SINGLE NUCLEOTIDE POLYMORPHISMs (SNPs) DISCOVERY IN THE TRANSCRIPTOMES OF MARBLED LUNGFISH (*Protopterus aethiopicus)* IN UGANDA

MARY NJERI MACHARIA, (BSc. WILDLIFE MANAGEMENT AND CONSERVATION)

A RESEARCH THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE MASTER OF SCIENCE DEGREE IN ANIMAL GENETICS AND BREEDING IN THE DEPARTMENT OF ANIMAL PRODUCTION, FACULTY OF VETERINARY MEDICINE, THE COLLEGE OF AGRICULTURE AND VETERINARY SCIENCES,

UNIVERSITY OF NAIROBI

September 2017

DECLARATION

I Mary Njeri Macharia, declare that this dissertation is my own work. It has not been submitted for a degree or any other qualifications in any other University.

Signature..... Date.....

This thesis has been submitted for examination with our approval as University supervisors.

Signature..... Date.....

Dr. Amimo Joshua Oluoch (PhD)

Department of Animal Production

The University of Nairobi

Signature..... Date.....

Dr. Rawlynce Cheruiyot Bett (PhD)

Department of Animal Production

The University of Nairobi

Signature..... Date.....

Prof. Morris Agaba

Research Chair, Genetics and Genomics

The Nelson Mandela African Institute of Science and Technology, Arusha, Tanzania

DEDICATION

I dedicate this thesis to my Father, Julius Macharia, Mother, Eunice Macharia and my siblings. Thank you for your encouragement.

ACKNOWLEDGEMENTS

I am greatly indebted to the following personalities and institutions for their contribution to my research described in this thesis.

I sincerely acknowledge my supervisors Dr. Joshua Amimo, Dr. Rawlynce Bett, Prof. Morris Agaba for their unconditional support, patience and guidance throughout this study.

I also extend my special thanks to the Biosciences Eastern and Central Africa of International Livestock Research Centre (BecA ILRI Hub) for funding this project.

I am grateful to Dr. John Walakira from National Fisheries Research Institute (NAFIRIRI) in Uganda for incorporating me in this ongoing project.

I am thankful to the BecA ILRI Hub; especially Joyce Njuguna, Dr. Stephen Opiyo, Senior scientist, Ohio State University, USA and other Bioinformaticians for their assistance in data analysis.

I am also grateful to Dr. Felix Kibegwa and Evan Cheruiyot Kiptoo from the University of Nairobi for their support in the bioinformatics analysis of the project work.

Thank you to my parents (Julius and Eunice Macharia) and all my siblings, who were always there for me, for their moral and financial support and encouragement throughout this time.

And above all, I give thanks to Almighty God for the gift of life and protection during the entire period of my research.

Table of Contents

DECLARATION ii
DEDICATION iii
ACKNOWLEDGEMENTS iv
LIST OF TABLES ix
LIST OF FIGURES x
LIST OF ABBREVIATIONS AND ACRONYMS xi
ABSTRACT xiii
CHAPTER ONE 1
1.0 INTRODUCTION 1
1.1 Background information 1
1.2 Problem statement 2
1.3 Justification 2
1.4 Objectives 3
1.4.1 General objective 3
1.4.2 Specific objectives 4
CHAPTER TWO 5
2.0 LITERATURE REVIEW 5
2.1 Fish production 5
2.1.1 Fish production in Africa and East Africa 5
2.1.2 Fish production in Uganda 6
2.2 Origin and distribution of the marbled lungfish 7
2.3 Significance characteristics of the marbled lungfish 8
2.3.1 The economic and social importance of marbled lungfish 9
2.4 Other lungfish species 10
2.4.1 Queensland lungfish (*Neoceratodus forsteri*) 10
2.4.2 The South American lungfish (*Lepidosiren paradoxa*) 10
2.4.3 The gilled African Lungfish (*Protopterus amphibius)* 10
2.4.4 The West African Lungfish (*Protopterus annectens*) 11
2.4.5 The Spotted African lungfish (*Protopterus dolloi*) 11
2.5 Genetic diversity 11

2.5.1 Genetic variation in marbled lungfish in Africa 12
2.5.2 Conservation of genetically threatened wild species stocks....13
2.6 Molecular markers in evaluating genetic diversity of populations 15
2.6.1 Restriction fragment length polymorphisms (RFLPs) 15
2.6.2 Random amplified polymorphic DNA (RADP) 16
2.6.3 Mitochondrial DNA (mtDNA)....16
2.6.4 Microsatellites....16
2.6.5 Amplified fragment length polymorphisms (AFLP)....17
2.6.6 Single Nucleotide Polymorphisms (SNPs)....17
2.7 Next Generation Sequencing (NGS)....19
2.8 The use of RNA-Sequencing technology in transcriptomics studies 19
2.8.1. RNA-Sequenced data 20
2.9 Bioinformatics analysis in genomic studies....20
2.10 Heterozygosity (H) 21
2.11 Single Nucleotide Polymorphisms (SNPs) genotyping 21
2.12 Polymerase Chain Reaction (PCR) - Restriction Fragment Length Polymorphism (RFLP) procedure for genotyping Single Nucleotide Polymorphisms 21
CHAPTER THREE....23
3.0 MATERIALS AND METHODS....23
3.1 Study sites 23
3.2 Sampling 24
3.2.1 Sample size determination....24
3.2.2 Sample collection....24
3.2.3 Sample dissection for RNA extraction....25
3.3 RNA extraction....25
3.4 Library preparation and sequencing 25
3.5 Bioinformatics analysis of the RNA-Seq data 26
3.5.1 Quality control of the raw RNA-Seq data 26
3.5.2 *De novo* assembly and reference mapping 27
3.5.3 SNP detection....28
3.5.4 Single Nucleotide Polymorphism annotation 28

3.6 Calculation of Heterozygosity (H) values 29
3.7 Calculation of Guanine-cytosine (GC) content 29
3.8 Population structure 29
CHAPTER FOUR 31
4.0 RESULTS 31
4.1 Transcriptomes of the marbled lungfish 31
4.1.1 Generation of high quality reads 31
4.1.2 *De novo* assembly and Blastn results of the *de no* assembled contigs. 31
4.1.3 Read mapping and visualization 34
4.2 Single Nucleotide Polymorphisms identification 36
4.2.1 Single Nucleotide Polymorphisms contigs annotation 37
4.3 A catalog of Single Nucleotide Polymorphisms discovered 39
4.4. Population structure 39
4.4.1 Pairwise population relatedness 39
4.4.2 Admixture structure for population analysis 40
4.4.3 Distance matrix for the marbled lungfish from the six lakes 42
4.4.4 Principal Component Analysis (PCA) 43
CHAPTER FIVE 45
5.0 DISCUSSION 45
5.1 Transcriptomes of marbled lungfish for SNP discovery 45
5.1.1 Single Nucleotide Polymorphisms selection based on the Heterozygosity values 48
5.1.2 Genetic diversity interpretation using the heterozygosity values 48
5.2 Population structure 49
5.2.1 Hardy Weinberg Equilibrium 49
5.2.2 Admixture structure 49
5.2.3 Principal Component Analysis and neighbor joining trees 50
CHAPTER SIX 51
6.0 CONCLUSION AND RECOMMENDATIONS 51
6.1 Conclusion 51
6.2 Recommendations 52
REFERENCES 53

APPENDICES 72
APPENDIX I 72
APPENDIX II 74

LIST OF TABLES

Table 3.1 Adapter sequences for every sample from each of the six lakes.....26

Table 4. 1Statistics of the percentages of assigned and unassigned contigs from the six studied lakes.....32

Table 4. 2 A summary of the gene identified from the annotation of the marbled lungfish contigs against the Uniprot and Tremble fasta databases.....38

Table 4. 3 The values of the Identity by State (IBS) across the six lakes 40

Table 4. 4 Summary of the statistics of the RNA-Sequenced data before and after the quality control.....72

Table 4. 5 The selected panel of Single Nucleotide Polymorphisms to guide in genetic diversity analysis and breeding purposes.....74

LIST OF FIGURES

Figure 2. 1 A schematic diagram showing the appearance of the marbled lungfish 9

Figure 3. 1 Map of Uganda showing the selected study lakes----------24

Figure 4. 1 Blastn results of the Bisina trinity results showing the similarity with known sequences on NCBI----------33

Figure 4. 2 Percentage of the query that matched to the marbled lungfish out of genus protopterus identified ----------34

Figure 4. 3 A read coverage plot on IGV for TR65600|c0_g1_i1 contig and the SNPs detected 35

Figure 4. 4 A summary chart of the genotype composition of the transitions and transversion SNP types identified----------36

Figure 4. 5 plots of genetic admixture proportions (Q) according to additional values of K (bis – Lake Bisina, edw – Lake Edward, geo – Lake George, kyo – Lake Kyoga, naw – Lake Nawampasa ; wam – Lake Wamala). ----------41

Figure 4. 6 The matrix distances between the selected six lakes----------42

Figure 4. 7 A neighbor joining tree for the six lake regions ----------43

Figure 4. 8 A PCA plot of the six study lakes in Uganda ----------44

LIST OF ABBREVIATIONS AND ACRONYMS

% - percentage

°*C* -degree Celsius

AC - Adenine /Cytosine

AG - Adenine/Guanine

AnGR - Animal Genetic Resources

AT - Adenine/ Thymine

BAM file (.bam) - the binary version of a SAM file

BLAST - Basic Local Alignment Search Tool

Blastn - Basic Local Alignment Search Tool against known nucleotides

Blastx - Basic Local Alignment Search Tool against known proteins

cDNA - complementary DNA

CG - Cytosine / Guanine

CT - Cytosine/Guanine

DNA - Deoxyribonucleic Acid

GC - guanine-cytosine

GT - Guanine/ Thymine

H- Heterozygosity

HWE - Hardy-Weinberg Equilibrium

IBS - Identical By State

IGV - Integrated Genome Viewer

Kg - Kilograms

Km - Kilometres

m^{2-}- square metre

Mins - Minutes

Ml - Milliliter

MtDNA - Mitochondrial DNA

NaHCO3 - sodium hydrogen carbonate

NCBI - National Center for Biotechnology Information

NGS- Next Generation Sequencing

PCA - Principal Component Analysis

PCR - polymerase chain reaction

pH - potential of Hydrogen

QC - Quality Control

RADP - Random Amplified Polymorphic DNA

RNA - ribonucleic acid

RT - Reverse Transcriptase

SAM - Sequence Alignment Map

SNPs - Single Nucleotide Polymorphisms

VCF - Variant Call Format

ABSTRACT

The marbled lungfish (*Protopterus aethiopicus*) is a significant fish species that contributes greatly to the world's fishery production. There has been limited genetic improvement of the species. This has necessitated the need to develop Single Nucleotide Polymorphism (SNPs) molecular markers that could be precise for use in their selective breeding and genetic diversity studies. Genomic selection based on informative SNP markers would play a major role in the shift to appropriate breeding strategies. This would improve productivity and conservation of the marbled lungfish natural stocks. The current reduced population size of the marbled lungfish would cause genetic diversity loss within the species. Availability of SNP markers for the marbled lungfish species is suitable for genetic improvement of the breeds selected for aquaculture use. In the present study, total RNA was extracted from 18 individual marbled lungfish from Lake Wamala, Kyoga, Edward, Nawampasa, George and Bisina. Three samples represented population within each lake. Library preparation was done using the Illumina protocol and quantified before being sequenced. The reads were *de novo* assembled to create a marbled lungfish transcriptomes reference genome that identified a total number of 5,961 SNPs. Of the 5,961 SNPs obtained, only 1979 SNPs had a heterozygosity value of 0.2 - 0.5 indicating relatively low genetic variation among the marbled lungfish population. One hundred and ninety eight (198) SNPs that had maximum heterozygosity values of 0.5 and flanking sequences of 140 base pairs of 40-60% Guanine-Cytosine content were selected as potential SNP panel for the lungfish studies. To facilitate the study of the population structure of the lungfish in Uganda, a total number of 4,565 SNPs data with a genotyping rate of 0.5 were selected. The Admixture, neighbor joining trees and the Principal Component Analysis (PCA) clustered the studied lakes into four groups: (a) Bisina (b) Edward and Kyoga, (c) Nawampasa and Wamala, and (d)

George. The 198 SNP panel selected could be used for monitoring the genetic diversity during conservation and management program of the wild stocks of lungfish. The research recommends further studies to validate the identified SNP panel for use in characterization of the marbled lungfish.

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background information

The marbled lungfish *(Protopterus aethiopicus)* are significant animal genetic resources (AnGR) in the ecosystem of many African countries (Walakira *et al*., 2012). They are valuable to a vast number of fish dependent communities but their population is currently reducing in number. In Uganda, the species have become of great interest for most communities as source of nutrients. Most of the wild marbled lungfish stock has been widely used for culture on the farms and are caught unsustainably at all sizes (Walakira *et al*., 2013). Recent studies show that marbled lungfish is appropriate for aquaculture as it can aestivate that enables it to survive for extended periods in little or no water (Walakira, 2013; Walakira *et al*., 2014). This is a unique characteristic desirable for aquaculture practices. But heavy reliance on the wild marbled lungfish necessitates the need to develop molecular markers that would guide in molecular breeding and diversity studies of the species. The SNP data would also be vital for genomic discovery of the marbled lungfish since its whole genome has not yet been sequenced and made available publicly. This research used marbled lungfish from lakes Wamala, Bisina, Edward, George, Wamala, Kyoga and Nawampasa that add significantly to fish production in Uganda (Walakira *et al*., 2012).

1.2 Problem statement

Lungfish are in the least concern species list of the International Union Conservation Nature (IUCN) red list (Fishbase team RMCA *et al*., 2016). The populations are rapidly declining due to practices like over exploitation, indiscriminate fishing, environmental degradation and large-scale use of wetlands for agricultural practices (Balirwa *et al*., 2003; Walakira *et al*., 2016). Furthermore, climate change continues to influence regional rainfall patterns and temperature regimes (Bisina, 2009). This directly affects both wild and cultured fish stocks as the reduced rainfall pattern cause increased habitat destruction (Barange *et al*., 2014) and lead to low dissolved oxygen that reduces the fish production (Musinguzi *et al*., 2016). Most marbled lungfish farmers in Uganda mainly access the seed from natural environments which further leads to wild lungfish population decline (Walakira *et al*., 2015). Appropriate intervention using advanced aquaculture technologies like the use of reliable molecular markers would ensure increased production of the marbled lungfish. Therefore, this study aimed to detect SNPs which could provide baseline information to guide farmers in the species breeding and diversity practices. As a result, there would be improved productivity to meet the demands.

1.3 Justification

Many molecular markers methods have been utilized in the identification of genetic variation across populations (Duran *et al*., 2009; Gautier *et al*., 2013). SNPs markers have been employed in the molecular ecology field (Morin *et al*., 2007), genetics studies in fisheries (Kochzius and universietiet, 2009) and the development of aquaculture programs (Liu *et al*., 2011). SNPs provide broad information on genetic variation across the species genome (Houston *et al*., 2014) which together with the phenotype and pedigree increases accuracy of selection during breeding.

SNPs have successfully been applied for speciation analysis in fish genome due to their occurrence throughout the genome (Liu *et al*., 2004).

The developments of suitable SNP assays have positively modified the genomic manipulation of the farmed fish (Yu *et al*., 2014). Atlantic salmon is an evident of successful discovery of true bi-allelic SNPs using the Next generation sequencing platforms (Houston *et al*., 2014). Such development of SNP molecular markers is critical for genetic improvement programs because they increase the economic profits of the aquaculture systems through selective breeding (Gjedrem, 2012). Efficient aquaculture practices means that every heritable and economically important trait, such as growth rate and resistance to disease, should be considered in the breeding goals. Therefore, a transcriptomics based sequencing approach employed in this study identifies a significant number of polymorphisms (Djari *et al*., 2013). The identified SNP markers could be utilized as a tool to enhance improved accuracy of selection. This increases the genetic gain and reduces the levels of inbreeding which is significant to population management. Consequently, the natural stocks could be protected through this intervention.

1.4 Objectives

The overall goal of this research is to contribute towards the development of improved breeding practices for marbled lungfish for better livelihoods through improved household nutrition, food security and income.

1.4.1 General objective

The aim of this research is to identify suitable SNP markers using marbled lungfish transcriptomes from Uganda, Africa.

1.4.2 Specific objectives

1. To catalog Single Nucleotide Polymorphisms in the wild marbled lungfish and identify the most useful SNPs.

2. To explore the population structure of the marbled lungfish across Lake Wamala, Kyoga, Edward, Nawampasa, George and Bisina using the generated SNPs data.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 Fish production

Fisheries and aquaculture are crucial sources of food, nutrition and revenue generation for hundreds millions of communities around the universe (FAO, 2016). According to FAO's new State of World Fisheries and Aquaculture report (2016), the global fish supply is 20 kilograms (kg) per person as recorded in 2014 and the number is expected to rise with time due to tremendous growth in aquaculture. The world's total fishery capture was 93.4 million tonnes in 2014, with about 11.9 million tonnes originating from inland waters and 81.5 million tonnes originating from marine waters (FAO, 2016). Fish is the highly traded food resource globally with over half of the fish exchange originating from the developing nations. There is increased potential of fish production globally due to availability of water sources and improved breeding strategies (Kobayashi et *al*., 2015). Therefore, the fishing industry has intervened in proving the status of certain fish stocks through various fishery management practices. This is geared towards promoted food security and adequate nutrient supply for the global population.

