequences. Molecular phylogenetics and evolution, 20(3): 361-374.

Nelson, M.N. et al., 2010. Aligning a new reference genetic map of Lupinus angustifolius with the genome sequence of the model legume, Lotus japonicus. DNA research, 17(2): 73-83.

Njiru, M., Mkumbo, O., Van der Knaap, M., 2010. Some possible factors leading to decline in fish species in Lake Victoria. Aquatic Ecosystem Health & Management, 13(1): 3-10.

O'brien, P., 1991. Molecular mechanisms of quinone cytotoxicity. Chemico-biological interactions, 80(1): 1-41.

O'connell, M., Wright, J.M., 1997. Microsatellite DNA in fishes. Reviews in Fish biology and fisheries, 7(3): 331-363.

Ogutu-Ohwayo, R., 1990. The decline of the native fishes of lakes Victoria and Kyoga (East Africa) and the impact of introduced species, especially the Nile perch, Lates niloticus, and the Nile tilapia, Oreochromis niloticus. Environmental biology of fishes, 27(2): 81-96.

Ostrander, E.A., Jong, P.M., Rine, J., Duyk, G., 1992. Construction of small-insert genomic DNA libraries highly enriched for microsatellite repeat sequences. Proceedings of the National Academy of Sciences, 89(8): 3419-3423.

Peakall, R., Smouse, P.E., 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Resources, 6(1): 288-295.

Pouyaud, L., Agnèse, 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.

Primmer, C., Møller, A., Ellegren, H., 1996. A wide-range survey of cross-species microsatellite amplification in birds. Molecular Ecology, 5(3): 365-378.

Rico, C., Rico, I., Hewitt, G., 1996. 470 million years of conservation of microsatellite loci among fish species. Proceedings of the Royal Society of London B: Biological Sciences, 263(1370): 549-557.

Rognon, X., Andriamanga, M., McAndrew, B., Guyomard, R., 1996. Allozyme variation in natural and cultured populations in two tilapia species: Oreochromis niloticus and Tilapia zillii. Heredity, 76(6): 640.

Ros, A.F., Zeilstra, I., Oliveira, R.F., 2003. Mate choice in the Galilee St. Peter's fish, Sarotherodon galilaeus. Behaviour, 140(8): 1173-1188.

Selkoe, K.A., Toonen, R.J., 2006. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecology letters, 9(5): 615-629.

Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nature biotechnology, 26(10): 1135-1145.

Spataru, P., 1978. Food and feeding habits of Tilapia zillii (Gervais)(Cichlidae) in Lake Kinneret (Israel). Aquaculture, 14(4): 327-338.

Suresh, E., Tiwari, V., Sekar, M., Sankar, M., Kathirvelpandian, A., 2013. Genetic structure of populations of Mugil cephalus using RAPD markers. African Journal of Biotechnology, 12(44): 6262-6266.

Tautz, D., 1989. Hypervariabflity of simple sequences as a general source for polymorphic DNA markers. Nucleic acids research, 17(16): 6463-6471.

Tóth, G., Gáspári, Z., Jurka, J., 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome research, 10(7): 967-981.

Trewavas, E., Teugels, G.G., 1991. Sarotherodon. Check-list of the Freshwater Fishes of Africa, CLOFFA, 4: 425-437.

Turcatti, G., Romieu, A., Fedurco, M., Tairi, A.-P., 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic acids research, 36(4): e25-e25.

Turner, G.F., Seehausen, O., Knight, M.E., Allender, C.J., Robinson, R.L., 2001. How many species of cichlid fishes are there in African lakes? Molecular Ecology, 10(3): 793-806.

Untergasser, A. et al., 2012. Primer3—new capabilities and interfaces. Nucleic acids research, 40(15): e115-e115.

Vrijenhoek, R., 1998. Conservation genetics of freshwater fish. Journal of fish Biology, 53(sA): 394-412.

Wasonga, A., Daniel, W., Brian, O., 2017. Interspecific Hybridization of Tilapiines in Lake Victoria, Kenya. J Fisheries Livest Prod, 5(235): 2.

Weber, J.L., Wong, C., 1993. Mutation of human short tandem repeats. Human molecular genetics, 2(8): 1123-1128.

Weyl, O.L., Ribbink, A.J., Tweddle, D., 2010. Lake Malawi: fishes, fisheries, biodiversity, health and habitat. Aquatic Ecosystem Health & Management, 13(3): 241-254.

Wright, J., 1993. DNA fingerprinting of fishes. Biochemistry and molecular biology of fishes, 2: 57-91.

Xiong, J., 2006. Essential bioinformatics. Cambridge University Press.

Zalapa, J.E. et al., 2012. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. American journal of botany, 99(2): 193-208.

Zane, L., Bargelloni, L., Patarnello, T., 2002. Strategies for microsatellite isolation: a review. Molecular ecology, 11(1): 1-16.

Zhang, J., Kobert, K., Flouri, T., Stamatakis, A., 2013. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics, 30(5): 614-620.

Zucchi, M., Brondani, R., Pinheiro, J., Brondani, C., Vencovsky, R., 2002. Transferability of microsatellite markers from Eucalyptus spp. to Eugenia dysenterica (Myrtaceae family). Molecular Ecology Resources, 2(4): 512-513.

