# Characterisation of SNPs associated with growth rate in dusky kob (*Argyrosomus japonicus*), using exome sequencing

by

**Tassin Jackson**

*Thesis presented in partial fulfilment of the requirements for the degree of Master of Science at Stellenbosch University*

**Supervisor: Clint Rhode, Ph.D., Pr.Sci.Nat.**

**Department of Genetics**

March 2020

# Declaration

---

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

March 2020

# Abstract

Marine living-resources such as dusky kob, (*Argyrosomus japonicus*) are particularly vulnerable to overfishing as this species has been targeted for decades by commercial, recreational and subsistence fisheries, which has led to the steady decline in the natural populations. A shift towards aquaculture as a sustainable alternative supply to the market has been initiated, with considerable efforts being made to understand the fundamental role that genes play in the biological processes influencing complex traits such as growth rate. Although a few studies have been conducted on the species, they have been hindered by the limited number of genomic resources, which is an issue that affects many non-model species. Therefore, this study aimed to investigate the transferability of a model organism's exon capture kit in a non-model species for the development of SNP markers associated with growth. By using 16 dusky kob individuals for exome sequencing this study was able to capture 6,623 of the 346,263 exons found within the model organisms, zebrafish, as well as a large number of exons that could potentially be species-specific. Overall, the exome data proved to be a valuable resource for the identification of variants, with variant detection identifying 4.5 million potential molecular makers with a total of 2.8 million putative SNPs and 3,276 tandem repeats. These variants were spread across the exome regions with a SNP occurring approximately every 1000 nt. Using the candidate gene approach and a selection of 15 gene regions, 263 putative SNPs were identified, of which 38 SNPs in nine genes were confirmed using Sanger sequencing and identified as having a potential association to the trait of interest. Association of these markers was analysed by performing both case-control and quantitative analyses using 80 individuals (classified as large and small) of dusky kob. These analyses were able to identify eight SNPs in three key genes. This study demonstrated the ability of exon capture to be customised for cross-species capture to assist in molecular marker discovery for non-model organisms with limited or no genomic resources. Resources which could be used for the development of markers which could assist in the implementation of marker assisted selection (MAS), which will aid in the development and effective utilisation of the species.

# Opsomming

Mariene lewende hulpbronne soos die boerkabeljou (*Argyrosomus japonicus*) is veral kwesbaar vir oorbevissing, aangesien hierdie spesie al dekades lank deur kommersiële, ontspannings- en bestaansvisserye geteiken word, wat gelei het tot die bestendige afname van die natuurlike populasie. Akwakultuur bied 'n volhoubare alternatiewe oplossing aan die mark, en toenemende pogings word aangewend om die fundamentele rol van gene in biologiese prosesse van komplekse eienskappe, soos groeitempo, te verstaan. Ongelukkig word studies in hierdie spesie, net soos in ander nie-modelspesies, belemmer deur die beskikbaarheid van ‘n beperkte aantal genomiese hulpbronne. Daarom het hierdie studie ten doel gehad om die oordraagbaarheid van die eksonvangsstel (“exon capture kit”) van 'n modelorganisme in 'n nie-modelspesie te ondersoek, met die oog op die ontwikkeling van ENP-merkers wat met groeitempo geassosieer word. Hierdie studie het deur middel van eksoomvolgordebepaling op 16 boerkabeljou individue daarin geslaag om 6,623 uit 346,263 eksone van die model organisme, zebravis, sowel as ’n groot aantal moontlike spesie-spesfieke eksone vas te vang. Die ontdekking van 4.5 miljoen potensiele molekulere merkers, waarvan 2.8 miljoen moontlike ENP merkers en 3,726 tandem herhalings, dui daarop aan dat die eksoomdata ’n waardevolle hulpbron vir die identifisering van genetiese variasie is. Hierdie variante was verspreid oor die eksoomareas, met ’n ENP wat ongeveer elke 1000 nt voorkom. Met behulp van die kandidaatgeenbenadering en ’n seleksie wat 15 geenstreke behels, is 263 veronderstelde ENPs geidentifiseer, waarvan 38 ENPs in nege gene van sanger-volgorde bevestig was, en getoon het om moontlike assosiasie met die eienskap van belangstelling, groei, te toon. Bimodale gevallestudie en kwantitatiewe analises is uitgevoer deur gebruik te maak van 80 boerkabeljou individue (wat geklassifiseer is as klein en groot) om die assosiasie tussen merkers en groei te ondersoek. Hierdie analises het gelei tot die identifisering van ag ENPs in drie sleutelgene. Hierdie studie het getoon dat dit moontlik is om ’n eksonvangsstel aan te pas vir gebruik in ander spesies om te help met die ontdekking van molekulere merkers in nie-model organismes met beperkte of geen genomiese hulpbronne. Daarmee help hierdie studie om genomiese hulpbronne op te bou, wat kan lei tot die ontwikkeling van molekulere merkers wat gebruik kan word om merker bemiddelde seleksie (MBS) toe te pas, om sodoende die optimale benutting van hierdie spesie te bereik.