2.1.1 Fish production in Africa and East Africa

Fishing in Africa has improved from 4,175,000 tonnes in 2000 to 5,674,000 tonnes in 2014 (FAO, 2016). Out of this 91,000 tonnes and 284, 000 tonnes originate from the farmed fishes in 2000 and 2014 respectively. This shows vigorous increase in the aquaculture production. The East African region has various leading freshwaters lakes on the world that harbor considerable fishery resources. Lake Victoria is shared by Uganda (45%), Tanzania (49%) and Kenya (6%) (Nunan, 2014). It covers approximately 68,000 km^2 surface area and 3,400 km shoreline length.

Lake Victoria is a habitat for nearly 350 fish species. Some of the species include *Haplochromis*, Nile perch, *Mormyrus*, Tilapia, *Labeo*, Dagaa, *Protopterus*, *Clarius*, *Schilbe* and *Synodontis* caught for commercial and domestic purposes. Other lakes like Lake Tanganyika in Tanzania, Turkana in Kenya and Kyoga in Uganda offer excellent grounds for fishing. Aquaculture practices in the East Africa region is growing. According to (FAO, 2012) the countries need to embrace sustainable fishing practices and maintain the ecological balance in the river and the lake ecosystems. This will arrest the declining fish production in East Africa caused by indiscriminate fishing activities. For example, in 2011, Uganda inland fishing contributed approximately 437,415 tonnes of fish but declined to 407,638 tonnes in 2012. Kenya inland fishing is approximately 123,861 tonnes.

2.1.2 Fish production in Uganda

In Uganda, fishery is mainly in fresh waters lakes (Rutaisire *et al*., 2017). Lake Victoria and river Nile are dominant in fish production (Gordon *et al*., 2015). Over 365 fish species has been reported in all the Ugandan lakes; the most important being Nile tilapia, *haplochromis*, Sardine-like *Rastrineobola sp*, *Bagrus docmac*, African catfish (*Clarias gariepinus*) and lungfish (*Protopterus sp*) for commercial and subsistence exploitation (Ssebisubi, 2011).

According to the Department of Fisheries Resources (2004), aquaculture practices in Uganda yield about 15,000 tonnes of fish from the subsistence and commercial producers while others originate from the stocked community reservoirs and minor lakes. According to (Mwanja, 2007), Uganda has nearly 20,000 fish ponds each within an area of 500. The Ministry of Agriculture, Animal Industries and Fisheries of Uganda (2011) show that this occupation has increased from the 50 m^2 to 200 m^2 than it was early 1960s. The subsistence and upcoming commercial fish

farmers add to this production with approximately 1500 kg and 15000 kg respectively for every hectare per year.

According to NaFIRRI Annual Report (2009/2010), the demand for fish continues to increase in the market every year. This has attracted the government intervention for promoted production via adequate supply of aquaculture fisheries. The phenomenon has caused the shifting of 20-30 percent smallholder substance ponds into profitable small-scale production units. The farmers dwell mostly on the pond culture system though the cage culture system is slowly being adopted by the emerging commercial fish farmers (Walakira *et al*., 2015).

The National Aquaculture Sector Overview (2005) demonstrated that the major issue is inadequate technical support and management which leads to a high number of individuals practicing subsistence farming. Farmers use the limited resources to maximize their production. Nevertheless, Uganda has a high potential for increased aquaculture productivity since 20% of its land is occupied by water (Walakira *et al*., 2012). Therefore, aquaculture practices with genetically improved fish seeds would boost fish productivity levels (Kjaer *et al*., 2012).

2.2 Origin and distribution of the marbled lungfish

Lungfish are lobe-finned freshwater fish that constitute the Dipnoi subclass (Greenwood, 1966). The marbled lungfish belong to kingdom animalia, phylum Chordata, Class Sarcopterygii, order Lepidosireniformes and family Protopteridae. The African lungfish species present today includes *Protopterus amphibious*, *Protopterus annectens, Protopterus aethiopicus*, and *Protopterus dolloi* (Biscotti *et al*., 2016). Marbled lungfish (*Protopterus aethiopicus*) inhabits African countries which include Kenya, Uganda, Ethiopia, Sudan, and the Democratic Republic of Congo (Gosse, 1984). Their habitats include River Nile and lakes like Kyoga, Albert, Edward, No, Nabugando, Victoria, Tanganyika, Wamala, Bisina, Nawampasa and George among others

(Gosse, 1984). Marbled Lungfish have also been introduction into areas outside their natural distribution. The first introduction was Lake Mohasi in Upper Akagera system, Rwanda, from Lake Edward in Uganda around 1988-1989. It has since dispersed to various sections of upper Akagera (De Vos *et al*., 2001). The next introduction was in Lake Baringo, Kenya, in 1975 from the Winam Gulf area of Lake Victoria (Mlewa *et al*., 2006). Other living lungfish species include the *Lepidosiren paradoxa* and *Neoceratodus forster*.

2.3 Significance characteristics of the marbled lungfish

Marbled lungfish are elongated, cylindrical and possess even body structure with highly embedded scales (Bailey, 1994). They have a long tail tapered at the end. The adult marbled lungfish live in swamps, riverbeds, floodplains and river deltas throughout its range while the young members of the species often live in between the roots of papyrus plants (Gosse, 1984). These habitats are likely to dry up in the dry season though the marbled lungfish have evolved as obligate breathers (Bailey, 1994). They can, therefore, endure extended periods out of the water where they remain holed up in burrows in the dry mud. Their air bladder has thus developed into lungs to breath air (Greenwood, 1958).

Their life cycle involves external fertilization, and they breed during flood season (Bailey, 1994). The males prepare a pit nest that can be used by one or more females to lay their eggs (Greenwood, 1966). The female leaves the nest to be guarded by the male for eight weeks where the male regularly aerates the water in the nest for the survival of newly-laid eggs (Greenwood, 1966). Lungfish have a predatory behavior, with the diet of adults consisting primarily of mollusks though they also eat small fishes and insects (Gosse, 1984). The diet of juveniles consists almost entirely of insects (Witte and Winter, 1995). The adults can also survive an extended period without food as they are autophagy (Babiker, 1979).

Figure 2. 1 A schematic diagram showing the appearance of the marbled lungfish

Source: http://www.tropical-fish-keeping.com/wp-content/uploads/2015/10/Marbled-Lungfish-Protopterus-aethiopicus...jpg

2.3.1 The economic and social importance of marbled lungfish

Marbled lungfish are highly exploited for both commercial and subsistence purposes in most African countries that make them highly threatened (Walakira, 2015). Both cultured and wild fish species contribute to fish production in Uganda. The fishery industry is among the most highly valued sectors in the economy contributing approximately 6% of its GDP with 2.8% channeled to national accounts (Ssebisubi, 2011). According to the Ministry of Agriculture, Animal Industries and Fisheries (2011), the marbled lungfish adds to this with a production of 5-10% towards total fish catch. It contributes to foreign exchange income, source of animal

protein, has social, cultural values and provides income for many livelihoods through employment (Garner *et al*., 2006).

2.4 Other lungfish species

2.4.1 Queensland lungfish (*Neoceratodus forsteri*)

Queensland lungfish is prevalent to Australia and its fossils show that it has not changed for approximately 100 million years (Allen *et al*., 2002). It is a highly primitive air-breathing lungfish. It can stay for prolonged period outside water when kept moist though it cannot live on entire water deprivation like the African lungfish.

2.4.2 The South American lungfish (*Lepidosiren paradoxa*)

The South American lungfish is located in swamps, lower Parana River basins and the slowly moving waters of the Paraguay and Amazon in South America. It is an obligate air-breather and is speckled with gold on a black background when immature. This lightens to a gray or brown color when it matures. It has premaxiallary and maxillary bones and has an autostylic jaw suspension. It possesses an elongated body that can reach up to 125 centimeters. The species have lean and narrow pectoral fins while its pelvic fins seem large outlying on its back. These fins connect to the shoulder by one bone (Shubin, 2008). They have reduced gills that are non-functional in the mature fish (Bruton, 1998).

2.4.3 The gilled African Lungfish (*Protopterus amphibius)*

According to the Fishbase.org, *Protopterus amphibius* is found in the East Africa. It has about 44cm length that makes it the smallest extant lungfish around the globe. It is evenly blue or slate grey with small black dots as well as a pale grey belly.

2.4.4 The West African Lungfish (*Protopterus annectens*)

The West African Lungfish is located in West Africa. It is characterized by a large snout, tiny eyes, two pairs of elongated filamentous fins and eel-like body. Its pectoral fins are thrice its head length and have a basal fridge. The pelvic fins appear almost double its head length.

2.4.5 The Spotted African lungfish (*Protopterus dolloi*)

The Spotted African lungfish inhabits Africa in the Kouilou-Niari and Congo River basins of the Republic of the Congo, and Ogowe River basin in Gabon. It covers itself with desiccated mucus layer for aestivation on land (Brien, 1959). It can reach up to 130 long.

2.5 Genetic diversity

Genetic diversity describes the total number of genetic traits in a species genetic makeup (Avise, 1989). It serves as a means for the populations to adjust to the changes in the environments. Increased variation shows a higher likelihood that individuals will survive and produce alleles suitable for the environment. Hence, the population will survive for many generations due to the success of such individuals. Diversity within species helps in maintaining the diversity among species (Avise, 2012). Changes in the genetic diversity cause loss of the biological diversity. A higher level of genetic diversity is linked to a species' adaptive capability while the fitness of the population is closely linked to its heterozygosity (Reed and Frankham, 2003). Inbreeding is linked to genetic diversity loss that reduces the reproductive fitness of any population. This is so because of the reduction in its heterozygosity from the mating of individuals with close genetic relatedness (Frankham *et al*., 2002). Newman and Pilson (1997) explain that this situation might eventually cause inbreeding depression in small population sizes threatening their survival.

Genetic diversity is significant during the selection of animals in breeding programs to improve the population as only individuals with superior traits are selected as parents for the next

generation. The information is crucial for germplasm conservation and the species identification (Ellegren and Galtier, 2016). A population indicating increased genetic diversity has elevated chance of large potential genetic gain in every generation (Wright, 1922). Reduced genetic variation shows high level of homozygosity of the population (Peischl and Excoffier, 2015). Too much inbreeding reduces the genetic diversity of the populations. Inbreeding coefficient describes the chances where two alleles positioned at a specific locus, randomly chosen, are similar by descent (Wright, 1922). Genetic variation can be indicated by change within the species sequence (Duran *et al*., 2009).

Maintaining biodiversity genetic diversity has been promoted by several agencies such as the Rio Conference that advocates for sustainable development goals through the conventions on the conservation of biodiversity, desertification and enhanced measures to discourage ineffective climate change impacts (Hens, 1996). The activities of these conventions operate in similar ecosystems and addresses interdependent challenges (Malanczuk, 1995). The Convention on biological diversity endeavors to conserve the biological diversity, efficiently utilize its constituents and reasonably allocate the returns of the genetic resources. It also involves the UN Convention to Combat Desertification Conference that addresses drought issues in various countries.

2.5.1 Genetic variation in marbled lungfish in Africa

Marbled lungfish possess the biggest genome of every known vertebrate, at approximately 130,000 Mb in size (Thomson, 1972). The established karyotypes of marbled lungfish are identified as diploid number of 2n = 28 chromosomes (Omer and Abukashawa, 2012). Information on extant variability among marbled lungfish population is not available. Marbled lungfish genetic composition has been evaluated in the species found in Lake Victoria and Lake

Baringo in Kenya using their phenotypic features (Garner *et al.*, 2006). The study found increased genetic variation in the mitochondrial control regions in the sampled population from Lake Victoria, Kanyaboli and Nabugabo (Garner *et al*., 2006). In Lake Baringo, there was no mtDNA genetic variation in the population as only three individuals had been introduced. Walakira, *et al*. (2016) determined the genetic diversity of the marbled lungfish in Uganda using analysis of D-loop sequences. The research explored on the nucleotide diversity and haplotype diversity (Walakira, *et al*., 2016). Another study on the Australian lungfish developed microsatellite markers using samples Burnett and Mary Rivers to determine the population genetic structure (Hughes *et al*., 2015). The study found that the population had average heterozygosity values of 0.39 that implied high genetic diversity among the study populations.

2.5.2 Conservation of genetically threatened wild species stocks

The Global Plan of Action for animal genetic resources (AnGR) and the Interlaken Declaration outlines the successful approaches for conservation, sustainable exploitation and development of AnGR for improved food productivity (Hoffmann, 2010). It targets increased food security, improved dietary and increased rural growth. The Interlaken Declaration acknowledges the sovereignty rights over the AnGR for the production of food (Mowbray, 2010). It ensures that individual states help in the conservation plans of the genetic resources in within and across states. The Interlaken Declaration recognizes that the present reduction in animal genetic resources can impede the efforts towards food security (Mowbray, 2010). This calls for further efforts to protect, build up and suitably exploit the accessible resources. The global Plan of Action suggests certain actions to overturn the continuing patterns of exhaustion and under-consumption of animal genetic resources. It entails Strategic Priorities for Action contributing

towards international efforts to boost food security and guarantee sustainable expansion (Hoffmann, 2010).

Conservation is a major Strategic Priority Area that is of interest in global plan of action for AnGR (Hoffmann, 2010). Increased human population over the past years has greatly led to drastic habitat loss, land degradation and conversion of wetlands to other agricultural activities. This has led to reduced or extinction of some vital species like the Tecopa Pupfish from California in 1970 (Noecker, 1998); (Miller *et al*., 1989); the species became extinct due to the interference with the hot springs of the Amargosa River (Levitt, 1981).This has over the past years been of concern where molecular markers have been developed as conservation tool. Vulnerability of the species also results from the stochastic factors that include both biotic like predation, competition and pathogenic agents and abiotic factors which includes droughts and floods. Others like population stochasticity may result from low birth and death rates, and improper sex ratios. Inbreeding depression and population bottlenecks are genetic stochasticity which results due to lowered species gene frequency. As inbreeding affect animals' fitness and survival various molecular techniques have assisted developing breeding technologies that reduce the level of inbreeding among and within populations (Avise, 2012). Both wild and captive species are prone to inbreeding depression though wild stocks with very low genetic diversity are at a very high risk of extinction (Saccheri *et al*., 1998). It is crucial to define management measures within species, identify invasive species and levels of inbreeding to conserve a species of interest.

Genetic variation within organisms is significant as it helps in their conservation and management strategies. It reveals their response to changing environmental conditions and other anthropogenic factors (Avise, 1989; Scheiner, 1993). Maintaining genetic diversity of a species

ensures food security for a long term. Hanotte and Jianlin (2005) illustrate that maintained genetic diversity helps preserve certain social-economic and cultural values of the species and prevent species extinction. Conservation of the genetic diversity of a wild stock would include selection and breeding strategies imposed for aquaculture of a species to improve the productivity (Boettcher *et al.,* 2010). This has resulted in a wide variety of molecular markers being employed as tools that provide baseline information for conservation of a species, for example, population phylogeny, inbreeding coefficient and the frequencies of genotypes.

2.6 Molecular markers in evaluating genetic diversity of populations

Genetic diversity can be determined using various genetic, phenotypic or environmental parameters (Ellegren and Galtier, 2016). In most molecular fields, extensive application of molecular markers has helped in population genetic diversity studies and in identification of imperative production traits. Molecular techniques available for genotyping include:

2.6.1 Restriction fragment length polymorphisms (RFLPs)

RFLPs approach involves digestion of double-stranded DNA with atleast one restriction endonucleases to produce DNA segments which are separated by their molecular weight through electrophoresis (Rasmussen, 2012). The created bands are scored to derive the frequency data for genetic diversity analysis. RFLP has several disadvantages. It requires multiple steps and takes weeks to produce outcomes compared to PCR methods. It also needs a large DNA sample and the isolation could be laborious and time-consuming (Skolnick and White, 1982). It is disadvantaged due to the low level of heterozygosity with many being diallec thus less informative for assessment of genetic variation. They do not have a resolving power when populations of study are closely related because variations results from mutations at the restriction regions which are very low (Rasmussen, 2012).

2.6.2 Random amplified polymorphic DNA (RADP)

RADPs are DNA polymorphisms group that are considered easiest to detect. They are constructed based on the PCR amplification of arbitrary DNA fragments with individual short primers with eight to ten base pairs of unpredictable sequence (Lynch and Milligan, 1994). Agarose electrophoresis reveals highly variable band patterns with every arbitrary primer generating various patterns of bands which are scored to produce information on individuals. It is useful in estimating genetic variability and in trait mapping (Lynch and Milligan, 1994).

It is difficult to evaluate evolutionary histories because RAPD data does not clearly outline how the genetic variation is generated. It has another limitation where the obtained fingerprint patterns are uncertain and it has a dominant inheritance where the heterozygote and homozygotes cannot be distinguished. However, this technology is best suited for DNA fingerprinting (Kumar and Gurusubramanian, 2011).

2.6.3 Mitochondrial DNA (mtDNA)

Mitochondrial DNA refers to a greatly conserved molecule with genes densely arranged together where some overlap (Anderson *et al*., 1982). It is maternally inherited. It has non-coding region, D-loop, which is indicated to evolve at higher rate than the other mtDNA parts (Aquadro *et al*., 1984) . It is preferred when detecting variations between individuals at mtDNA level in research studies. MtDNA possesses a high copy number for each cell and the mutation rates are high (Aquadro *et al*., 1984).

2.6.4 Microsatellites

Microsatellites are repeats of two to six base-pair nucleotides in the genomic DNA which are detected by gel electrophoresis (Ellegren, 2004). They are scored to provide genetic assessment frequency data (Vos *et al*. 1995). They are useful in differention of closely related populations.

Microsatellites show that they are easy to identify and polymorphism detection requires small amount of DNA (Govindaraj *et al*., 2015). They are highly informative with many alleles and have high reproducibility (Ellegren, 2004).