8.0 Appendices

Appendix 1. Tables showing the composition of some of the chemicals used in this study

1 x TAE, pH 8.3; 1 liter
(40 mM Tris, 20 mM acetate, 1 mM EDTA); 1000ml

- 800 ml H_2O
- 4.84 g Tris base (M_{rel}=121.14)
- 1.14 ml glacial acetic acid
- 2 ml 0.5 M EDTA pH 8.0
- add H_2O to final volume 1 liter

Note: if necessary adjust pH to 8.3 with HCl, autoclave

Wash Buffer
10 mM Tris-HCl, pH 7.5 with 80% EtOH

DNA loading dye; 1 ml
(15% Ficoll)

- 150 mg Ficoll-Typ 400
- 1 ml 1xTAE
- einige Kristalle Bromphenolblau
- 1 µl Safe Green (1:10.000) for intas stain 1:100.000?)

1 M Tris-HCl, pH7.4 – 8.5; 1000 ml

1. 600 ml H_2O
2. 121.1g Tris base (M_{rel}=121.14)
3. adjust pH to **7.4** by adding 70 ml 37.5 % HCl

   for pH 7.6, 8.0, **8.5** add 60 ml, 42 ml, 15 ml

   (105 ml, 90 ml, 63 ml, 22 ml 25 % HCl)

1. add H_2O to final volume 1 liter

autoclave

0,5 M EDTA, pH8; 500ml

1. 400 ml H_2O
2. 73.06 (M=292.25) or 93.05 g EDTA-Na_2 (M_{rel}=372.24) adjust pH to 8.0 with ~14 g NaOH pellets
3. add H_2O to final volume 500ml

Note: EDTA will not completely dissolve until pH8
autoclave

Elution Buffer; 250 ml
10 mM Tris-HCl, pH 7.5

- 1:100 dilution of a 1 M Tris-HCl pH7.5

or 1:5 dilution of a 50 mM Tris-HCl pH7.5

or

- 250 ml H_2O
- 0.303 g Tris base (Mrel=121.14)
- adjust pH with HCl

Lysis Buffer NEW; 120 ml
2% SDS, 2%PVP 40, 250 mM NaCl, 200 M Tris HCl, 5mM EDTA, pH8.0

- 71 ml H_2O
- 1.76 g NaCl
- 24 ml 1 M Tris HCl pH8.0
- 1.2 ml 0.5M EDTA pH8.0
- autoclave for 20 min 121°C

then add

- 2.4 g PVP 40

Binding Buffer; 250 ml
2 M GuHCl in 95 % EtOH

- 237.5 ml EtOH abs.
- 12.5 ml H_2O
- 47.765 g guanidine hydrochloride (M_{rel}=95.53)

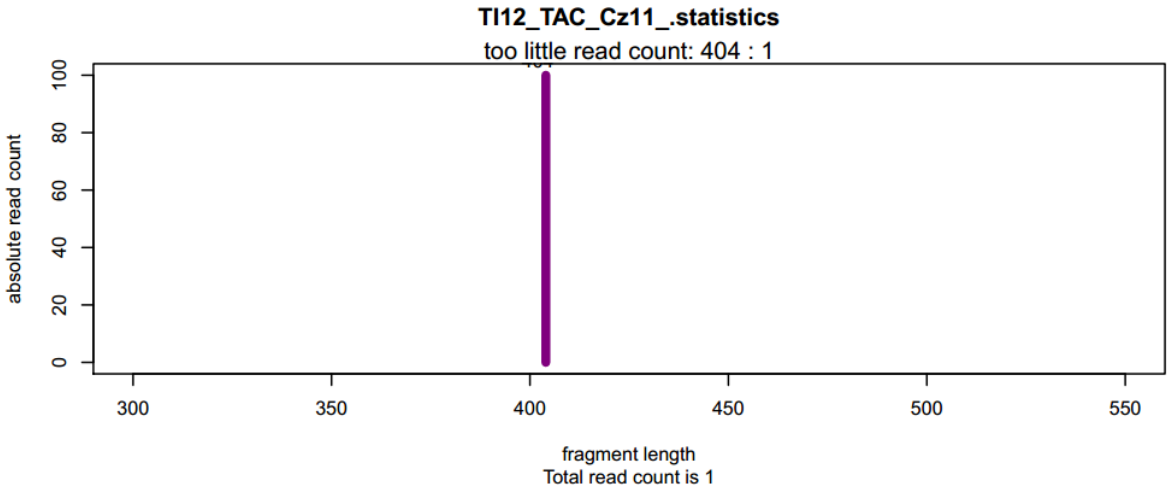
Note: do not autoclave or filter sterilize