# Acknowledgements

I would like to extend my gratitude to the Department of Science and Technology, the National Research Foundation of South Africa, and Stellenbosch University for financial support. My gratitude also goes out to the members of the Molecular Breeding and Biodiversity research group for all their help and support. To my supervisor Dr Clint Rhode who always pushed me to be the best that I can be. Thank you for all the knowledge and inspiration over the last few years. Finally, I would like to thank my family, partner and friends for their support, particularly during the final stages of thesis writing, I could not have done it without each and every one of you.

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| % | Percentage |
| (Pty) | Property Limited |
| > | Greater than |
| < | Less than |
| ≥ | Greater than or equal to |
| ~ | Approximately |
| µg/l | micrograms per litre |
| µM | micromolar |
| ng/µl | nanogram per microlitre |
| 5' | Five prime |
| 3' | Three prime |
| A | Adenine |
| AFLP | Amplified Fragment Length Polymorphism |
| AIC | Akaike information criterion |
| BAC | Bacterial artificial chromosome |
| bp | Base pair |
| BIC | Bayesian information criterion |
| *BMP2A* | Bone morphogenic protein 2 a |
| Bn | Billion |
| C | Cytosine |
| °C | Degree Celsius |
| cDNA | Complementary Deoxyribonucleic Acid |
| chr | chromosome |
| CI/s | Confidence interval/s |
| cm | centimetre |
| CTAB | Cetyl Trimethylammonium Bromide |
| CV | Coefficient of variance |
| DAFF | Department of Agriculture, Forestry and Fisheries |
| ddRAD | Double digest restriction-site associated deoxyribonucleic acid |

| | |
|---|---|
| dph | days post hatch |
| DNA | Deoxyribonucleic acid |
| dNTP | Deoxynucleotide triphosphate |
| *e.g.* | exempli gratia (for example) |
| EST | Expressed sequence tag |
| EtBr | Ethidium bromide |
| et al. | et alii (and others) |
| etc | et cetera |
| E-value | Expectation value |
| ezRAD | Novel strategy for restriction site–associated DNA |
| F | Forward primer |
| F1 | First-generation |
| FAO | Food and Agriculture Organisation |
| FBC | Family bias corrected cohort |
| G | Guanine |
| g | Grams |
| GAS | Gene-assisted selection |
| GB | Giga bases |
| GBS | Genotyping by Sequencing |
| GO | Gene ontology |
| h2 | (narrow-sense) heritability |
| HRM | High Resolution Melt |
| hrs | hours |
| HTS | High Throughput Sequencing |
| HWE | Hardy-Weinberg Equilibrium |
| *i.e.* | id est (that is to say) |
| ISP | Ion Sphere Particle |
| K | Fulton’s conditioning factor |
| kg | kilogram |
| KOG | EuKaryotic Orthologous Groups |
| KW | Kruskal-Wallis |

| | |
|---|---|
| L | Litres |
| LD | Linkage Disequilibrium |
| LD-MAS | Linkage disequilibrium with QTL |
| LE-MAS | Linkage disequilibrium with quantitative trait |
| Ls | Standard length |
| M | Million |
| MAF | Minor allele frequency |
| MAS | Marker-Assisted Selection |
| m | meters |
| min | minutes |
| ml | millilitres |
| mm | millimetres |
| mM | millimolar |
| mtDNA | mitochondrial Deoxyribonucleic Acid |
| mya | million years ago |
| *MYOD1* | Myogenic differentiation factor 1 |
| n | sample size |
| NCBI | National Centre for Biotechnology Information |
| ng | nanogram(s) |
| NR | Non-redundant |
| nsSNPs | Non-synonymous Single Nucleotide Polymorphism |
| nt | nucleotides |
| OR | Odds ratio |
| p | page |
| PCR | Polymerase Chain Reaction |
| PIC | Polymorphism Information Content |
| p-value | Probability value |
| QC | Quality Control |
| QTL/s | Quantitative trait locus/loci |
| r | Relatedness |
| r | Correlation coefficient |