2.6.5 Amplified fragment length polymorphisms (AFLP)

AFLP require PCR amplification for detection of restriction fragments (Vos *et al*., 1995). Digestion of the genomic DNA is by two different restriction endonucleases. The resultant fragments are subjected to PCR amplification. The procedure is advantageous since little quantity of DNA is required and its fingerprint traces are greatly reproducible consisting of several markers. Therefore, they are better for genetic studies of individuals with close genetic relatedness compared to other techniques like RAPDs (Russell *et al*., 1997). They provide significant information on probes with multiple locus and have high degree of resolution thus is reliable for many genetic studies (Yu *et al*., 2014).

2.6.6 Single Nucleotide Polymorphisms (SNPs)

SNPs occur at differing frequencies in various chromosome sections and may exist within coding and non-coding gene sections (Jiang, 2013). SNPs may be grouped in accordance to their nucleotide substitutions, such as transitions where purine changes to another purine and pyrimidine to pyrimidine. The transversions means pyrimidine change to purine and vice versa. Genetic studies in molecular fields using SNPs would provide significant insight into associations, movement and evolution of natural populations (Morin *et al*., 2004; Collins *et al*., 1997). Relative to other markers, SNPs have a higher advantage (Slate *et al*., 2009); (Diopere *et al*., 2013). The benefits include;

1. They occur in large numbers and are widely dispersed throughout the complete genome.

2. They are very stable genetically with high accuracy as they do not alter much from one generation to another thus are reliable and efficient for population studies.
3. They ensure quick high-throughput genotyping.
4. They are co-dominance thus it is easy to distinguish heterozygote from homozygote

SNP markers have a disadvantage where they present less information in comparison with microsatellites DNA markers due to their biallelic nature. Thus, more SNP numbers are required for accurate and precise comparison of genetic variation in a population as it has only two alleles compared with such other markers with many alleles (Liu *et al*., 2011).

2.6.6.1 Single Nucleotide Polymorphisms in determination of species genetic diversity and breeding

RNA sequenced data has been widely used to generate million numbers of reads that are used for transcriptomics analysis of individuals (Mutz *et al*., 2013). These reads have been utilized for the detection of SNPs and to explore some of the genes involved from the specific dataset (Ekblom and Galindo, 2011). The identified variations among species represent their genetic diversity which would have a significance impact on their future resilience and adaptive capacity. More conservation genetics have been successful from SNPs discovered using Next Generation Sequencing technologies (Kumar *et al*., 2012).

A population indicating high levels of genetic variation has a significant chance of large potential genetic gains in every generation (Caballero and Toro, 2002). Too much inbreeding reduces the genetic diversity of the populations. Inbreeding coefficient describes the chances that two alleles at a indiscriminately selected locus are similar by ancestry (Wright, 1922). SNP markers, therefore, act as a valid tool used in genotyping and for selective breeding programs to improve the population as only individuals with superior traits are chosen (Varshney *et al*., 2016).

2.7 Next Generation Sequencing (NGS)

Next Generation Sequencing (NGS) platforms entail the Roche 454, Illumina and ABI SOLID approaches (Harismendy *et al*., 2009). Every NGS platform generates a particular pattern showing the sequence coverage consistent between the samples. Notably, they help determine more than 95% of variant alleles (Kilian and Graner, 2012). The decreased costs related to the NGS platforms has promoted the development of SNP markers from the non-model species transcriptomes for genetics and genomics research (DeFaveri *et al*., 2011).

SNP discovery research based on NGS derived data has a variety of approaches though this study utilizes the pooled RNA-Seq on all the samples to detect a vast number of polymorphisms that are then subjected to validation (Lamaze *et al*., 2012). The NGS has relied heavily on the transcriptomes for the biodiversity research since the transcripts represent extensive information of the species genome (Gayral *et al*., 2011).

Illumina NGS sequencing approach covers a greater read depth than 454 sequencing platform and enhances the discovery of true SNPs (Yu *et al*., 2014). Despite NGS analysis having more advantages compared to Sanger sequencing, it faces sequencing errors which require bioinformatics computations to allow separation of true polymorphisms from machine artifacts (Helyar *et al*., 2011). It has largely involved filtering of the detected SNPs using the read depth coverage, and the heterozygosity degrees.

2.8 The use of RNA-Sequencing technology in transcriptomics studies

Transcriptomics is the study of all transcripts in a specific cell at a specified developmental phase or physiological circumstance (Wang *et al*., 2009). High-throughput sequencing methods have supported research of dynamic transcriptomes as it reveals RNA presence in a genome at certain time (Qian *et al*., 2014) through RNA-sequencing. It has enabled study of various

biological processes in fish like development, growth traits and diseases. In other studies, such transcriptomics data have helped identify breeding pedigree, species genomic divergence and to identify relationship between genotypes and phenotypes of these organisms (Feltus, 2004). Transcriptomes have also enhanced study of the level of expression of genes in a whole genome and in annotation of these genes (Costa *et al*., 2010).

2.8.1. RNA-Sequenced data

RNA sequenced data is a significant tool that helps to analyze species transcriptomes (Wang *et al*., 2010). A sample RNA could produce millions of reads for use in transcripts *de novo* assembly, gene annotation and the discovery of genomic structural polymorphisms (Maher *et al*., 2009). It has an excellent dynamic range of expression degrees for transcript discovery, and a greater coverage is essential to detect rare transcripts (Qian *et al*., 2014). RNA-seq can be applied in deep sequencing technologies, that is sequencing a genomic region several times, to precisely measure the transcripts' intensity and their isoform compared to other procedures (Wang *et al*., 2010). In other studies, RNA-seq remains important to identify SNPs directly linked to traits of interest like growth and those associated with certain disease like Aeromonas *sp*., cestodes, and Aspergilus *sp* (Yaez *et al*., 2014). The synonymous and non-synonymous SNPs could have an impact on the protein activity directly (Yu *et al*., 2014).

2.9 Bioinformatics analysis in genomic studies

Many sequencing centers are generation of large biological data that necessitates the need to use various bioinformatics pipelines for the computational analysis and practical interpretation of available data (Fallis, 2013). Approach towards certain biological information depends largely on the set objectives for each study. Bioinformatics has led to great success of genomic studies,

for example, in non-model species it allows contigs assembly for detection of SNPs. Bioinformatics workflows consists of a variety of tools that are easy to follow the guide.

2.10 Heterozygosity (H)

The level of heterozygosity of markers is used to define the quality of a polymorphism (Reed and Frankham, 2003). Heterozygosity is vital in determining the suitable SNPs for use as a breeding tool. The locus heterozygosity defines the likelihood for an individual being heterozygous for a specific locus in a given population (Gregorius, 1978).

H is calculated as

$$H = 1 - \sum_{i=1}^{l} P_i^2$$

P_i represents the frequency for the ith allele among the total of 1 allele.

2.11 Single Nucleotide Polymorphisms (SNPs) genotyping

SNPs genotyping is the process of measuring the genetic variations of SNPs (Perkel, 2008). It can incorporate phylogenetic inferences based on genetic distance methods and allele frequencies. Clustering can be done using Principal Component Analysis (PCA) (Holland, 2016) or Admixture algorithms (Alexander and Lange, 2011) when there is correct choice of similarity matrix. Allele frequencies can also be used to estimate the size and change of a population.

2.12 Polymerase Chain Reaction (PCR) - Restriction Fragment Length Polymorphism (RFLP) procedure for genotyping Single Nucleotide Polymorphisms

SNP genotyping is a vital procedure in genetic studies of diverse populations which tends to show short homologous DNA sequences variations (Perkel, 2008). Genotyping allows the researcher to explore the detected polymorphisms in the DNA on a molecular level which has

been highly facilitated by the NGS platforms (DePristo *et al*., 2011). Genotyping help to analyze the multiple SNP markers and allows detections of the outliers which would give a vital insight into the functional result of the genetic alterations generated (Shen *et al*., 2010). Genotyping could be on the whole genome or the targeted sequences based on the research purpose.

PCR-RFLP technique has helped to detect both the intraspecies and interspecies genetic variation (Rasmussen, 2012). SNP assays that results from an RFLP assay requires that the chosen restriction enzymes recognizes only an individual SNP containing sequence (Zhang *et al*., 2005). The restriction enzymes cuts at specific positions resulting into small sequence fragments of varying lengths and then separated through Agarose gel electrophoresis and are then transferred to a membrane (Rasmussen, 2012). Every single fragment length is considered as an allele which is utilized in genetic studies. The technique exploits the fact that micro variations such as SNPs are consistently linked with disrupting or creation of a restriction enzyme recognition enzyme site (Narayanan, 1991).

CHAPTER THREE

3.0 MATERIALS AND METHODS

3.1 Study sites

Uganda is located in East Africa and is approximately 241,038 square kilometers. Water covers 43,938 square kilometers and the land occupying 197,100 square kilometers. The natural waters cover nearly 18% of the total Ugandan area. Fisheries are significant for commercial and subsistence livelihood. The selected study sites were fresh water lakes which included Lakes Wamala, Nawampasa, Kyoga, Bisina, Edward, and George. The species of interest is the marbled lungfish. Lake Wamala is located in Central Uganda across Mubende, Mityana, and Mpigi areas in the Central Uganda. River Kibimba drains Lake Wamala into Katonga River, which in turn flows into Lake Victoria covering a total area of 250 square kilometers (FAO, 1990). Lake George and Edward are located in the extreme West of the country joined by Kazinga channel. Lake Kyoga is to the North of Lake Victoria in Central Uganda, 914 meters above sea level. Lake Kyoga has shallow swampland fingerling extensions that include Lake Kwania, Lake Bisina, and Opeta. According to Ramsar Convention on Wetlands (2013), Lake Bisina wetland is found in Eastern Uganda and has been listed as Wetland of International Importance under Ramsar Convention since 2006. Ramsar Convention aims at conservation and wise exploitation of all the wetlands by the local and national actions along with global collaboration to reach sustainable growth (Matthews, 1993). This means that Lake Bisina is an important wetland that requires conservation and wise use. Lake Bisina is recognized by Birdlife International as Important Bird Areas and provides a vital habitat for threatened plants and animals (Bisina, 2009). It drains into Lake Kyoga. Lake Nawampasa is situated in Eastern

Uganda between Nkone, Buyumba, and Irundu nearby to Namasajeri Island 1034 meters above sea level.

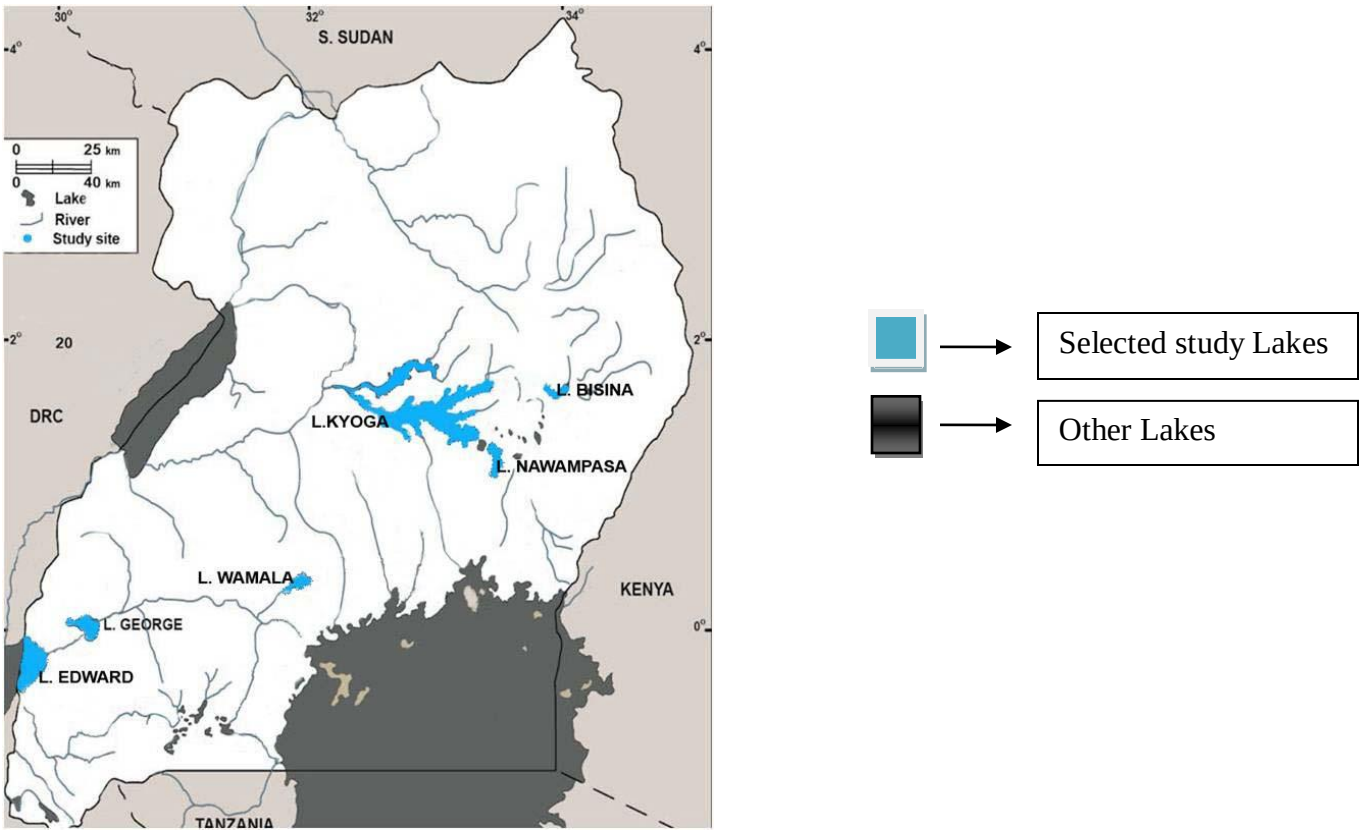


Figure 3. 1 Map of Uganda showing the selected study lakes

3.2 Sampling

3.2.1 Sample size determination

A total number of 18 whole fish were sampled from six lakes in Uganda (Bisina, Edward, George, Kyoga, Nawampasa, and Wamala) shown in Figure 3.1. Each lake was represented by three whole fish samples.

3.2.2 Sample collection

Fish samples were harvested using basket traps and conditioned for handling in *happa* materials at the shores of the lake.

3.2.3 Sample dissection for RNA extraction

Fish samples were anesthetized with 100mg/ml Tricaine methane sulfonate (MS-222) buffered with 0.2 ml $NaHCO_3$, pH = 7 for handling. Clips of the left pectoral fin were aseptically collected from each fish sample and transported in Qiagen RNA*later* solution at room temperature (26-29°C) to BecA-ILRI-Hub and kept at -80°C.

3.3 RNA extraction

RNA was extracted from the fin tissue of the eighteen samples. The fin tissues were crushed into powder under frozen temperatures using Liquid Nitrogen. Homogenization was done using TRizol Reagent 1ml to 100mg of grounded tissue and mixed systematically by pipetting repeatedly (Portillo *et al*., 2006). Approximately 0.2 ml of chloroform was added for each ml of the TRizol Reagent and then shaken vigorously for about 15 secs. The mixture was incubated at room temp (15-30^{o}C) and centrifuged at 1200 g for 10 mins at 4^{o}C. Later the aqueous phase was transferred to a 1.5 ml micro centrifuge tube. Isopropyl alcohol was added equally and mixed gently. Incubation followed at room temperatures for 10 minutes. The RNA was precipitated forming a gel-like pellet on side. Centrifugation was done at 1200 g for 15 mins and the supernatant was removed and discarded. The precipitated RNA was washed by adding 1ml of 75% ethanol and incubating it for 10 mins. The supernatant was discarded and the RNA pellet was air-dried at room temperature for 15-20 mins. The RNA pellets were then dissolved in 110 µl RNAase free water. An aliquot of 10 µl was used for quality control check, quantity and agilent analysis using Agarose Gel Electrophoresis (Masek *et al*., 2005).

3.4 Library preparation and sequencing

Six libraries were made using Illumina protocol (Illumina, 2011). Each of the pools was digested using the enzymatic method. This gave fragments of 200 - 250 base pairs which was the desired

length. RNA was fragmented before converting it into cDNA through controlled digestion of the RNA using magnesium. Adenylation, ligation and reverse transcription (RT) followed. Oligonucleotide adapters were attached to the ends of the target sequences as shown in Table 3.1. Time adjustments were made regularly during the digestion to ensure desired lengths of the RNA library sections. Finally, the libraries were quantified for sequencing. Sequencing was done using Miseq sequencer to provide raw RNA-seq data.

Table 3.1 Adapter sequences for every sample from each of the six lakes

Lake	Sample Index	ADAPTORS
BISINA	1	CGATGT
	2	TGACCA
	3	ACAGTG
EDWARD	4	GCCAAT
	5	CAGATC
	6	TAGCTT
GEORGE	7	GGCTAC
	8	CTTGTA
	9	AGTCAA
KYOGA	10	AGTTCC
	11	ATGTCA
	12	CCGTCC
NAWAMPASA	13	GTCCGC
	14	GTGAAA
	15	GTGGCC
WAMALA	16	GTTTCG
	17	CGTACG
	18	GAGTGG

3.5 Bioinformatics analysis of the RNA-Seq data

3.5.1 Quality control of the raw RNA-Seq data

After the data had been derived from Illumina sequencing, the "fastq" text files were downloaded and the low quality bases and the adapter sequences were removed using fastx-clipper. Quality

control followed using fastQC/0.11.3 toolkit (Patel and Jain, 2012) to find out the number of the forward and the reverse sequence, the guanine-cytosine content (GC) percentage before and after quality control and their lengths. The reads with a Qphred score below 22 base pairs and a length 30 base pairs were removed as they are regarded to be of bad quality. FastQC was used to import data from BAM, Sequence Alignment Map (SAM) or FastQ files, give summary graphs and tables for easy access to the data offline. The box plots showed the distribution of the quality scores and the bases. Additionally, the box plots showed the duplicates and the singletons. Dynamic trimming of the reads was done using SolexaQA, a tool for quality analysis and visualization (Cox *et al.*, 2010). The sequences were further trimmed by clipping using fastx-trimmer and then final quality control was done with fast QC to ensure satisfactory of the reads quality. The low quality, over-represented and adapter sequences were eliminated (Qian *et al.*, 2014).