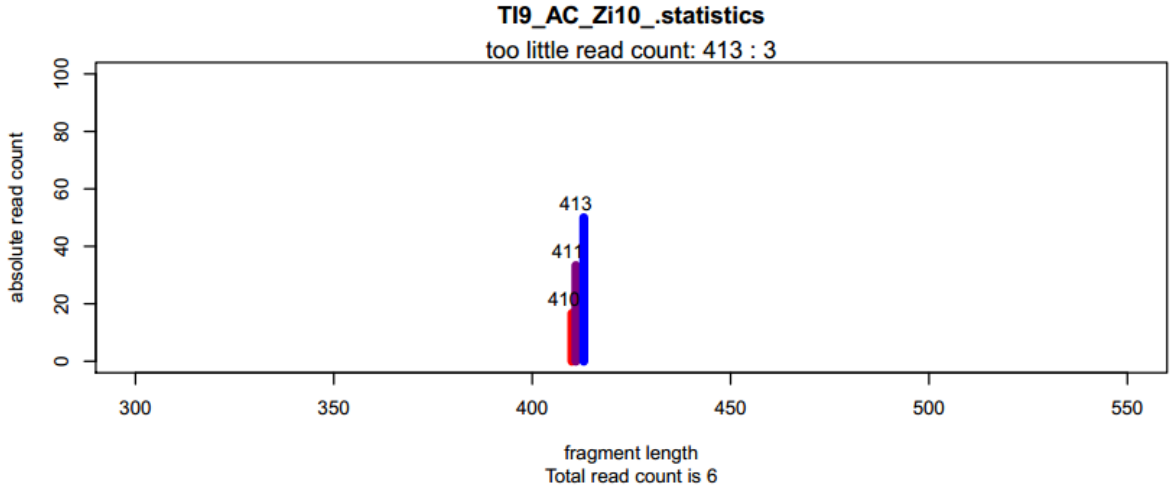
70 % EtOH prepare always fresh and 70ml + 30ml (not topping off to 100 ml)

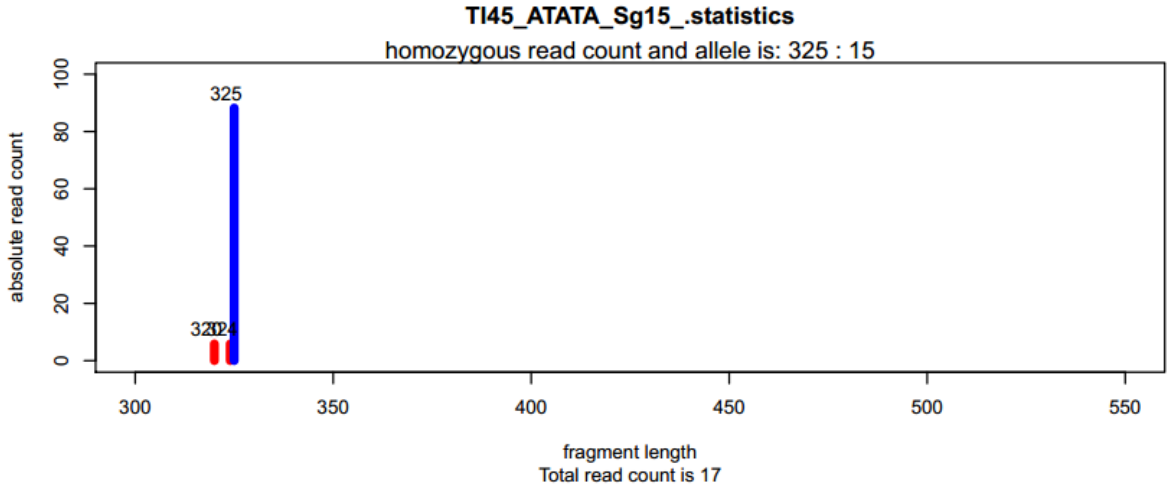
10X TAE Running Buffer, 1 liter

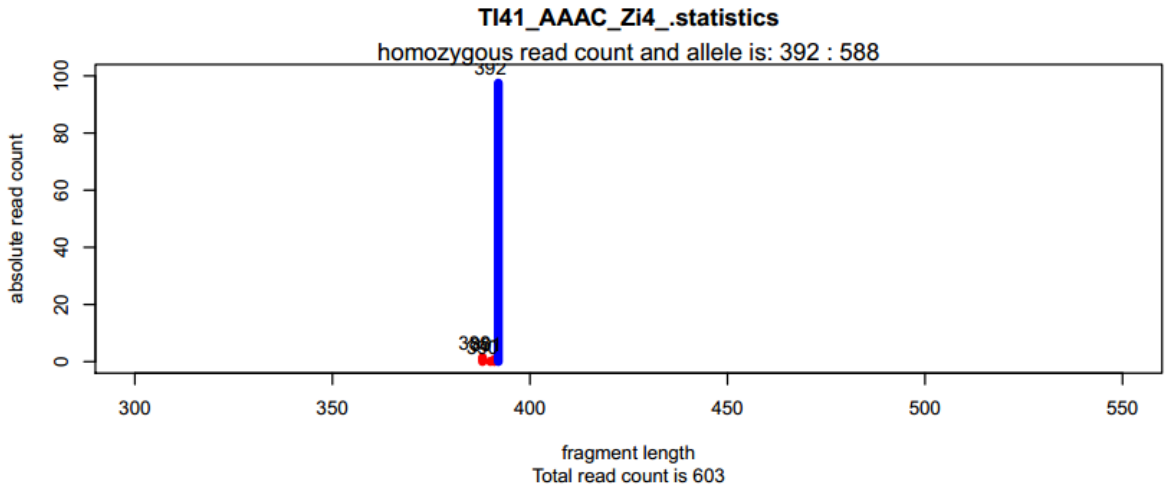
- Tris-base (M_{rel}= 121.14) 48.4 g
- Glacial Acetic Acid 10.0 ml
- EDTA (Free Acid, f.w. 292.25) 2.92 g
- DI Water to 1.0 L

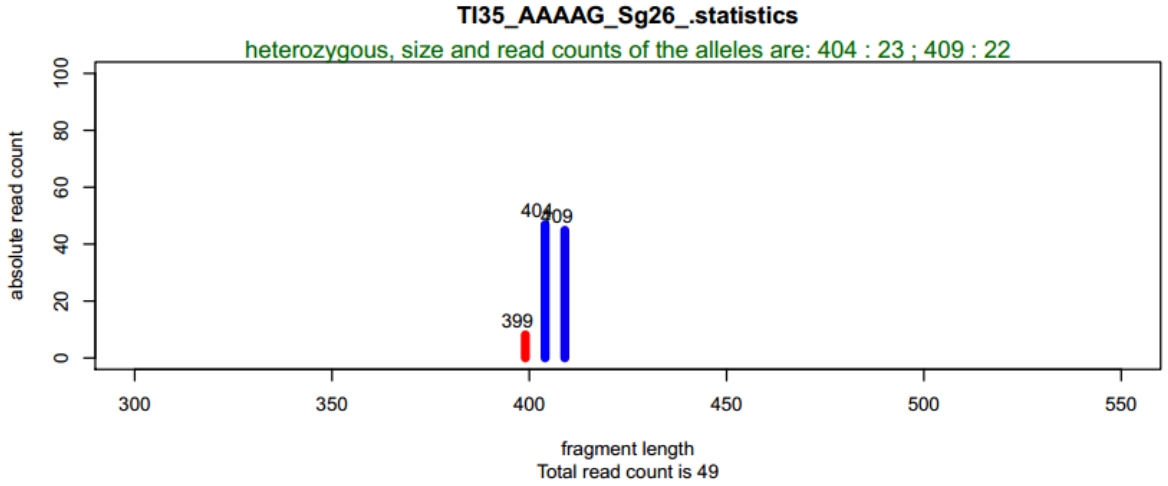
Store at room temperature

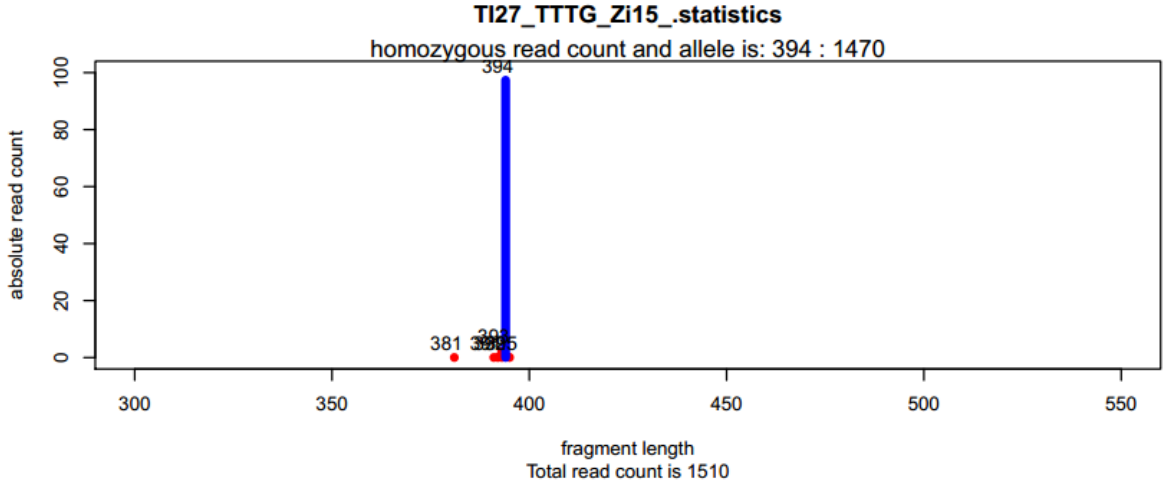

TI12_TAC_Cz11_.statistics
too little read count: 404 : 1
absolute read count
0
20
40
60
80
100
300
350
400
450
500
550
fragment length
Total read count is 1