3.5.2 *De novo* assembly and reference mapping

All the cleaned reads were a representative of the total mRNA present during the sampling of the marbled lungfish. They were combined to generate a *de novo* assembly using Trinity/v2.0.6 software (Grabherr *et al.*, 2011). Three of the sequences from each of the lakes were pooled for use in the assembly. *De novo* assembly was performed from the combined sequences from each lake where six fasta files were obtained which was later concatenated to form a complex fasta file similar to (Yang *et al.*, 2014) study. All the fasta sequences below 250 base pairs together with the duplicates were filtered. This created a catalog of contigs of maximum coverage level to determine polymorphism as marbled lungfish annotated reference genome is not yet available publicly. During the assembly, there was an allowance of 5 mismatches in each of the reads and the reads that matched to many contigs were ignored. Blastn was done on the *de novo* assembled

contigs to generate the percentage of similarity with the available nucleotide using BLAST (Basic Local Alignment Search Tool) on the National Center for Biotechnology Information (NCBI) (Wit *et al.*, 2012).

3.5.3 SNP detection

The *de novo* assembled reference was then indexed using Bowtie2 software (Langmead and Salzberg, 2012). The short reads from the 18 samples were then aligned against the reference contigs using Bowtie2 creating SAM files. The mapping statistics were analyzed using samtools flagstat parameter. Read groups were appended to each of the SAM files for each of the six lakes using Picard tools (Li et al., 2009) to relate the reads to the individuals after the bam files were pooled for SNP calling. The software package Samtools (Li *et al*., 2009) was used to convert sequence alignments from SAM format to their binary equivalent BAM format files. Duplicates were removed using the Picard toolkit.

The pooled bam file was sorted and indexed using samtools and a summary of the reads aligning to the contigs was created by Samtools pileup. Variant calling and filtering was done using bcftools to ensure that only the reliable SNPs were obtained. A variant call format file (VCF) was created (Danecek *et al*., 2011) and custom Perl scripts applied to derive the SNP genotypes. SNP analysis was done using vcftools/0.1.12b tools and SNiplay(Dereeper *et al*., 2011) where the total counts of the SNPs and the transition transversion ratios were identified respectively.

3.5.4 Single Nucleotide Polymorphism annotation

All the contigs containing the predicted SNPs were extracted and Blastx search was carried out to obtain the similar protein sequences from the Uniprot's SwissProt database and the protein databases from TrEMBL (Altschul *et al*., 1997). Blastx annotated the combine contigs with similarity to known databases of proteins, genes, and their functions. A threshold E-value was set

at 1e-20 to assess the significance of the matches. A percentage of above 70% similarity was also considered to ensure generation of the significant similarity.

3.6 Calculation of Heterozygosity (H) values

The Heterozygosity values of the total number of SNPs from the 18 samples across the six Ugandan Lakes were calculated using R Package version 0.7 as described by (Graebner *et al*., 2016). This was done because heterozygosity is averaged over several loci to provide an estimate of genetic diversity along the genome (Lewontin and Hubby, 1966).

3.7 Calculation of Guanine-cytosine (GC) content

Guanine-cytosine (GC) content describes the percentage of the guanine or cytosine bases on the particular RNA/DNA fragment. The GC pair gets bound using three hydrogen bonds making RNA structures more tolerant to increased temperatures. RNA strand with higher G-C content is more stable compared to one with low GC content (Benjamini and Speed, 2012). As a result, the primers GC-content have been utilized in the prediction of their annealing temperatures to the RNA/DNA template in PCR practical. An increased GC-content illustrates a higher melting temperature. In this research, the GC content was calculated for the specific SNP flank sequence designed at 140 base pairs.

3.8 Population structure

Using the Variant Call Format (VCF) file obtained in section 3.5.3, the population structure of the marbled lungfish was determined. The VCF file was converted to plink files using the plink software version1.9 (Purcell *et al*., 2007). The plink files included the binary ped file (.bed) that stored the genotype information. The pedigree information was found on a (.fam) file while the (.bim) file contained the allele names.

The plink ped file was used to detect for the population deviation from Hardy Weinberg Equilibrium (HWE). The ped and map file were used inAlequin3.5 (Excoffier and Lischer, 2010) to construct PCA plots for the six lake. To validate this, the VCF file was converted into a Hapmap file using Tassel software (Bradbury *et al*., 2007) to derive distance matrix values which were used to construct a neighbor joining tree. Admixture structure was constructed using a Alequin3.5 to show the population substructure (Corander and Marttinen, 2006). The Admixture plots identify genetically homogenous groups of individuals (Falush *et al*., 2016). The admixture algorithm also detects the true number of clusters (K) in individuals' samples when the trend of dispersal among the populations is not homogenous.

CHAPTER FOUR

4.0 RESULTS

4.1 Transcriptomes of the marbled lungfish

4.1.1 Generation of high quality reads

Reads with Phred quality score (Q score) above 22 and 30 base pairs length were all regarded as high quality reads. After quality filter using the NGS Quality Control Toolkit, a total number of 71191331 good quality reads were derived from the eighteen marbled lungfish transcriptomes across the six lakes as shown in Appendix I.

4.1.2 *De novo* assembly and Blastn results of the *de no* assembled contigs.

De novo assembly gave a total number of 753802 contigs. The *de novo* assembled contigs from the marbled lungfish were classified by Blastn search against NCBI (Wit *et al.*, 2012). Blastn results indicated that the unassigned contigs had the highest percentage compared to the assigned contigs as shown in Table 4.1. The percentage that was assigned to the marbled lungfish was displayed using the Krona software (Figure 4.1 and 4.2)(Ondov *et al.*, 2011).

Table 4. 1 Statistics of the percentages of assigned and unassigned contigs from the six studied lakes

Lake	total contigs	Unassigned contigs	Percentage of the assigned contigs	Percentage of unassigned contigs
Bisina	10100	8541	5%	85
Edward	8224	7248	2%	88
George	10518	8974	5%	85
Kyoga	8807	7688	3%	87
Nawampasa	9202	8049	3%	87
Wamala	7247	6205	4%	86

Blastn results demonstrated extremely low percentages of the assigned contigs for the marbled lungfish similar to the already known nucleotides in the NCBI. This reflected that most of the nucleotides found in the marbled lungfish are uncharacterized.

Figure 4. 1 Blastn results of the Bisina trinity results showing the similarity with known sequences on NCBI

The Krona results indicated that approximately 4% of the total contigs had a significant similarity to the known sequences in the NCBI representing the *P. aethiopicus.*

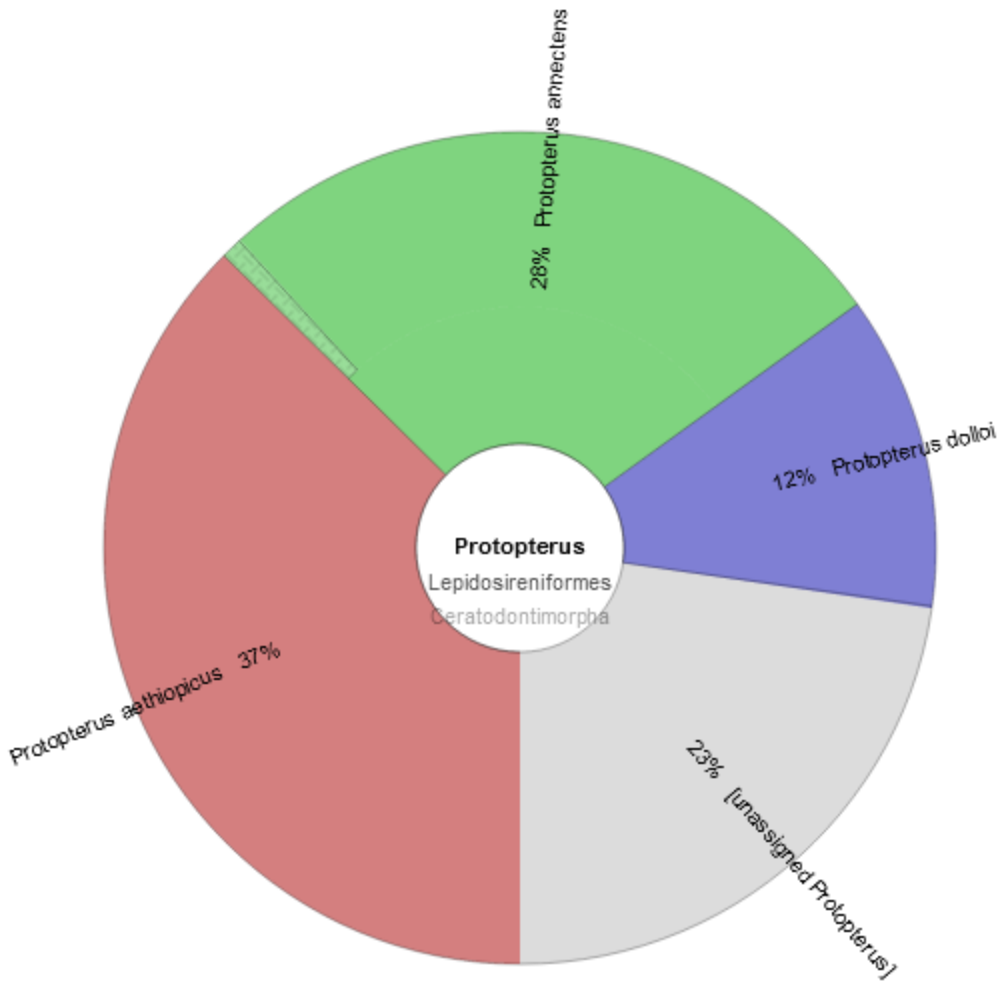


Figure 4. 2 Percentage of the query that matched to the marbled lungfish out of genus protopterus identified

The visualization of the *Protopterus* group indicated that *P. aethiopicus* (37%) had the highest representation of the similar sequences on the NCBI.

4.1.3 Read mapping and visualization

A total number of 41,075,265 reads which was an equivalent of 63.77% of reads that mapped to the assembled lungfish reference contig. This allowed the discovery of Single Nucleotide Polymorphisms. The SNPs were visualized using Integrated Genome Viewer (IGV) software

(Thorvaldsdóttir *et al*., 2013) that illustrated the reads coverage and the various positions that the SNPs occurred. Figure 4.3 shows the view of a section of the alignment of the reads to the assembled contigs at a specific point. The example shows a mismatch of two bases at read group sample 3.2 obtained from Lake Bisina which represents 0.5 % mismatch at the read coverage position. The contig had a length of 966 base pairs with maximum number of 4,668 reads. The red colored alleles show an alteration of the bases from the reference contig. Therefore, these are profound examples of the occurrence of specific variant at certain positions. The specific SNP position shows the genotypes at the particular row height with the sample identity (Thorvaldsdóttir *et al*., 2013). The grey vertical bars represent the read coverage histogram. The grey horizontal bars represent the reads and the blank grey space represents the sequence part with no reads.

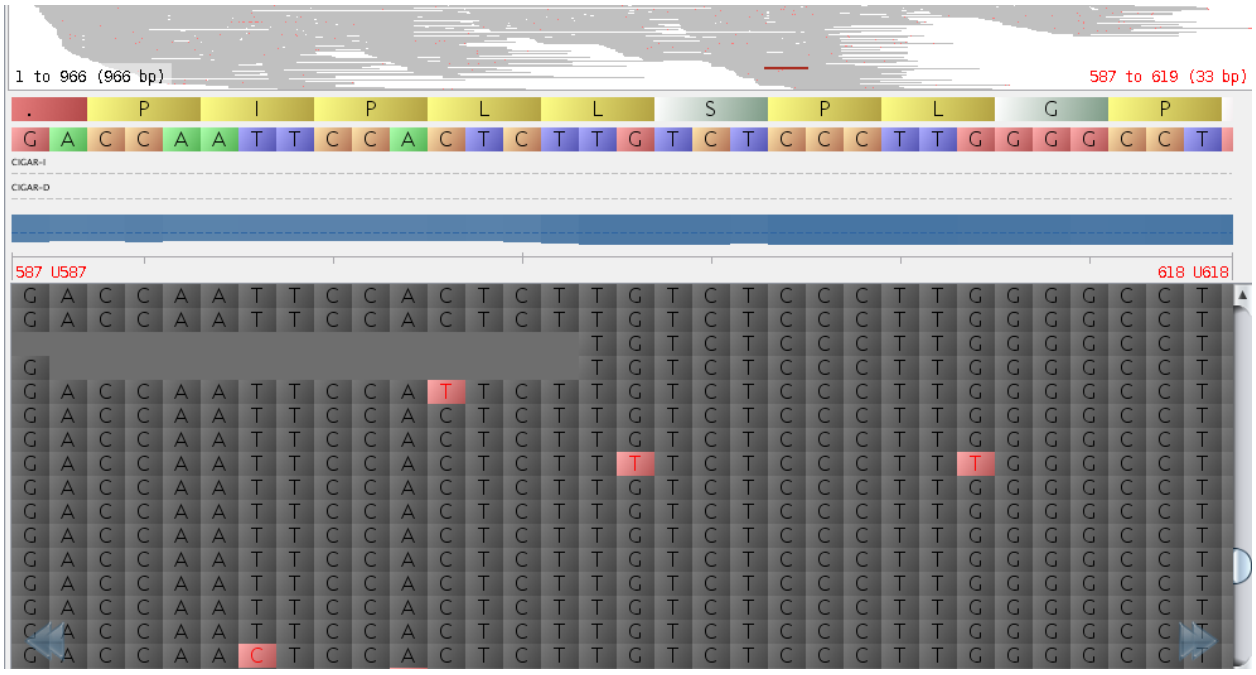


Figure 4. 3 A read coverage plot on IGV for TR65600|c0_g1_i1 contig and the SNPs detected

4.2 Single Nucleotide Polymorphisms identification

A total number of 5,961 SNPs were detected with a transition transversion ratio of 1.73. SNiplay software (Dereeper *et al*., 2011) was used to obtain the statistics of the VCF file where a total number of 3,812 SNPs were categorized as transition SNP types which composed of 564 GT, 366 CG, 615 AT and 604 AC genotypes. There were 2,149 transversions having 1938 AG genotypes and 1874 CT representatives.

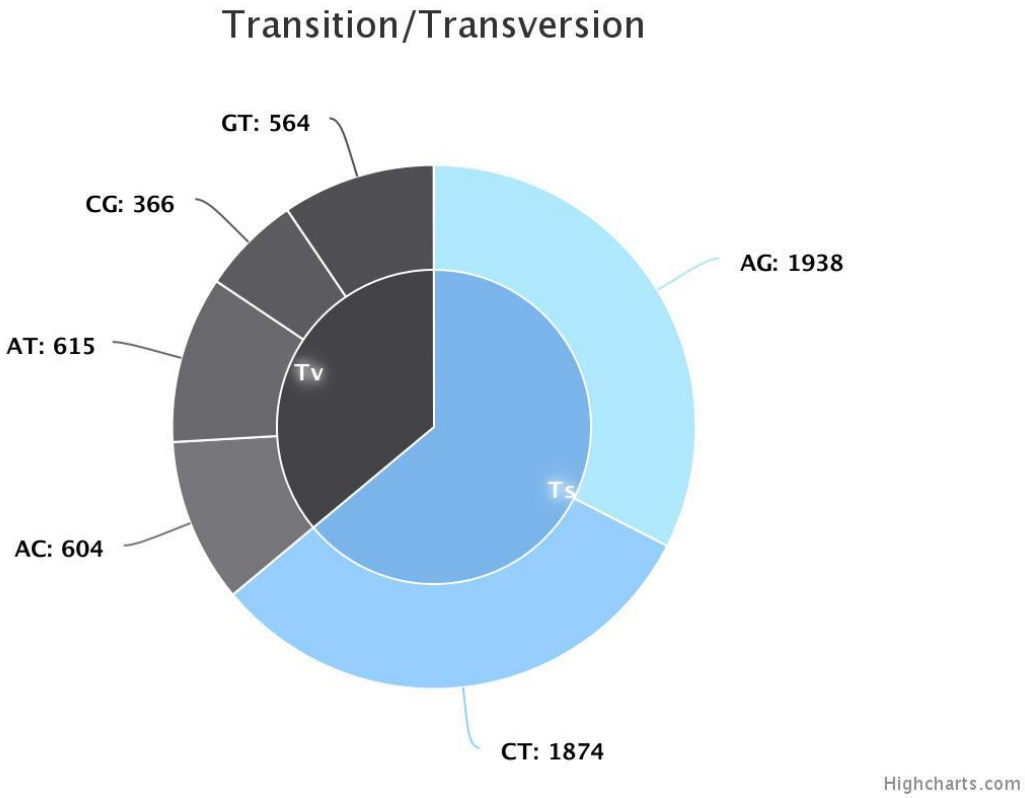


Figure 4. 4 A summary chart of the genotype composition of the transitions and transversion SNP types identified

The Minor allele frequencies (MAF) of the SNPs were determined. A total number of 943 SNPs had a MAF 0.033, 1096 SNPS had a MAF of 1352 SNP had a MAF of 0.08, 961 SNP had a MAF of 0.17, 392 SNPs had a MAF of 0.25, 456 SNPs had a MAF of 0.33 while 761 SNPs had the higher MAF of 0.452.