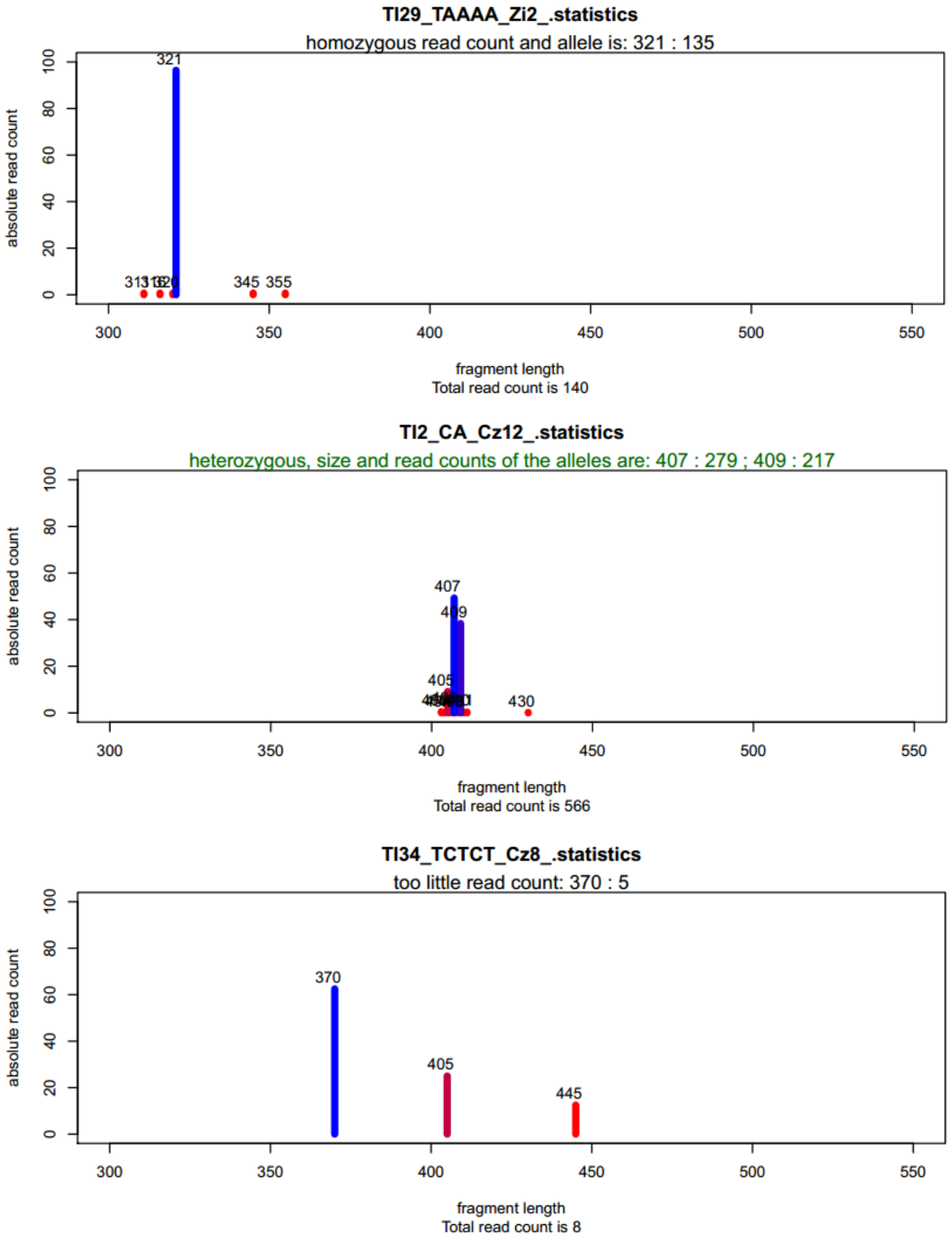

TI9_AC_Zi10_.statistics
too little read count: 413 : 3
absolute read count
0
20
40
60
80
100
413
411
410
300
350
400
450
500
550
fragment length
Total read count is 6


TI45_ATATA_Sg15_.statistics
homozygous read count and allele is: 325 : 15
absolute read count
0
20
40
60
80
100
325
300
350
400
450
500
550
fragment length
Total read count is 17


TI41_AAAC_Zi4_.statistics
homozygous read count and allele is: 392 : 588
absolute read count
fragment length
Total read count is 603


TI35_AAAAG_Sg26_.statistics
heterozygous, size and read counts of the alleles are: 404 : 23 ; 409 : 22
absolute read count
fragment length
Total read count is 49


TI27_TTTG_Zi15_.statistics
homozygous read count and allele is: 394 : 1470
absolute read count
fragment length
Total read count is 1510

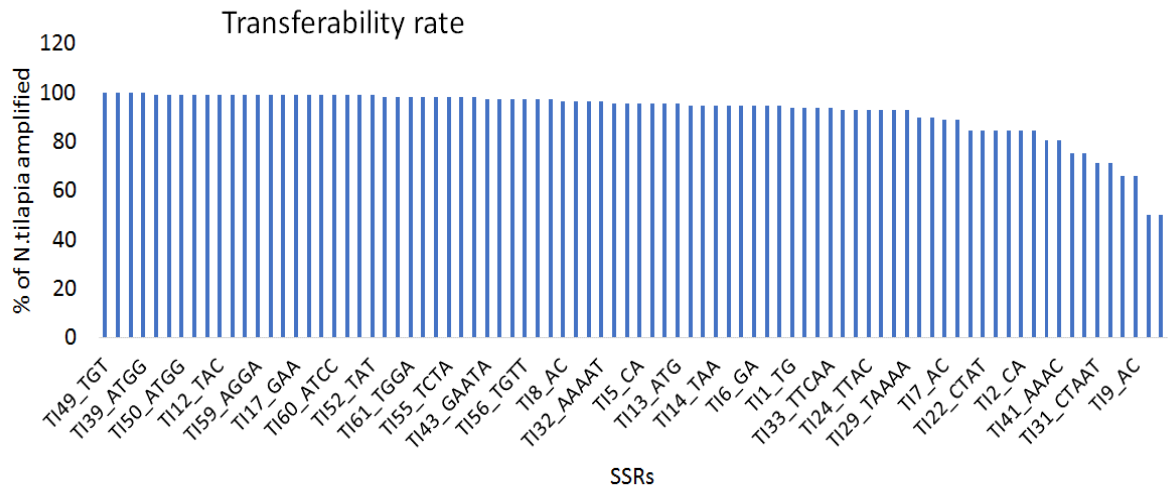


Appendix 2. Some sequence length results showing how the convectional method measured fragment length. The results are for the following microsatellite markers: TI12_TAC, TI9_AC, TI48_ATATA, TI41_AAAC, TI35_AAAAG, TI27_TTTG, TI29_TAAAA, TI2_CA, TI34_TCTCT

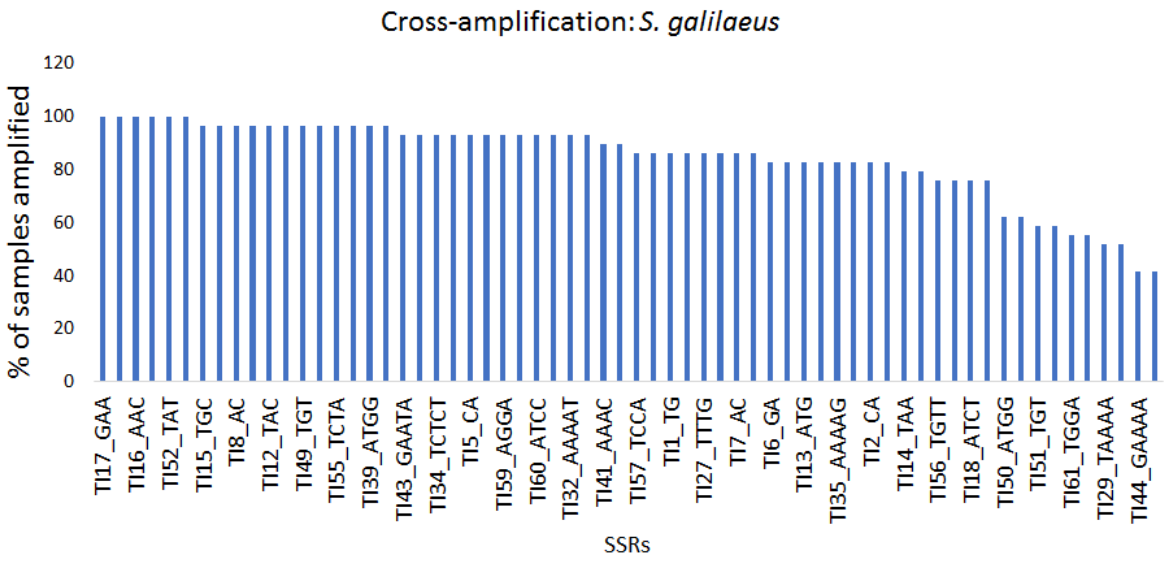
Appendix 3. Table of primers sequences that completely failed and eliminated from further tests.

Name	Primer sequence (5'- 3')	Repeat motif	Size (bp)
TI36_ATT	F: GCCGTAATGGAGCTGACAGA R: CCAAGATGTCGGCAAACTGC	$(ATT)_{17}$	160-210
TI37_TTA	F: GCATGCACTAAACCACGCAT R: CGAGACTGTGGCGGATTAGG	$(TTA)_{16}$	291-338
TI38_TAT	F: ACTCCACACAGTGAACTACTCT R: TGAGACTCTCACGTAGGCCA	$(TAT)_{14}$	80-121
TI45_ATATA	F: CCTGCTGAAGCTAAACCTGC R: TCAAAGGACATTATGGTCTGACT	$(ATATA)_8$	274-313
TI46_TAT	F: ACTCCACACAGTGAACTACTCT R: TGAGACTCTCACGTAGGCCA	$(TAT)1_4$	80-121

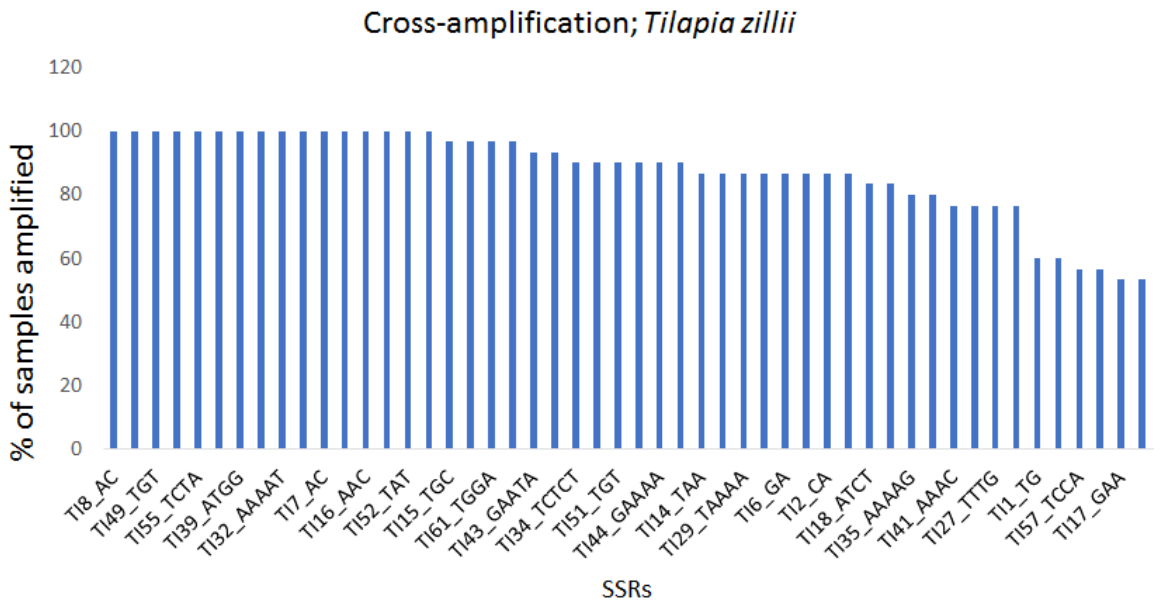
Transferability and cross-species amplification rates