4.2.1 Single Nucleotide Polymorphisms contigs annotation

Functional annotation to detect the SNP contigs similarity to the identified proteins in the Uniprot and Tremble databases (Altschul *et al*., 1997) illustrated that most of the significant hits were uncharacterized. Nevertheless, some had higher percentages of hit to the known proteins in the databases. Only the SNP contigs that had a significant hit were considered for further analysis to check on the associated gene function. The resulting significant sequences indicated the score percentages of similarity and the gaps that exists between the query sequence and the known sequences. Gene function of some of the significant hits was evaluated using NCBI Blast2GO as shown in Table 4.2.

Table 4. 2 A summary of the gene identified from the annotation of the marbled lungfish contigs against the Uniprot and Tremble fasta databases

Gene ID	Description	Function
Poldip3	polymerase (DNA-directed), delta interacting protein 3	Specifically targets S6 kinase 1 and regulates cell growth.
Su(P)	Putative uncharacterized protein Su(P)	Cell redox homeostasis, Oxidation-reduction process, electron carrier activity, protein disulfide oxidoreductase
Dyak\ref(2)P	GE13269 gene product from transcript GE13269-RA	autophagy related gene
TCM_008841	Rubber elongation factor protein (REF), putative isoform 1	Molecular function- Translation elongation factor activity
SmD2	Probable small nuclear ribonucleo-protein Sm D2	Molecular function- Poly(A) RNA binding Biological function- Mitotic nuclear division
hypothetical protein AMK59_6360	hypothetical protein AMK59_6360	Molecular function- protein binding
Cbn-snr-4	CBN-SNR-4 protein	RNA Splicing
NECAME_08035	LSM domain protein	Protein domain specific binding
gb:eh507706	Uncharacterized protein	Calcium and Ion channel activity,
LOC106603799	Small nuclear ribonucleoprotein Sm D2-like	Small molecular binding
ref-2	REgulator of Fusion	Molecular function-metal ion binding and nucleic binding Biological function-DNA-templated transcription, initiation
Dgri\GH14279	GH14279	RNA splicing

4.3 A catalog of Single Nucleotide Polymorphisms discovered

The level of heterozygosity of markers is used to define the quality of a polymorphism and is a vital determinant of the suitable SNPs for use in breeding procedures (Reed and Frankham, 2003; Lewontin and Hubby, 1966). The locus heterozygosity defines the likelihood for an individual being heterozygous for a specific locus in a given population (Gregorius, 1978). This study relied on the heterozygosity values and the percentage of the GC content to show the measures of the quality of the variation of the SNP markers. A total number of 1979 SNPs which is an equivalent of 29.6% of the total 5,961 discovered SNPs had H values of between 0.2 and 0.5 values. In addition, the 29.6% SNPs had satisfactory GC content of 40-60%. SNP flanking sequences of length 140 base pairs were extracted from the RNA sequences where the specific polymorphism was detected. Selection was done based on their GC content and those SNPs whose flanking sequences attained a GC content of 40-60% were chosen as potential SNPs for genetic diversity studies and breeding. From the 1979 SNPs selected based on GC content and H values, a set of 198 SNPs with the maximum H value of 0.5 was selected to develop a SNPs panel for genetic diversity studies and breeding purposes. The SNP panel is shown in Appendix II.

4.4.Population structure

4.4.1 Pairwise population relatedness

All the six lakes population passed the test where after filtering the population for HWE using Plink software (Purcell, 2007), the HWE statistic kept 6 out of 6 populations for the biallelic loci. The SNP data generated was used to estimate genetic relatedness among the marbled lungfish across the six study Lakes in Uganda. Plink software showed the identical by state (IBS) values for estimating pairwise relatedness among the individuals as shown in Table 4.3. Individuals

from Lake Bisina portrayed a very close genetic relatedness with an IBS value of **0.9194015**. On the other hand, the least similar individuals were a comparison of individuals from Lake Wamala and Kyoga with an IBS value of **- 0.0188359**. Individuals from Lake George and Bisina were highly similar amongst themselves but very distinct compared to individuals from the other four lakes. The average relatedness across all the species was 0.10747097.

Table 4. 3 The values of the Identity by State (IBS) across the six lakes

Lake	Bisina	Edward	George	Kyoga	Nawampasa	Wamala
Bisina	**0.9194015**	0.0590479	0.026233	0.0212926	0.05513227	0.06460211
Edward	0.0590479	0.4119598	0.008586	0.0460569	0.01953575	0.0440788
George	0.0262393	0.0085836	0.515903	0.0138788	0.02648222	0.0280764
Kyoga	0.0212926	0.0460569	0.013878	0.3699302	0.01776149	**-0.0188359**
Nawampasa	0.0551327	0.0195355	0.026482	0.0177619	0.43778764	0.0195957
Wamala	0.0646021	0.0440788	0.028074	- 0.0188359	0.01959527	0.3509239

4.4.2 Admixture structure for population analysis

Model-based clustering was employed to further predict the minimum number of subpopulations (k) that would explain genetic variation across the six lakes. The admixture structure showed that there are four subpopulations in the marbled lungfish population across the six lakes as shown in Figure 4.5. This was done using 4,565 SNPs filtered at genotyping rate of 0.5 percent to allow accuracy of the results using Plink toolkit. At K6, Lakes Bisina and George seemed to have pure population while the other lakes have admixture populations. This infers that there is a lot of genetic variation among individuals in these lakes compared to Lakes Bisina and George. The figure demonstrates six populations. It appeared that Bisina population is more pure compared to the others. The population from the other lakes is highly mixed. At K2, Edward and George are grouped together but at K3, K4, K5 and K6 they are separated. It shows that Edward and Bisina

are two different populations. At K2, Wamala and Nawampasa are grouped together but with further groupings in K3, K4, K5 and K6, the populations are grouped separately.

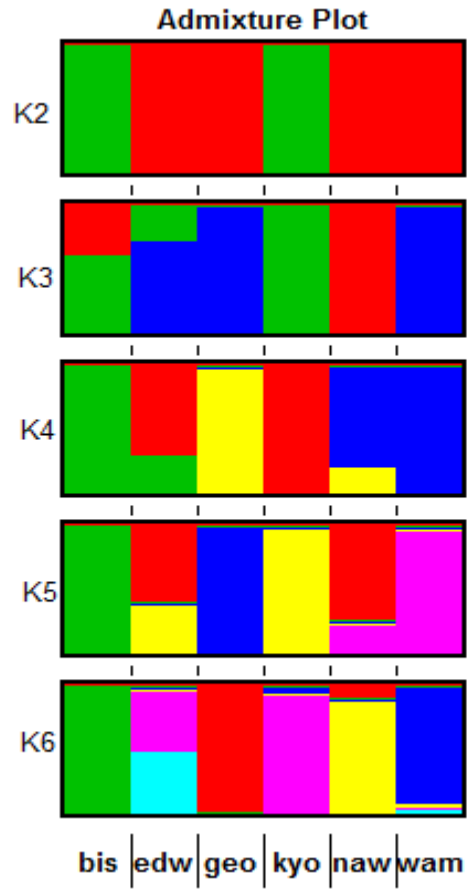
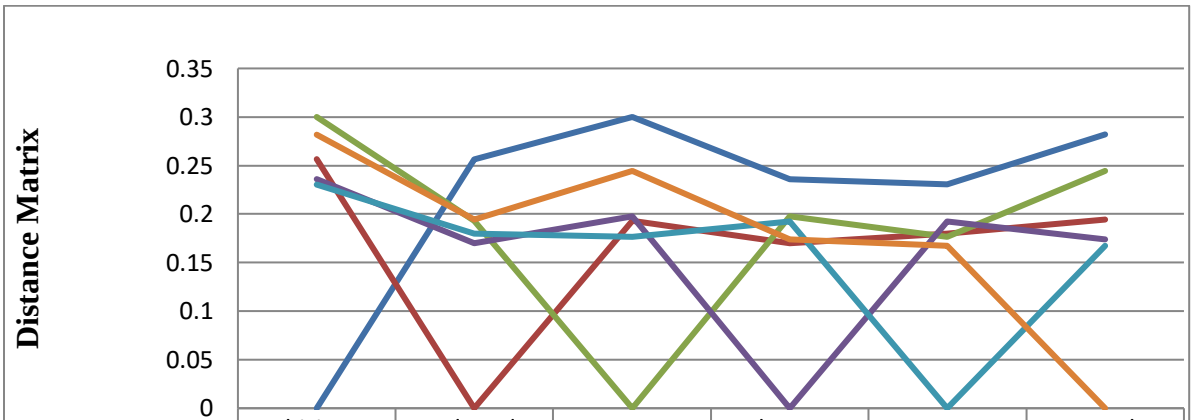


Figure 4. 5plots of genetic admixture proportions (Q) according to additional values of K (bis – Lake Bisina, edw – Lake Edward, geo – Lake George, kyo – Lake Kyoga, naw – Lake Nawampasa ; wam – Lake Wamala).

4.4.3 Distance matrix for the marbled lungfish from the six lakes

Hapmap files were generated from the Variant Call Format file using Tassel software (Bradbury *et al.*, 2007). The files were then used to estimate the population distance matrix using the genetic distance from sequences as shown in Figure 4.6. The results showed a high genetic distance between populations in Lake Bisina and George **(0.3)**. Lake Wamala and Nawampasa populations showed the smallest distance between them **(0.16746988)**. Lake Kyoga and Edward showed a relatively close genetic distance **(0.170254403)**. The average distance matrix was **0.14797638**.



	bisina	edward	george	kyoga	nawampasa	wamala
bisina	0	0.256548536	0.3	0.236	0.230337079	0.281783681
edward	0.256548536	0	0.192965779	0.170254403	0.180124224	0.194587629
george	0.3	0.192965779	0	0.19785575	0.176252319	0.244332494
kyoga	0.236	0.170254403	0.19785575	0	0.192307692	0.173684211
nawampasa	0.230337079	0.180124224	0.176252319	0.192307692	0	0.16746988
wamala	0.281783681	0.194587629	0.244332494	0.173684211	0.16746988	0

Figure 4. 6 The matrix distances between the selected six lakes

The distance matrix data was used to construct a neighbor-joining tree that clustered the lakes into four groups as shown in Figure 4.7. The phylogenetic tree shows operational taxonomic units that comprises of one node joining them (Saitou and Nei, 1987). The six populations were clustered into four groups, (a) Bisina, (b) Edward and Kyoga, (c) Nawampasa and Wamala, and (d) George. Nawampasa and Wamala and Kyoga and Bisina were the closest groups.

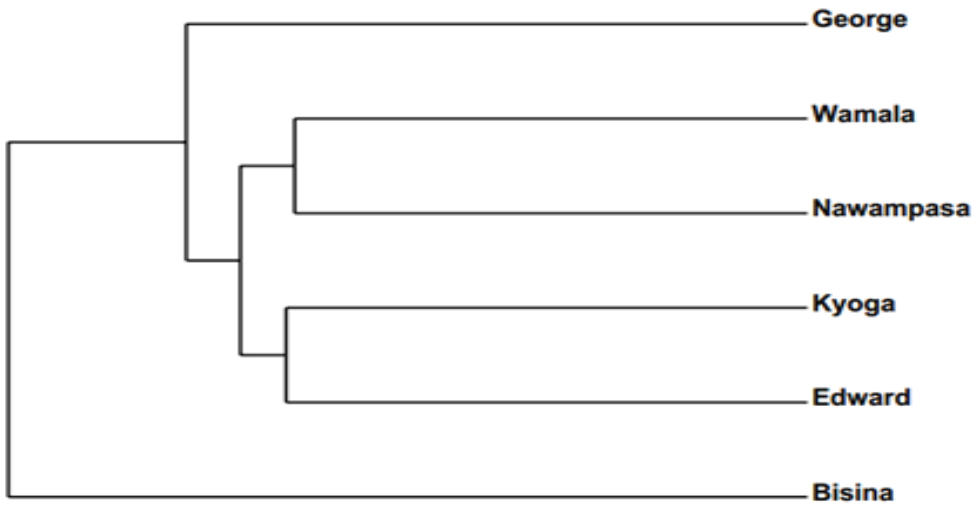


Figure 4. 7 A neighbor joining tree for the six lake regions

4.4.4 Principal Component Analysis (PCA)

Further clustering was done using Principal Component Analysis which also predicted four clusters as in neighbor joining tree (a) Bisina, (b) Edward and Kyoga, (c) Nawampasa and Wamala, and (d) George (Figure 4.8). These indicate that species from Lake Bisina and George have high genetic divergence from the other lakes. Species from Lake George and Bisina are not closely related genetically with the other species from the other lakes. Lake Wamala and Nawampasa are grouped together and the same to Lake Kyoga and Edward.

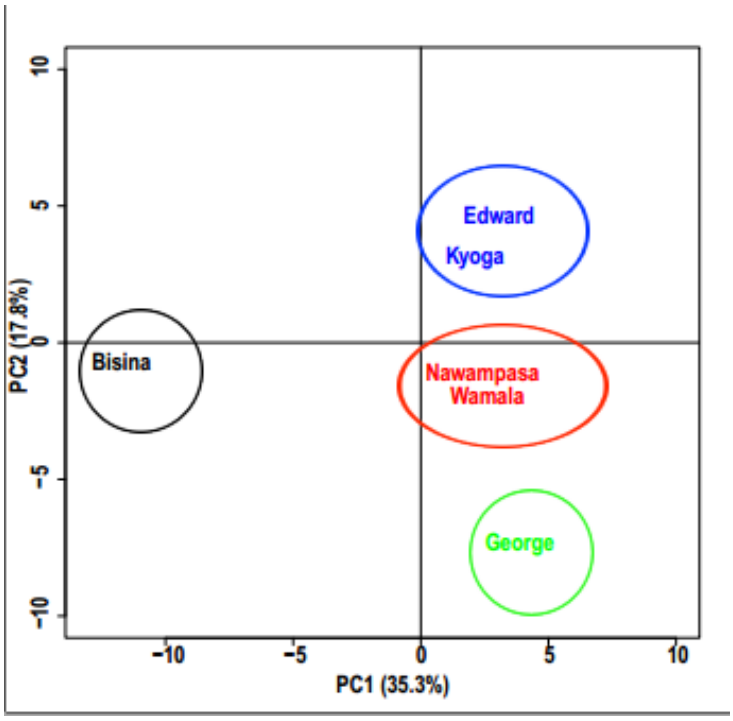


Figure 4. 8A PCA plot of the six study lakes in Uganda

CHAPTER FIVE

5.0 DISCUSSION

5.1 Transcriptomes of marbled lungfish for SNP discovery

Next generation sequencing have enabled thorough and efficient study of the transcriptomes of various species for the detection of SNPs (Jeukens *et al*., 2010 ; Peterson *et al*., 2014). Diopere *et al*. (2013) demonstrate that NGS approaches can be applied in evolutionary and fisheries studies to understand genomic and management aspects. The identified SNPs can be used to discover probable disease phenotypes and other economically valuable species traits of importance for sustainable management practices through genetic association studies as demonstrated by Ma *et al*. (2008). This study used illumina sequenced data for the detection of SNP markers that would be precise for improved conservation strategies of the marbled lungfish. The interpretation of the genetic variation of the cultured marbled lungfish could allow breeders to link particular traits with the identified SNPs markers. This would support determination of novel alleles that could be used to promote productivity. Breeding programs based on highly informative SNP markers would be precise to adjust to suitable aquaculture system for marbled lungfish (Altshuler *et al*., 2000). It would form basis for their conservation and support other genetic diversity studies of the species thereby promoting national growth through increased productivity (Aslam *et al*., 2012).

The study of marbled lungfish resulted in the discovery of 5,961 SNPs to represent the genetic variation of the population. The rate of transition: transversion was 1.73. Only 1979 out of the 5,961 SNPs attained heterozygosity values between 0.2 - 0.5 that showed more SNPs were moderately low polymorphic while a lesser number had maximum heterozygosity values. The study relied on the SNPs with the maximum heterozygosity value (0.5) which would provide

appropriate information for the marbled lungfish. Kooloos et al. (2009) has demonstrated that a primary selection criterion for SNPs uses a range of (0.400 - 0.480). The selection criteria used in the marbled lungfish study to select suitable SNPs is within the range.

In *L. vannamei* study (Yu *et al*., 2014), discovery of 96,040 SNPs with a transition: transversion of 2.0 were indicated as reliable for use during the study of *L. vannamei* in genome wide association research and high density linkage map construction. The transition: transversion ratio is within the range of the current study of marbled lungfish (1.73). The MAF of the highest number of SNPs from *L. vannamei* was within a range of 0.45 - 0.5.

SNP minor allele frequencies have also been utilized to determine the genetic variation in pigs (Ramos *et al*., 2009). The study used 158 pigs to detect SNPs using *de novo* assembled reference genome. The average MAF for the selected SNPs in the pig study was 0.274. Validation of the SNP60 Bead chip in pig was done for use as a reliable tool in further pig studies. The study indicated the use of NGS as an appropriate way to generate reliable SNPs. In the marbled lungfish study, selected SNPs (198) had a MAF of 0.25 - 0.452 that is within a similar range with the research on pig. In marbled lungfish, about 2,570 had a MAF of 0.17 - 0.452 and 3391 SNPs had 0.003 -0.08 that indicated a relatively low genetic variation within the populations.