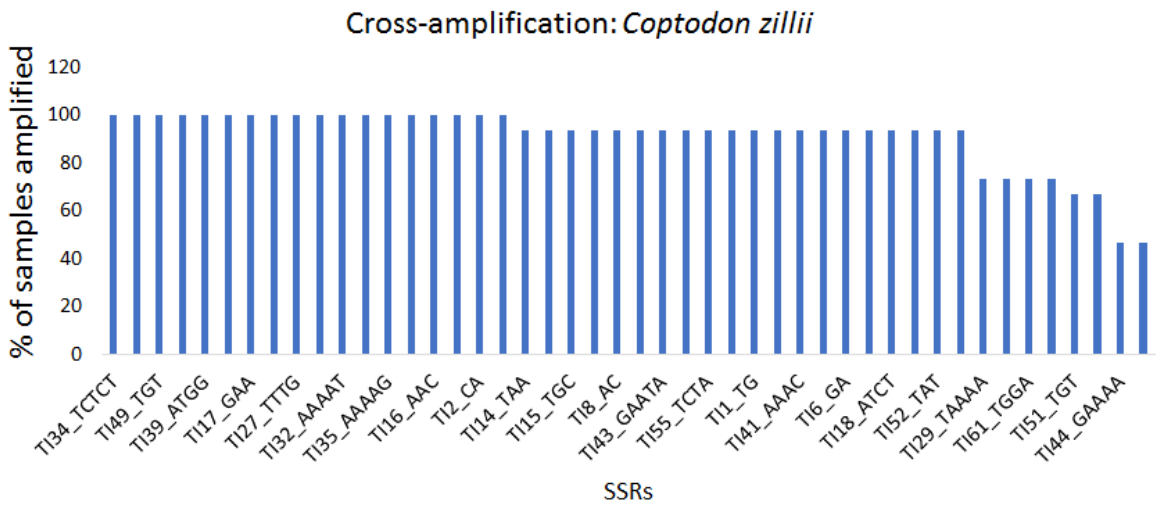
Appendix 4. Graph showing the transferability rate of the developed SSRs. The success rate was 100%, i.e., all the microsatellites were amplified in Nile tilapia, with the least amplifying in 50% of the samples.



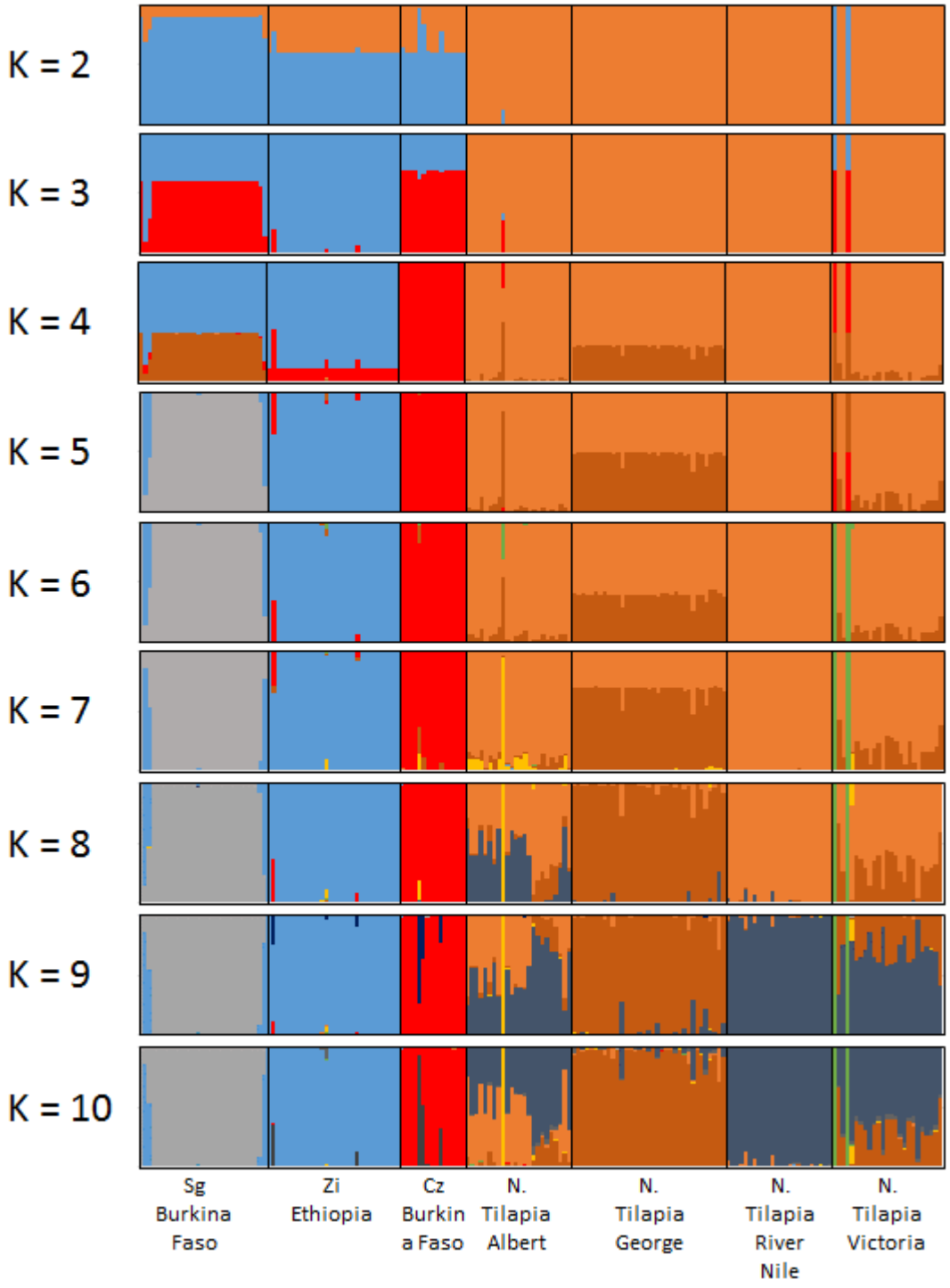
Appendix 5. Graph showing amplification rate in *Sarotherodon galilaeus*. Thirty two out of the forty-two developed SSRs were amplified