Transcriptomes have also been linked to the traits of interest like growth advantages (Diopere *et al*., 2013). The study used a candidate gene approach in sole (*Solea solea L.)* to detect novel SNPs in or around the genes related to significant life history characteristics of the species (Diopere *et al*., 2013). Screening was done to a total number of 76 candidate genes associated with growth and maturation. The study identified a total number of 22 informative SNPs for use as a tool in evolutionary signature of over-exploitation of sole fish and for purposes of marker-

assisted selection. Diopere *et al*., (2013) show that transcriptomes are important for SNPs discovery.

Research on Atlantic salmon is an evident of successful discovery of true bi-allelic SNPs using the Next generation sequencing platforms (Houston *et al*., 2014). The study used 283 Atlantic salmon samples giving approximately 400K putative SNPs that would be used for further studies in aquaculture breeding practices of salmoid species through genomic selection. Comparing our research on marbled lungfish to this Atlantic salmon study, the 198 SNPs panel detected could be validated and used in genetic mapping and linkage research (Suh and Vijg, 2005). It would also form a great basis for the conservation of the wild stocks.

Other related studies include the development of high-density SNP chips in cattle. The study demonstrates the significance of NGS technology in SNP detection and genomic selection (Matukumalli *et al*., 2009). The study identified 54,001 SNP loci with an average MAF of 0.24 - 0.27. The assay was reported to be suitable for the mapping disease genes and QTL in cattle.

SNP contigs annotation using blastx showed some of the similar proteins and their relative uses. The proteins functions identified ranged from biological, molecular and chemical operations which could have developmental purposes in marbled lungfish growth. The proteins could link certain SNPs to particular traits for use when choosing fish for breeding. The main interest was to identify some of the genes that could be linked to some of the biochemical functions in the marbled lungfish. Dyak\ref (2) P gene was found to have an association with autophagy characteristic of the marbled lungfish. This could be the contributing factor for the marbled lungfish aestivation. Poldip3 gene was identified that specifically targets S6 kinase 1 and regulates cell growth. Cbn-snr-4 and Dgri\GH14279 were linked to RNA splicing.

5.1.1 Single Nucleotide Polymorphisms selection based on the Heterozygosity values

Next generation sequencing is a sufficient resource to identify putative SNP markers from the species transcriptomes. This study identified 5,961 SNPs from a pool of 18 samples where only 5,112 had H values. The SNPs with the maximum heterozygosity showed higher polymorphic levels in the transcriptomes. The SNPs could be possibly significant for genomic selection of the marbled lungfish. A related study on catfish species filtered SNPs within a range of 0.28 – 0.31 for use in the genetic studies (Liu *et al.*, 2011). This is within the range of the MAF for the marbled lungfish research where 2,570 SNPs out of 5,196 SNPs had a MAF of 0.17 - 0.452. The study on marbled lungfish selected SNPs with a flanking sequence of GC percentage content between 40 - 60%. Liu *et al*. (2014) used flanking sequences of GC content within a range of 30%-70% suitable for designing a SNP assay. The selected SNPs could guide in marbled lungfish conservation and management practices.

5.1.2 Genetic diversity interpretation using the heterozygosity values

Genetic variation of the marbled lungfish was determined by the heterozygosity characteristic of the SNPs discovered in this study. The detected SNPs will be important in the genetic diversity studies and could be a clear guide for genomic selection during aquaculture practices. The heterozygosity values demonstrate the degree of polymorphism of the SNPs (Gregorius, 1978). The low genetic diversity among the populations within Lakes Bisina and George indicates high levels of inbreeding for the marbled lungfish in the lakes (Gregorius, 1978). Anthropogenic factors might also have resulted in decreased population causing inbreeding, especially Lake George. Low genetic diversity within Lake Bisina populations can be explained by the fact that it is under Ramsar Convention that helps in conserving it as a bird area that has diverse flora and fauna (Ramsar Convention on Wetlands, 2013).

5.2 Population structure

5.2.1 Hardy Weinberg Equilibrium

The HWE test for the pooled populations showed no significant deviation from the principle. It proved that the genetic polymorphism in the population would remain constant from one generation to the other when there will be no disruptive elements (Wigginton et al., 2005). The principle suggests that the genotypes and the allele frequencies would remain invariable since they are in equilibrium. Populations would go against the HWE if the evolutionary factors would lead to the change of the allele frequencies. Disruption elements like mutations, genetic drift, recombination during sexual reproduction, natural selection and gene flow at this study sites can change equilibrium (Wigginton et al., 2005). This means that mating is not random, genetic drift further reduces the population size, mutation takes place and all the members of the population do not survive and have unequal reproduction rates. Such elements would change the relative allele frequencies in the populations from one generation to the other.

5.2.2 Admixture structure

The admixture structures have been widely utilized to explore population genetic structure (Falush *et al*., 2016). The accuracy of genetic mixture occurs when highly informative markers are determined with desirable statistical approaches. The current study admixture structure gave an illustration of a close genetic relation for species within Lake Bisina and George. The other lakes showed relatively high genetic variation of individuals within the lakes. Close genetic diversity infers that the populations could have higher inbreeding levels while high genetic diversity means reduced inbreeding.

5.2.3 Principal Component Analysis and neighbor joining trees

The VCF SNP data file in the form of plink files were used to determine the population structure across the six lakes. The lakes were clustered into four groups using the Principal Component Analysis structure and the neighbor joining trees.

CHAPTER SIX

6.0 CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

A total number of 198 SNPs out of 5,961 putative SNPs were detected to have a higher level of heterozygosity and were considered informative which could provide general information for improved genetics studies and breeding programs of marbled lungfish. The SNP data provided appropriate basis for population diversity analysis and comparison of the variation among the lakes using SNPs (Aslam *et al*., 2012). The SNP chip will be publicly availed to guide in various genetic and linkage studies of the marbled lungfishes.

The six populations used in this study were put into four clusters (a) Bisina, (b) Edward and Kyoga, (c) Nawampasa and Wamala, and (d) George (Figure 4.8) using Admixture plots, PCA, IBS, neighbor adjoining trees and their distance matrix. The study illustrated a close genetic relatedness for species within Lake Bisina and George. The low genetic diversity among these two populations indicates high levels of inbreeding for the marbled lungfish in the lakes. Lake Bisina's low genetic diversity can be explained by the fact that it is under Ramsar Convention that helps in conserving it as a bird area with diverse flora and fauna (Ramsar Convention on Wetlands, 2013). Lake George low genetic diversity could infer that anthropogenic factors might have interrupted the wetland causing a decrease in their population that could results in inbreeding. Clusters (b) Edward and Kyoga and (c) Nawampasa and Wamala, (Figure 4.8) showed relatively high genetic variation of individuals within the lakes suggesting that anthropogenic factors might have led to gene flow among the fish in the lakes.

6.2 Recommendations

This study recommends that the identified SNPs should be validated to form a significant basis for the diversity studies of *P. aethiopicus.* The SNP would further be used for the characterization of the marbled lungfish that is critical for improved productivity (Groenen *et al*., 2011). This would allow the actual linkage of the specific panel SNPs to the traits of interests during aquaculture (Houston *et al*., 2014). Additionally, this study recommends similar studies using more samples than the currently used to help to track marbled lungfish gene flow across the lakes. A deeper sequencing at genome level ought to be done to provide valuable information on marbled lungfish genetic diversity studies. The estimated low genetic diversity in Lake Bisina and George means that there is the need to conserve the wild stocks within these lakes to maintain genetic diversity (Charlesworth, 2003). This could reduce the levels of inbreeding. The estimated genetic diversity among populations within clusters (b) Edward and Kyoga and (c) Nawampasa and Wamala, (Figure 4.8) must also be maintained.

REFERENCES

Alexander, D.H. and Lange, K., 2011. Enhancements to the Admixture algorithm for individual ancestry estimation. *BMC bioinformatics*, *12*(1), p.246.

Allen, G.R., Midgley, S.H. and Allen, M., 2002. *Field guide to the freshwater fishes of Australia*. Western Australian Museum.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), pp.3389-3402.

Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L. and Lander, E.S., 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, *407*(6803), pp.513-516.

Anderson, S., De Bruijn, M.H.L., Coulson, A.R., Eperon, I.C., Sanger, F. and Young, I.G., 1982. Complete sequence of bovine mitochondrial DNA conserved features of the mammalian mitochondrial genome. *Journal of molecular biology*, *156*(4), pp.683-717.

Aquadro, C.F., Kaplan, N. and Risko, K.J., 1984. An analysis of the dynamics of mammalian mitochondrial DNA sequence evolution. *Molecular biology and evolution*, *1*(5), pp.423-434.

Aslam, M.L., Bastiaansen, J.W., Elferink, M.G., Megens, H.J., Crooijmans, R.P., Blomberg, L.A., Fleischer, R.C., Van Tassell, C.P., Sonstegard, T.S., Schroeder, S.G. and Groenen, M.A., 2012. Whole genome SNP discovery and analysis of genetic diversity in Turkey (Meleagris gallopavo). *BMC genomics*, *13*(1), p.391.

Avise, J.C., 1989. A role for molecular genetics in the recognition and conservation of endangered species. *Trends in Ecology & Evolution*, *4*(9), pp.279-281.

Avise, J.C., 2012. *Molecular markers, natural history and evolution*. Springer Science & Business Media.

Babiker, M.M. 1979. Respiratory behaviour, oxygen consumption and relative dependence on aerial respiration in the African lungfish (Protopterus annectens, Owen) and air-breathing teleosts (Clarius lazera, C.). Hydrobiologia 65: 177–187.

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

Balirwa, J.S., Chapman, C.A., Chapman, L.J., Cowx, I.G., Geheb, K., Kaufman, L., Lowe-McConnell, R.H., Seehausen, O., Wanink, J.H., Welcomme, R.L. and Witte, F., 2003. Biodiversity and fishery sustainability in the Lake Victoria basin: an unexpected marriage?.*BioScience*, *53*(8), pp.703-716.

Barange, M., Merino, G., Blanchard, J.L., Scholtens, J., Harle, J., Allison, E.H., Allen, J.I., Holt, J. and Jennings, S., 2014. Impacts of climate change on marine ecosystem production in societies dependent on fisheries. *Nature Climate Change*, *4*(3), pp.211-216.

Benjamini, Y. and Speed, T.P., 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research*, *40*(10), pp.e72-e72.

Biscotti, M.A., Gerdol, M., Canapa, A., Forconi, M., Olmo, E., Pallavicini, A., Barucca, M. and Schartl, M., 2016. The lungfish transcriptome: a glimpse into molecular evolution events at the transition from water to land. *Scientific reports*, *6*, p.21571.

Bisina, L., 2009. Ecological Baseline Surveys Of : Lake Bisina - Opeta Wetlands System Lake Mburo - Nakivali Wetlands System.

Boettcher, P.J., Tixier-Boichard, M., Toro, M.A., Simianer, H., Eding, H., Gandini, G., Joost, S., Garcia, D., Colli, L.I.C.I.A. and Ajmone-Marsan, P.A.O.L.O., 2010. Objectives, criteria

and methods for using molecular genetic data in priority setting for conservation of animal genetic resources. *Animal Genetics*, *41*(s1), pp.64-77.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. and Buckler, E.S., 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), pp.2633-2635.

Brien, P., 1959. Ethologie du Protopterus dolloi (Boulenger) et de seslarves. Signification des sacs pulmonaires des Dipneustes. *Ann. Soc. R. Zool. Belg*, *89*, pp.9-48.

Bruton, M.N., 1998. In Paxton, JR & Eschmeyer, WN. ed. Encyclopedia of Fishes.

Caballero, A. and Toro, M.A., 2002. Analysis of genetic diversity for the management of conserved subdivided populations. *Conservation Genetics*, *3*(3), pp.289-299.

Charlesworth, D., 2003. Effects of inbreeding on the genetic diversity of populations. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *358*(1434), pp.1051-1070.

Collins, F.S., Guyer, M.S. and Chakravarti, A., 1997. Variations on a theme: cataloging human DNA sequence variation. *Science*, *278*(5343), pp.1580-1581.

Corander, J. and Marttinen, P., 2006. Bayesian identification of admixture events using multilocus molecular markers. *Molecular ecology*, *15*(10), pp.2833-2843.

Costa, V., Angelini, C., De Feis, I. and Ciccodicola, A., 2010. Uncovering the complexity of transcriptomes with RNA-Seq. *BioMed Research International*, *2010*.

Cox, M.P., Peterson, D.A. and Biggs, P.J., 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC bioinformatics*, *11*(1), p.485.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. and McVean, G., 2011. The variant call format and VCFtools. *Bioinformatics*, *27*(15), pp.2156-2158.

De Vos, L., Snoeks, J. and van den Audenaerde, D.T., 2001. An annotated checklist of the fishes of Rwanda (East Central Africa), with historical data on introductions of commercially important species. *Journal of East African Natural History*, *90*(1), pp.41-68.

DeFaveri, J., Shikano, T., Shimada, Y., Goto, A. and Merilä, J., 2011. Global analysis of genes involved in freshwater adaptation in three spine sticklebacks (*Gasterosteus aculeatus*). *Evolution*, *65*(6), pp.1800-1807.

Department of Fisheries Resources. 2004. Fisheries Sector Strategic Plan: Ministry of Agriculture, Animal Industry and Fisheries, Entebbe, Uganda.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M. and McKenna, A., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, *43*(5), pp.491-498.

Dereeper, A., Nicolas, S., Le Cunff, L., Bacilieri, R., Doligez, A., Peros, J.P., Ruiz, M. and This, P., 2011. SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC bioinformatics*, *12*(1), p.134.

Diopere, E., Hellemans, B., Volckaert, F.A. and Maes, G.E., 2013. Identification and validation of single nucleotide polymorphisms in growth-and maturation-related candidate genes in sole (Soleasolea L.). *Marine genomics*, *9*, pp.33-38.

Djari, A., Esquerré, D., Weiss, B., Martins, F., Meersseman, C., Boussaha, M., Klopp, C. and Rocha, D., 2013. Gene-based single nucleotide polymorphism discovery in bovine muscle using next-generation transcriptomic sequencing. *BMC genomics*, *14*(1), p.307.

Duran, C., Appleby, N., Edwards, D. and Batley, J., 2009. Molecular genetic markers: discovery, applications, data storage and visualisation. *Current Bioinformatics*, *4*(1), pp.16-27.

Ekblom, R. and Galindo, J., 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, *107*(1), pp.1-15.

Ellegren, H. and Galtier, N., 2016. Determinants of genetic diversity. *Nature Reviews Genetics*, *17*(7), pp.422-433.

Ellegren, H., 2004. Microsatellites: simple sequences with complex evolution. *Nature reviews genetics*, *5*(6), pp.435-445.

Excoffier, L. and Lischer, H.E., 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, *10*(3), pp.564-567.

Fallis, A.., 2013. *Bioinformatics and functional genomics*. Journal of Chemical Information and Modeling.

Falush, D., van Dorp, L. and Lawson, D., 2016. A tutorial on how (not) to over-interpret Structure/Admixture bar plots. *BioRxiv*, p.066431.

FAO *(May 1990).* "The Mismanagement of Lake Wamala's Fish Resources". *Rome, Italy:* Food and Agricultural Organization *(FAO)*. Retrieved 22 July 2014.

FAO. (2012). *The State of World Fisheries and Aquaculture.* Italy: FAO.

FAO. (2016). *The State of World Fisheries and Aquaculture.* Italy: FAO.

Feltus, F.A., Wan, J., Schulze, S.R., Estill, J.C., Jiang, N. and Paterson, A.H., 2004. An SNP resource for rice genetics and breeding based on subspecies *indica* and *japonica* genome alignments. *Genome research*, *14*(9), pp.1812-1819.

FishBase team RMCA, Geelhand, D. & Hughes, A. 2016. *Protopterus aethiopicus ssp. aethiopicus*. The IUCN Red List of Threatened Species 2016: e.T183073A49783255. http://dx.doi.org/10.2305/IUCN.UK.2016-3.RLTS.T183073A49783255.en. Downloaded on 29 June 2017.

Fishbase.org (Retrieved April 21, 2017.)

Frankham, R., Briscoe, D.A. and Ballou, J.D., 2002. *Introduction to conservation genetics*. Cambridge university press.

Garner, S., Birt, T.P., Mlewa, C.M., Green, J.M., Seifert, A. and Friesen, V.L., 2006. Genetic variation in the marbled lungfish *Protopterus aethiopicus* in Lake Victoria and introduction to Lake Baringo, Kenya. *Journal of fish biology*, *69*(sb), pp.189-199.

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C. and Estoup, A., 2013. Estimation of population allele frequencies from next generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology*, *22*(14), pp.3766-3779.

Gayral, P., Weinert, L., Chiari, Y., Tsagkogeorga, G., Ballenghien, M. and Galtier, N., 2011. Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. *Molecular Ecology Resources*, *11*(4), pp.650-661.

Gjedrem, T., 2012. Genetic improvement for the development of efficient global aquaculture: a personal opinion review. *Aquaculture*, *344*, pp.12-22.

Gordon, D.V. and Maurice, S., 2015. Vertical and horizontal integration in the Uganda fish supply chain: Measuring for feedback effects to fishermen. *Aquaculture Economics & Management*, *19*(1), pp.29-50.

Gosse, J.P., 1984. Protopteridae. *Check-list of the Freshwater Fishes of Africa*, *1*, pp.8-17.

Govindaraj, M., Vetriventhan, M. and Srinivasan, M., 2015. Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genetics research international*, *2015*.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q. and Chen, Z., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, *29*(7), pp.644-652.

Graebner, R.C., Hayes, P.M., Hagerty, C.H. and Cuesta-Marcos, A., 2016. A comparison of polymorphism information content and mean of transformed kinships as criteria for selecting informative subsets of barley (Hordeum vulgare L. sl) from the USDA Barley Core Collection. *Genetic resources and crop evolution*, *63*(3), pp.477-482.

Greenwood, P.H. 1958. Reproduction in the East African lung-fish *Protopterus aethiopicus* (Heckel). Proceedings of the Zoological Society of London 130: 547-567.

Greenwood, P.H., 1966. *The fishes of Uganda*. Kampala: Uganda Society.

Gregorius, H.-R. 1978. The concept of genetic diversity and its formal relationship to heterozygosity and genetic distance. Math. Bioscience 41: 253-271.

Groenen, M.A., Megens, H.J., Zare, Y., Warren, W.C., Hillier, L.W., Crooijmans, R.P., Vereijken, A., Okimoto, R., Muir, W.M. and Cheng, H.H., 2011. The development and characterization of a 60K SNP chip for chicken. *BMC genomics*, *12*(1), p.274.

Hanotte, O. and Jianlin, H., 2006. Genetic characterization of livestock populations and its use in conservation decision-making. *The Role of Biotechnology in Exploring and Protecting Agricultural Genetic Resources. Food and Agriculture Organization of the United Nations, Rome*, pp.89-96.

Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S. and Frazer, K.A., 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome biology*, *10*(3), p.R32.

Helyar, S.J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M.I., Ogden, R., Limborg, M.T., Cariani, A., Maes, G.E., Diopere, E., Carvalho, G.R. and Nielsen, E.E., 2011. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, *11*(s1), pp.123-136.

Hens, L., 1996. The Rio conference and thereafter. *Sustainable development*, pp.81-109.

Hoffmann, I., 2010. Climate change and the characterization, breeding and conservation of animal genetic resources. *Animal genetics*, *41*(s1), pp.32-46.

Holland, S.M., 2016. Principal components analysis (PCA). *Department of Geology, University of Georgia, Athens, GA*, pp.30602-2501.

Houston, R.D., Taggart, J.B., Cézard, T., Bekaert, M., Lowe, N.R., Downing, A., Talbot, R., Bishop, S.C., Archibald, A.L., Bron, J.E. and Penman, D.J., 2014. Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). *BMC genomics*, *15*(1), p.90.

Hughes, J.M., Schmidt, D.J., Huey, J.A., Real, K.M., Espinoza, T., McDougall, A., Kind, P.K., Brooks, S. and Roberts, D.T., 2015. Extremely low microsatellite diversity but distinct

population structure in a long-lived threatened species, the Australian lungfish Neoceratodus forsteri (Dipnoi). *PloS one*, *10*(4), p.e0121858.

Illumina, 2011. Quality Scores for Next-Generation Sequencing. *Http://Res.Illumina.Com/Documents/Products/Technotes/Technote_Q-Scores.Pdf*, pp.1–2.

Jeukens, J., Renaut, S., S., ST-CYR, J.É.R.Ô.M.E., Nolte, A.W. and Bernatchez, L., 2010. The transcriptomics of sympatric dwarf and normal lake whitefish (Coregonus clupeaformis spp., Salmonidae) divergence as revealed by next-generation sequencing. *Molecular ecology*, *19*(24), pp.5389-5403.

Jiang, G.L., 2013. Molecular markers and marker-assisted breeding in plants. In *Plant breeding from laboratories to fields*. Intech.

Kilian, B. and Graner, A., 2012. NGS technologies for analyzing germplasm diversity in genebanks. *Briefings in functional genomics*, *11*(1), pp.38-50.

Kjaer, A.M., Muhumuza, F., Mwebaze, T. and Katusiimeh, M., 2012. *The political economy of the fisheries sector in Uganda: Ruling elites, implementation costs and industry interests* (No. 2012: 04). DIIS Working Paper.

Kobayashi, M., Msangi, S., Batka, M., Vannuccini, S., Dey, M.M. and Anderson, J.L., 2015. Fish to 2030: the role and opportunity for aquaculture. *Aquaculture economics & management*, *19*(3), pp.282-300.

Kochzius, M. and Universiteit, V., 2009. Trends in fishery genetics. In *The Future of Fisheries Science in North America* (pp. 453-493). Springer Netherlands.

Kooloos, W.M., Wessels, J.A., van der Straaten, T., Huizinga, T.W. and Guchelaar, H.J., 2009. Criteria for the selection of single nucleotide polymorphisms in pathway

pharmacogenetics: TNF inhibitors as a case study. *Drug discovery today*, *14*(17), pp.837-844.

Kumar, N.S. and Gurusubramanian, G., 2011. Random amplified polymorphic DNA (RAPD) markers and its applications. *Sci Vis*, *11*(3), pp.116-124.

Kumar, S., Banks, T.W. and Cloutier, S., 2012. SNP discovery through next-generation sequencing and its applications. *International Journal of Plant Genomics*, *2012*.

Lamaze, F.C., Sauvage, C., Marie, A., Garant, D. and Bernatchez, L., 2012. Dynamics of introgressive hybridization assessed by SNP population genomics of coding genes in stocked brook charr (*Salvelinus fontinalis*). *Molecular Ecology*, *21*(12), pp.2877-2895.

Langmead, B. and Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), pp.357-359.

Levitt, A., 1981. *"Tecopa Pupfish Declared Extinct--Removed From Endangered List". United States Fish and Wildlife Service*.

Lewontin, R.C. and Hubby, J.L., 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics*, *54*(2), p.595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), pp.2078-2079.

Liu, S., Sun, L., Li, Y., Sun, F., Jiang, Y., Zhang, Y., Zhang, J., Feng, J., Kaltenboeck, L., Kucuktas, H. and Liu, Z., 2014. Development of the catfish 250K SNP array for genome-wide association studies. *BMC research notes*, *7*(1), p.135.

Liu, S., Zhou, Z., Lu, J., Sun, F., Wang, S., Liu, H., Jiang, Y., Kucuktas, H., Kaltenboeck, L., Peatman, E. and Liu, Z., 2011. Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. *BMC genomics*, *12*(1), p.53.

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

Lynch, M. and Milligan, B.G., 1994. Analysis of population genetic structure with RAPD markers. *Molecular ecology*, *3*(2), pp.91-99.

Ma, L., Runesha, H.B., Dvorkin, D., Garbe, J.R. and Da, Y., 2008. Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies. *BMC bioinformatics*, *9*(1), p.315.

Maher, C.A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N. and Chinnaiyan, A.M., 2009. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, *458*(7234), pp.97-101.

Malanczuk, P., 1995. Sustainable development: some critical thoughts in the light of the Rio Conference. *Sustainable Development and Good Governance*, pp.32-33.

Masek, T., Vopalensky, V., Suchomelova, P. and Pospisek, M., 2005. Denaturing RNA electrophoresis in TAE agarose gels. *Analytical biochemistry*, *336*(1), pp.46-50.

Matthews, G.V.T., 1993. The Ramsar Convention on Wetlands: its history and development. Gland: Ramsar convention bureau.

Matukumalli, L.K., Lawley, C.T., Schnabel, R.D., Taylor, J.F., Allan, M.F., Heaton, M.P., O'connell, J., Moore, S.S., Smith, T.P., Sonstegard, T.S. and Van Tassell, C.P., 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PloS one*, *4*(4), p.e5350.

Miller, R.R., Williams, J.D. and Williams, J.E., 1989. Extinctions of North American fishes during the past century. *Fisheries, 14*(6), pp.22-38.

Ministry of Agriculture, Animal Industries and Fisheries (MAAIF). (2011). Policy statement of Ministry of Agriculture Animal Industry & Fisheries, Uganda Government 2010/2011.

Mlewa, C.M. and Green, J.M., 2006. Translocation of marbled African lungfish, Protopterusaethiopicus (Telostei: Protopteridae), and its fishery in Lake Baringo, Kenya. *African Journal of Aquatic Science, 31*(1), pp.131-136.

Morin, P.A., Hancock, B.L. and George, J.C., 2007. *Development and application of single nucleotide polymorphisms (SNPs) for bowhead whale population structure analysis*. Paper SC/59/BRG8 presented to the Scientific Committee of the International Whaling Commission.

Morin, P.A., Luikart, G. and Wayne, R.K., 2004. SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, *19*(4), pp.208-216.

Mowbray, A., The Interlaken Declaration^ The Beginning of a New Era for the European Court of Human Rights?'(2010). *Human Rights Law Review, 10,* p.519.

Musinguzi, L., Efitre, J., Odongkara, K., Ogutu-Ohwayo, R., Muyodi, F., Natugonza, V., Olokotum, M., Namboowa, S. and Naigaga, S., 2016. Fishers' perceptions of climate change, impacts on their livelihoods and adaptation strategies in environmental change hotspots: a case of Lake Wamala, Uganda. *Environment, Development and Sustainability*, *18*(4), pp.1255-1273.

Mutz, K.O., Heilkenbrinker, A., Lönne, M., Walter, J.G. and Stahl, F., 2013. Transcriptome analysis using next-generation sequencing. *Current opinion in biotechnology*, *24*(1), pp.22-30.

Mwanja, W. W.2007. Freshwater fish seed resources in Uganda, East Africa, pp461─ 471. In: M.G. Bondad –Reantaso (ed.). Assessment of Freshwater Fish Seed Resources for Sustainable Aquaculture. FAO Fisheries Technical Paper. No. 501. Rome, FAO. 2007.

NaFIRRI. National Fisheries Resources Research Institute Annual Report 2009/2010.

Narayanan, S., 1991. Applications of restriction fragment length polymorphism. *Annals of Clinical & Laboratory Science*, *21*(4), pp.291-296.

National Aquaculture Sector Overview. Uganda. National Aquaculture Sector Overview Fact Sheets. Text by Mwanja, W.W.In: *FAO Fisheries and Aquaculture Department* [online]. Rome. Updated 19 July 2005. [Cited 21 June 2017]. http://www.fao.org/fishery/countrysector/naso_uganda/en

Newman, D. and Pilson, D., 1997. Increased probability of extinction due to decreased genetic effective population size: experimental populations of Clarkia pulchella. *Evolution*, pp.354-362.

Noecker, R.J., 1998, January. Endangered species list revisions: a summary of delisting and downlisting. Congressional Research Service, Library of Congress.

Nunan, F., 2014. Wealth and welfare? Can fisheries management succeed in achieving multiple objectives? A case study of Lake Victoria, East Africa. *Fish and Fisheries*, *15*(1), pp.134-150.

Omer, A. and Abukashawa, S., 2012. Morphometric Traits and Karyotypic Features of the African Lungfish (Um Koro) *Protopterus annectens annectens* (Owen, 1839) and *Protopterus aethiopicus aethiopicus* (Heckel, 1851) in Sudan.

Ondov, B.D., Bergman, N.H. and Phillippy, A.M., 2011. Interactive metagenomic visualization in a Web browser. *BMC bioinformatics*, *12*(1), p.385.

Patel, R.K. and Jain, M., 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS one*, *7*(2), p.e30619.

Peischl, S. and Excoffier, L., 2015. Expansion load: recessive mutations and the role of standing genetic variation. *Molecular Ecology*, *24*(9), pp.2084-2094.

Perkel, J., 2008. SNP genotyping: six technologies that keyed a revolution. *Nature Methods*, *5*(5), pp.447-453.

Peterson, G.W., Dong, Y., Horbach, C. and Fu, Y.B., 2014. Genotyping-by-sequencing for plant genetic diversity analysis: a lab guide for SNP genotyping. *Diversity*, *6*(4), pp.665-680.

Portillo, M., Fenoll, C. and Escobar, C., 2006. Evaluation of different RNA extraction methods for small quantities of plant tissue: Combined effects of reagent type and homogenization procedure on RNA quality-integrity and yield. *Physiologia Plantarum*, *128*(1), pp.1-7.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. and Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), pp.559-575.

Qian, X., Ba, Y., Zhuang, Q. and Zhong, G., 2014. RNA-Seq technology and its application in fish transcriptomics. *Omics: a journal of integrative biology*, *18*(2), pp.98-110.

Ramos, A.M., Crooijmans, R.P., Affara, N.A., Amaral, A.J., Archibald, A.L., Beever, J.E., Bendixen, C., Churcher, C., Clark, R., Dehais, P. and Hansen, M.S., 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PloS one*, *4*(8), p.e6524.

Ramsar Convention on Wetlands.*"The Annotated Ramsar List: Uganda, "2013*. Retrieved *28 June* 2017.

Rasmussen, H.B., 2012. Restriction fragment length polymorphism analysis of PCR-amplified fragments (PCR-RFLP) and gel electrophoresis-valuable tool for genotyping and genetic fingerprinting. In *Gel Electrophoresis-Principles and Basics*. InTech.

Reed, D.H. and Frankham, R., 2003. Correlation between fitness and genetic diversity. *Conservation biology*, *17*(1), pp.230-237.

Russell, J.R., Fuller, J.D., Macaulay, M., Hatz, B.G., Jahoor, A., Powell, W. and Waugh, R., 1997. Direct comparison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. *TAG Theoretical and Applied Genetics*, *95*(4), pp.714-722.

Rutaisire, J., Nandi, S. and Sundaray, J.K., 2017. A review of Uganda and Indias freshwater aquaculture: Key practices and experience from each country. *Journal of Ecology and The Natural Environment*, *9*(2), pp.15-29.

Saccheri I, Kuussaari M, Kankare M, Vikman P, Fortelius W, Hanski I (1998). Inbreeding and extinction in a butterfly metapopulation. Nature 392: 491–493.

Saitou, N. and Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, *4*(4), pp.406-425.

Sánchez, C.C., Smith, T.P., Wiedmann, R.T., Vallejo, R.L., Salem, M., Yao, J. and Rexroad, C.E., 2009. Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *Bmc Genomics*, *10*(1), p.559.

Scheiner, S.M., 1993. Genetics and evolution of phenotypic plasticity. *Annual review of ecology and systematics*, *24*(1), pp.35-68.

Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E.A., Liu, Y., Weinstock, G.M., Wheeler, D.A., Gibbs, R.A. and Yu, F., 2010. A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome research*, *20*(2), pp.273-280.

Shubin, N., 2008. *Your inner fish: a journey into the 3.5-billion-year history of the human body*. Vintage.

Skolnick, M.H. and White, R., 1982. Strategies for detecting and characterizing restriction fragment length polymorphisms (RFLP's). *Cytogenetic and Genome Research*, *32*(1-4), pp.58-67.

Slate, J., Gratten, J., Beraldi, D., Stapley, J., Hale, M. and Pemberton, J.M., 2009. Gene mapping in the wild with SNPs: guidelines and future directions. *Genetica*, *136*(1), pp.97-107.

Ssebisubi, M., 2011. Analysis Of Small-Scale Fisheries' Value-Chains In Uganda. University of Akreyri.

Suh, Y. and Vijg, J., 2005. SNP discovery in associating genetic variation with human disease phenotypes. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *573*(1), pp.41-53.

Thomson, K.S., 1972. An attempt to reconstruct evolutionary changes in the cellular DNA content of lungfish. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology*, *180*(3), pp.363-371.

Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, *14*(2), pp.178-192.

Varshney, R.K., Singh, V.K., Hickey, J.M., Xun, X., Marshall, D.F., Wang, J., Edwards, D. and Ribaut, J.M., 2016. Analytical and decision support tools for genomics-assisted breeding. *Trends in plant science*, *21*(4), pp.354-363.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T.V.D., Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M. and Zabeau, M., 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic acids research*, *23*(21), pp.4407-4414.

Walakira, J., 2013. *Culturing African Lungfish (Protopterus sp) in Uganda: Prospects, Performance in tanks, potential pathogens, and toxicity of salt and formalin* (Doctoral dissertation).

Walakira, J., 2015. Prospects and Potential of the African Lungfish (*Protopterus Spp*).: an Alternative Source of Fish Farming Livelihoods in Subsaharan Marbled African Lungfish (*Protopterus aethiopicus*). (A report to the International Center for Aquaculture and Aquatic Environments, Auburn University)

Walakira, J., Atukunda, G., Molnar, J. and Veverica, K., 2012. Prospects and Potential for Aquaculture of African Lungfish in Uganda. *World Aquaculture*, *43*(3), p.38.

Walakira, J., Boyd, C. and Molnar, J.J., 2015. Development of Low-cost Captive Breeding and Hatching Technologies for the African Lungfish (*Protopterus aethiopicus* and *P.*

amphibius) to Improve Livelihoods, Nutrition, and Income for Vulnerable Communities in Uganda. (A report to Aquaculture Research and Development Center, Kajjansi, Uganda; National Fisheries Resources Research Institute, Kampala, Uganda; Auburn University, Auburn, Alabama, USA).

Walakira, J., Morris, A., Sarah, O., Molnar, J.J., Veverica, K., Anthony, W., Phyllis, A., Godfrey, K., Constatine, O., 2016. Guiding Captive Breeding of African Lungfish *Protopterusaethiopicus* In Uganda: Genetic Diversity and Sex Determination. (A report to Biosciences eastern and central Africa – International Livestock Research Centre (BecA-ILRI Hub)

Walakira, J.K., Molnar, J.J., Phelps, R. and Terhune, J., 2014. Culturing the African lungfish in Uganda: Effects of exogenous fish feed on growth performance in tanks. *Uganda Journal of Agricultural Sciences*, *15*(2), pp.137-155.

Wang, Z., Gerstein, M. and Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), pp.57-63.

Wigginton, J.E., Cutler, D.J. and Abecasis, G.R., 2005. A note on exact tests of Hardy-Weinberg equilibrium. *The American Journal of Human Genetics*, *76*(5), pp.887-893.