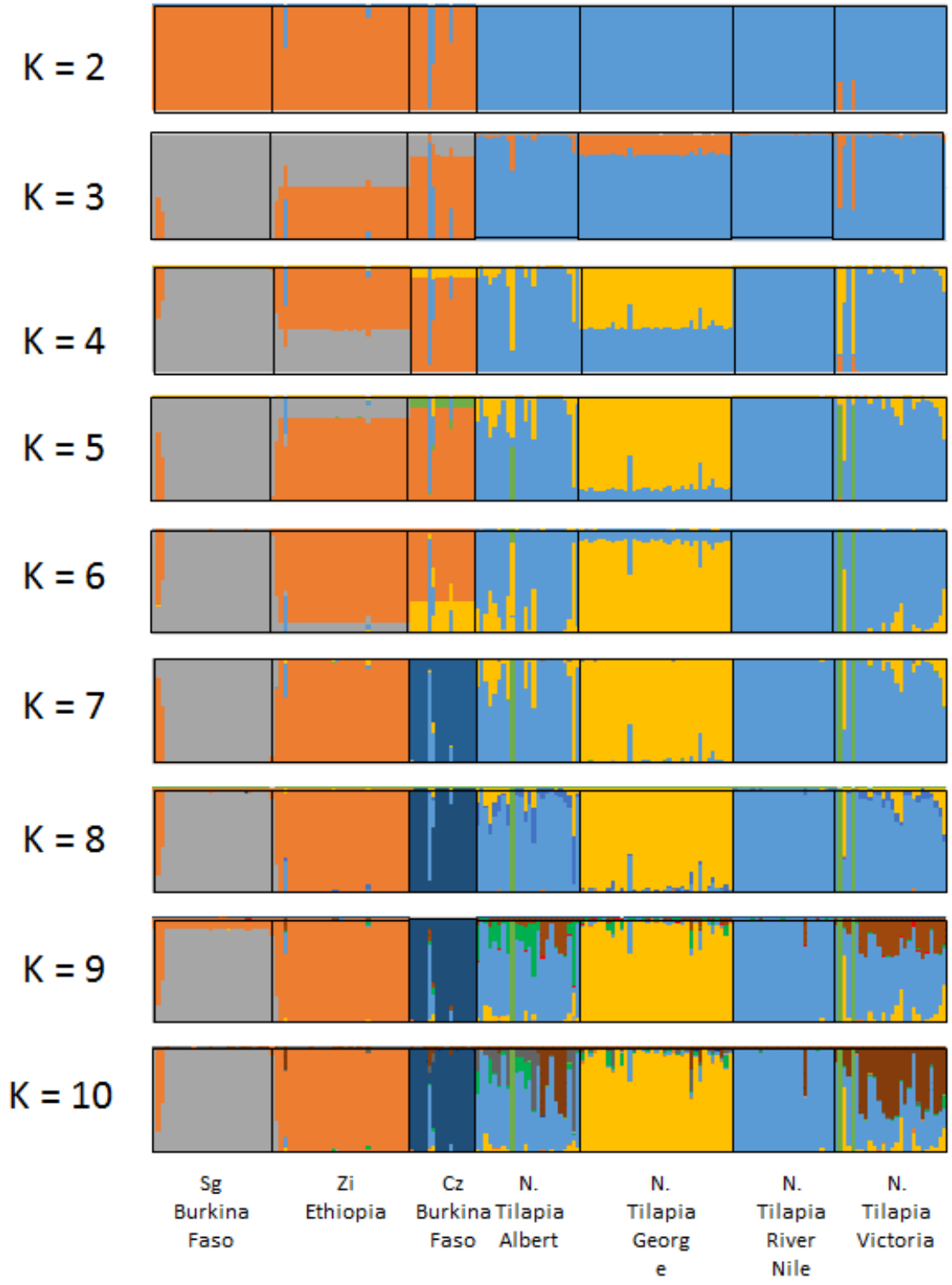
Appendix 6. Graph showing cross-amplification rate in *Tilapia zillii*. Twenty five of the 42 developed SSRs were amplified in *Tilapia zillii*



Appendix 7. Graph showing amplification rate in *Coptodon zillii*. Twenty three out of the forty-two developed SSRs were amplified here



Appendix 8. Bar plots showing the progression of K towards attainment of optimum K in Haplotype allele calling method in all populations. NB: The colours match to the colours in haplotype method above (Figure 12, haplotype method).



Appendix 9. Bar plots showing the progression of K towards attainment of optimum K in sequence length allele calling method in all populations. NB: The colours match to the colours in sequence length method above (Figure 12, sequence length)