Wit, P., Pespeni, M.H., Ladner, J.T., Barshis, D.J., Seneca, F., Jaris, H., Therkildsen, N.O., Morikawa, M. and Palumbi, S.R., 2012. The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular ecology resources*, *12*(6), pp.1058-1067.

Witte, F. and De Winter, W., 1995. Appendix II. Biology of the major fish species of Lake Victoria. *Fish stocks and fisheries of Lake Victoria-A handbook for field observations*, pp.301-320.

Wright, S., 1922. Coefficients of inbreeding and relationship. *The American Naturalist*, *56*(645), pp.330-338.

Yáñez, J.M., Houston, R.D. and Newman, S., 2014. Genetics and genomics of disease resistance in salmonid species. *Frontiers in genetics*, *5*, p.415.

Yang, Y., Xu, M., Luo, Q., Wang, J. and Li, H., 2014. *De novo* transcriptome analysis of Liriodendron chinense petals and leaves by Illumina sequencing. *Gene*, *534*(2), pp.155-162.

Yu, Y., Wei, J., Zhang, X., Liu, J., Liu, C., Li, F. and Xiang, J., 2014. SNP discovery in the transcriptome of white Pacific shrimp *Litopenaeus vannamei* by next generation sequencing. *PLoS One*, *9*(1), p.e87218.

Zhang, R., Zhu, Z., Zhu, H., Nguyen, T., Yao, F., Xia, K., Liang, D. and Liu, C., 2005. SNP Cutter: a comprehensive tool for SNP PCR.

APPENDICES

APPENDIX I

Table 4. 4 Summary of the statistics of the RNA-Sequenced data before and after the quality control

Lake	Filename	No of raw reads	No of reads after quality control	% GC content before	% GC content After	Sequence length before	Sequence length	Contigs$
Bisina	S1_R1.fastq	2878583	2672282	44	43	35-301	1-206	
	S1_R2.fastq	2878583	2519384	44	43	35-301	1-196	
	S2_R1.fastq	3166539	2946214	44	43	35-301	1-206	
	S2_R2.fastq	3166539	2785479	44	43	35-301	1-196	
	S3_R1.fastq	2213642	2072499	43	43	35-301	1-206	
	S3_R2.fastq	2213642	1932850	44	43	35-301	1-196	172903
Edward	S4_R1.fastq	1673223	1583050	43	43	35-301	1-206	
	S4_R2.fastq	1673223	1469248	43	42	35-301	1-196	
	S5_R1.fastq	1386748	1315541	43	42	35-301	1-206	
	S5_R2.fastq	1386748	1232226	43	42	35-301	1-196	
	S6_R1.fastq	2212235	2051722	43	43	35-305	1-206	
	S6_R2.fastq	2212235	1942903	43	43	35-301	1-196	127255
George	S7_R1.fastq	1474786	1403316	44	43	35-301	1-206	
	S7_R2.fastq	1474786	1258838	44	43	35-301	1-196	
	S8_R1.fastq	1819719	1711870	44	44	35-301	1-206	
	S8_R2.fastq	1819719	1597879	44	44	35-301	1-196	
	S9_R1.fastq	2613331	2448386	43	44	35-301	1-206	
	S9_R2.fastq	2613331	2273074	44	43	35-301	1-196	124536
Kyoga	S10_R1.fastq	2874105	2702067	43	43	35-301	1-206	
	S10_R2.fastq	2874105	2555715	44	43	35-301	1-196	
	S11_R1.fastq	1716553	1594853	43	43	35-301	1-206	
	S11_R2.fastq	1716553	1478019	43	43	35-301	1-196	
	S12_R1.fastq	1051271	951509	53	52	35-301	1-206	
	S12_R2.fastq	1051271	868551	54	52	35-301	1-196	99747
Nawampasa	S13_R1.fastq	2990300	2744651	44	43	35-301	1-206	
	S13_R2.fastq	2990300	2252437	44	42	35-301	1-196	
	S14_R1.fastq	1957897	1772344	44	44	35-301	1-206	
	S14_R2.fastq	1957897	1517508	44	43	35-301	1-196	
	S15_R1.fastq	2451753	2236813	44	43	35-301	1-206	
	S15_R2.fastq	2451753	1867161	44	42	35-301	1-196	106054
Wamala	S16_R1.fastq	2286263	2115702	44	43	35-301	1-206	
	S16_R2.fastq	2286263	1760346	44	43	35-301	1-196	
	S17_R1.fastq	3109839	2927513	44	44	35-301	1-206	
	S17_R2.fastq	3109838	2477115	44	43	35-301	1-196	
	S18_R1.fastq	2393689	2212939	44	43	35-301	1-206	
	S18_R2.fastq	2393689	1939327	44	43	35-301	1-196	123307
	Total	80540951	71191331					753802

S –sample, **R1.fastq**- Forward reads for specified sample (S), **R2.fastq**- Reverse reads for the specified sample (S), **No** – number, **% GC content** – percentage Guanine-cytosine content, Contig$ - total number of contigs for three sample pooled per lake.

APPENDIX II

Table 4. 5 The selected panel of Single Nucleotide Polymorphisms to guide in genetic diversity analysis and breeding purposes

SNP ID	H value	GC %	SNP ID	H value	GC Content
PetTR103219|C0_G3_I1_263	0.5	50.7	PetTR50205|C0_G1_I1_736	0.5	41.4
PetTR109486|C3_G9_I1_446	0.5	59.3	PetTR50466|C0_G1_I1_55	0.5	45.6
PetTR109726|C0_G1_I1_150	0.5	45.7	PetTR51535|C1_G11_I1_307	0.5	44.6
PetTR109726|C0_G1_I1_151	0.5	45.0	PetTR51535|C1_G11_I1_46	0.5	44.8
PetTR109726|C0_G1_I1_191	0.5	47.1	PetTR51535|C1_G35_I1_398	0.5	40.4
PetTR109726|C0_G1_I1_193	0.5	46.4	PetTR51535|C1_G35_I1_94	0.5	43.6
PetTR109726|C0_G1_I1_264	0.5	45.0	PetTR51603|C0_G2_I1_204	0.5	42.1
PetTR109726|C0_G1_I1_265	0.5	44.3	PetTR51603|C0_G2_I1_217	0.5	42.1
PetTR109726|C0_G1_I1_305	0.5	46.4	PetTR51833|C2_G7_I1_547P	0.5	40.0
PetTR109726|C0_G1_I1_307	0.5	45.7	PetTR51928|C0_G1_I1_597	0.5	46.4
PetTR109726|C0_G1_I1_337	0.5	43.7	PetTR51928|C0_G1_I1_602	0.5	46.4
PetTR112977|C6_G14_I1_658	0.5	47.4	PetTR51928|C0_G1_I1_628	0.5	48.6
PetTR112977|C6_G14_I2_139	0.5	44.3	PetTR51928|C0_G1_I1_648	0.5	45.7
PetTR114372|C0_G1_I1_228	0.5	54.3	PetTR51928|C0_G1_I1_655	0.5	45.7
PetTR11580|C0_G1_I2_271	0.5	40.0	PetTR52214|C73_G108_I1_131	0.5	43.6
PetTR120337|C0_G1_I1_181	0.5	49.3	PetTR52214|C73_G108_I1_149	0.5	45.0
PetTR120337|C0_G1_I1_182	0.5	50.0	PetTR52869|C0_G2_I1_791	0.5	44.3
PetTR120337|C0_G1_I1_183	0.5	50.0	PetTR53037|C0_G1_I1_152	0.5	50.7
PetTR120337|C0_G1_I1_184	0.5	50.0	PetTR53042|C0_G2_I1_161	0.5	42.9
PetTR120337|C0_G1_I1_191	0.5	48.6	PetTR5431|C0_G1_I1_272	0.5	48.6
PetTR120337|C0_G1_I1_196	0.5	48.6	PetTR55314|C0_G1_I1_358	0.5	40.0
PetTR120337|C0_G1_I1_197	0.5	49.3	PetTR55411|C1_G1_I1_318	0.5	52.9
PetTR121315|C0_G2_I1_1002	0.5	45.7	PetTR56556|C0_G1_I1_159	0.5	40.7
PetTR121315|C0_G2_I1_1011	0.5	44.3	PetTR57556|C0_G32_I1_109	0.5	42.1
PetTR121315|C0_G2_I1_1048	0.5	44.3	PetTR57556|C0_G32_I1_324	0.5	44.3
PetTR121315|C0_G2_I1_1124	0.5	49.3	PetTR57556|C0_G32_I1_63	0.5	46.6
PetTR121315|C0_G2_I1_1169	0.5	52.9	PetTR60691|C0_G2_I3_134	0.5	43.6
PetTR121315|C0_G2_I1_290	0.5	42.9	PetTR60691|C0_G2_I3_295	0.5	45.7
PetTR121315|C0_G2_I1_530	0.5	41.4	PetTR60691|C0_G2_I3_389	0.5	42.9
PetTR121315|C0_G2_I1_992	0.5	45.7	PetTR60691|C0_G4_I1_1172	0.5	40.0
PetTR121315|C0_G2_I1_993	0.5	45.7	PetTR60691|C0_G4_I1_1173	0.5	40.7
PetTR121884|C0_G1_I1_370	0.5	50.7	PetTR60691|C0_G4_I1_1185	0.5	43.6
PetTR122542|C0_G1_I1_645	0.5	48.6	PetTR60691|C0_G4_I1_927	0.5	45.7
PetTR122542|C0_G1_I1_661	0.5	47.9	PetTR60691|C0_G4_I1_938	0.5	47.1
PetTR124147|C0_G1_I1_383	0.5	45.0	PetTR60691|C0_G4_I1_939	0.5	46.4
PetTR12634|C0_G4_I1_199	0.5	50.0	PetTR61166|C0_G1_I1_625	0.5	50.0
PetTR12634|C0_G4_I2_369	0.5	49.1	PetTR61207|C0_G2_I1_468	0.5	47.9
PetTR12634|C0_G4_I3_150	0.5	50.0	PetTR61305|C0_G1_I1_142	0.5	48.6
PetTR12634|C0_G4_I3_156	0.5	51.4	PetTR61305|C0_G1_I1_143	0.5	48.6
PetTR15740|C0_G1_I3_405	0.5	42.9	PetTR62386|C1_G5_I5_150	0.5	40.0
PetTR17252|C0_G3_I1_2148	0.5	43.6	PetTR62386|C1_G5_I5_246	0.5	42.1
PetTR17686|C0_G1_I1_78	0.5	42.1	PetTR62386|C1_G5_I5_328	0.5	41.4
PetTR18251|C0_G3_I1_284	0.5	49.2	PetTR62386|C1_G5_I5_462	0.5	42.9
PetTR18251|C0_G3_I1_294	0.5	49.2	PetTR62386|C1_G5_I5_473	0.5	41.5
PetTR18412|C13_G9_I3_129	0.5	43.6	PetTR63514|C1_G3_I4_420	0.5	44.3
PetTR18412|C13_G9_I3_473	0.5	44.3	PetTR65319|C10_G1_I2_151	0.5	44.3
PetTR20340|C2_G8_I2_297	0.5	47.1	PetTR65319|C10_G1_I2_485	0.5	47.4
PetTR21209|C0_G1_I1_1749	0.5	42.1	PetTR67684|C0_G1_I1_2390	0.5	41.4
PetTR21359|C0_G1_I1_1151	0.5	40.7	PetTR67916|C0_G2_I2_169	0.5	42.1
PetTR23671|C0_G1_I2_302	0.5	57.1	PetTR67916|C0_G2_I2_502	0.5	41.4
PetTR23671|C0_G1_I2_303	0.5	57.1	PetTR68818|C0_G1_I1_1669	0.5	47.9
PetTR23671|C0_G1_I2_324	0.5	54.3	PetTR69558|C1_G4_I3_369	0.5	48.6
PetTR23671|C0_G1_I2_390	0.5	52.9	PetTR70076|C0_G5_I2_159	0.5	40.0
PetTR23671|C0_G1_I2_96	0.5	56.4	PetTR70076|C0_G5_I2_375	0.5	40.0
PetTR24831|C0_G2_I1_195	0.5	56.4	PetTR72395|C0_G2_I1_490	0.5	49.3
PetTR24831|C0_G2_I1_208	0.5	55.7	PetTR73294|C0_G1_I1_567	0.5	42.9
PetTR24924|C0_G2_I1_462	0.5	44.3	PetTR73308|C0_G1_I1_811	0.5	51.4
PetTR25439|C1_G5_I3_40	0.5	42.7	PetTR76132|C3_G1_I6_385	0.5	47.1
PetTR25439|C1_G5_I3_44	0.5	41.2	PetTR77355|C2_G2_I3_154	0.5	42.1

SNP ID	H value	GC %	SNP ID	H value	GC Content
PetTR25439|C1_G5_I3_45	0.5	40.9	PetTR77355|C2_G2_I3_89	0.5	42.1
PetTR26081|C0_G1_I1_1001	0.5	47.9	PetTR79684|C0_G1_I1_560	0.5	45.0
PetTR26081|C0_G1_I1_1006	0.5	49.3	PetTR81723|C1_G9_I5_511	0.5	47.9
PetTR26436|C7_G9_I1_44	0.5	41.2	PetTR82292|C0_G2_I8_255	0.5	40.7
PetTR29801|C0_G1_I1_704	0.5	54.3	PetTR82292|C0_G2_I8_521	0.5	40.7
PetTR30087|C0_G2_I2_251	0.5	43.6	PetTR8275|C0_G1_I1_364	0.5	41.4
PetTR32783|C0_G4_I1_418	0.5	46.4	PetTR8275|C0_G1_I1_386	0.5	43.6
PetTR32950|C1_G13_I8_1053	0.5	47.1	PetTR82862|C3_G17_I4_570	0.5	48.6
PetTR33432|C1_G2_I1_760	0.5	40.0	PetTR82862|C3_G17_I4_620	0.5	47.9
PetTR33432|C1_G2_I1_761	0.5	40.7	PetTR82862|C3_G17_I4_621	0.5	47.9
PetTR35485|C0_G1_I1_168	0.5	42.1	PetTR84011|C0_G1_I1_757	0.5	44.3
PetTR35485|C0_G1_I1_71	0.5	42.1	PetTR85519|C0_G1_I2_481	0.5	41.2
PetTR35527|C12_G4_I1_85	0.5	54.3	PetTR85563|C5_G2_I6_200	0.5	53.6
PetTR35544|C0_G1_I1_384	0.5	40.0	PetTR85563|C5_G2_I6_208	0.5	51.4
PetTR37080|C0_G1_I1_625	0.5	60.7	PetTR85563|C5_G2_I6_210	0.5	51.4
PetTR37796|C0_G6_I1_540	0.5	43.0	PetTR8749|C1_G28_I1_108	0.5	45.0
PetTR38968|C0_G17_I4_108	0.5	48.6	PetTR8749|C1_G28_I1_18	0.5	50.0
PetTR38968|C0_G17_I4_112	0.5	47.9	PetTR8749|C1_G28_I1_324	0.5	45.7
PetTR38968|C0_G17_I4_313	0.5	42.9	PetTR8749|C1_G28_I1_87	0.5	46.4
PetTR39150|C0_G1_I1_1319	0.5	48.6	PetTR87679|C0_G1_I1_728	0.5	44.3
PetTR41077|C0_G1_I1_1037	0.5	42.1	PetTR88807|C0_G1_I1_460	0.5	40.0
PetTR41528|C0_G1_I1_218	0.5	55.7	PetTR89325|C6_G7_I1_597	0.5	49.3
PetTR42123|C0_G2_I1_24	0.5	44.7	PetTR89629|C1_G1_I1_651	0.5	42.1
PetTR42123|C0_G2_I1_27	0.5	44.3	PetTR89629|C1_G1_I1_671	0.5	40.7
PetTR42136|C0_G1_I1_861	0.5	50.0	PetTR89629|C1_G1_I1_678	0.5	41.4
PetTR42164|C56_G90_I3_213	0.5	40.0	PetTR89629|C1_G1_I1_679	0.5	41.4
PetTR42164|C56_G90_I3_275	0.5	50.0	PetTR90340|C1_G2_I1_398	0.5	40.7
PetTR42164|C56_G90_I3_289	0.5	56.4	PetTR90340|C1_G2_I1_61	0.5	40.5
PetTR42164|C56_G90_I3_314	0.5	64.3	PetTR91000|C0_G1_I2_162	0.5	41.4
PetTR42164|C56_G90_I3_425	0.5	53.6	PetTR91000|C0_G1_I2_378	0.5	45.0
PetTR42164|C56_G90_I3_612	0.5	40.0	PetTR91000|C0_G1_I2_404	0.5	42.1
PetTR42164|C56_G90_I3_685	0.5	55.7	PetTR91000|C0_G1_I2_420	0.5	42.1
PetTR45227|C0_G1_I1_468	0.5	46.4	PetTR93311|C0_G1_I1_234	0.5	49.6
PetTR45885|C0_G1_I1_611	0.5	46.4	PetTR93490|C2_G2_I1_269	0.5	52.7
PetTR46179|C2_G17_I1_287	0.5	41.4	PetTR93964|C0_G2_I1_743	0.5	40.7
PetTR46179|C2_G17_I1_29	0.5	44.4	PetTR94129|C0_G1_I1_1350	0.5	45.0
PetTR47794|C0_G4_I1_23	0.5	43.0	PetTR94306|C0_G2_I2_86	0.5	40.0
PetTR47794|C0_G4_I1_355	0.5	45.7	PetTR94858|C0_G1_I1_789	0.5	47.9
PetTR49302|C0_G3_I1_166	0.5	44.3	PetTR95943|C0_G1_I1_315	0.5	48.6
PetTR49302|C0_G3_I1_176	0.5	44.3	PetTR97692|C0_G1_I1_454	0.5	51.4