| | |
|---|---|
| R | Reverse primer |
| R2 | Squared correlation coefficient |
| RAD-seq | Restriction site Associated DNA Sequencing |
| RAPD | Random Amplified Polymorphic DNA |
| RAS | Recirculating Aquaculture System |
| RFLP | Restriction Fragment Length Polymorphism |
| RNA | Ribonucleic acid |
| RRS | Reduced-Representation Sequencing |
| sec | second |
| SS | Solid Spine |
| SNP | Short Nucleotide Polymorphism |
| SSRs | Simple sequence repeats |
| STRs | Short tandem repeats |
| T | Thymine |
| t | tons |
| Ta | Annealing temperature |
| TDT | Transmission disequilibrium test |
| Tm | Melting temperature |
| TM | Trademark |
| *TNKSA* | Tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase a |
| ts/tv | Transition transversion ratio |
| USD | United States Dollar |
| UTRs | Untranslated regions |
| W | Bodyweight |
| WES | Whole-exome Sequencing |
| WGS | Whole Genome Sequencing |
| WGR | Whole Genome Resequencing |
| X | times |
| $\chi 2$ | chi-squared |

# CHAPTER 1

# Introduction: Literature Review, Aims and Objectives

---

## *1.1)* Species biology: An introduction to dusky kob *(Argyrosomus japonicus)*

### *1.1.1) Classification and Evolution of Dusky Kob*

As a member within the Actinopterygii class in the phylum Chordata, the Sciaenidae family is vast with about 280 species in 90 genera worldwide. They are primarily tropical and warm temperate coastal marine fishes with some species found to be confined to fresh water rivers (Chao et al., 2015). While the large majority live inshore over sandy or muddy bottoms, a few species are found in deep water and others have adapted to special habitats such as coral reefs and surf zones (Chao,1986). The genus *Argyrosomus*, found within the Sciaenidae family is represented by at least nine recognised species (Griffiths and Heemstra, 1995). The sciaenid species found in this genus all display a high degree of conservative morphology, which has resulted in the misidentification of many species, particularly those that inhabit a wide range of coastal areas. *Argyrosomus japonicus* has been known by at least 51 different common names and three trade names throughout its Indo-Pacific distribution occurring in the coastal waters of *i.e*., Australia, Africa, India, Pakistan, China, Korea and Japan (Bernatzeder and Britz, 2007; Griffiths and Heemstra, 1995; Kailola et al., 1993; Trewavas, 1977) (Figure 1.1). A study performed in 1990 indicated that *A. japonicus* had been misidentified and referred to as *A. hololepidotus* in both Australia and South Africa. This misidentification was discovered by pre-forming an in-depth study comparing the habitat distribution, morphometrics, otoliths and anatomical structure of the species within the genus. However, this was further complicated by the confusion of *A. japonicus* with *A. inodorus* (Griffiths and Heemstra, 1995) a species with which *A. japonicus* may occasionally hybridise within South Africa (Mirimin et al., 2014).

The wild populations of *A. japonicus* in South Africa and Australia have been considered conspecific as the populations could not be differentiated from one another following the revision of the genus *Argyrosomus* by Griffiths and Heemstra, (1995). The life history and biology of *A. japonicus* has been well studied in South Africa (Griffiths, 1996; Griffiths and Heemstra, 1995), and more recently in Australia (Bernatzeder and Britz, 2007; Ferguson et al., 2014; Silberschneider and Gray, 2007; Taylor et al., 2006). These studies have shown there to be significant differences in the life-history traits (*e.g.* growth, age at sexual maturity,

time of spawning) amongst the geographical locations. Using mitochondrial DNA, a study confirmed that there had been a long period of isolation between the South African and Australian populations, with each population potentially representative of a different species (Farmer, 2008). A revision of the taxonomy *A. japonicas* is, therefore, justified. For this thesis the focus will be on the South African *A. japonicus*, commonly known as dusky kob.

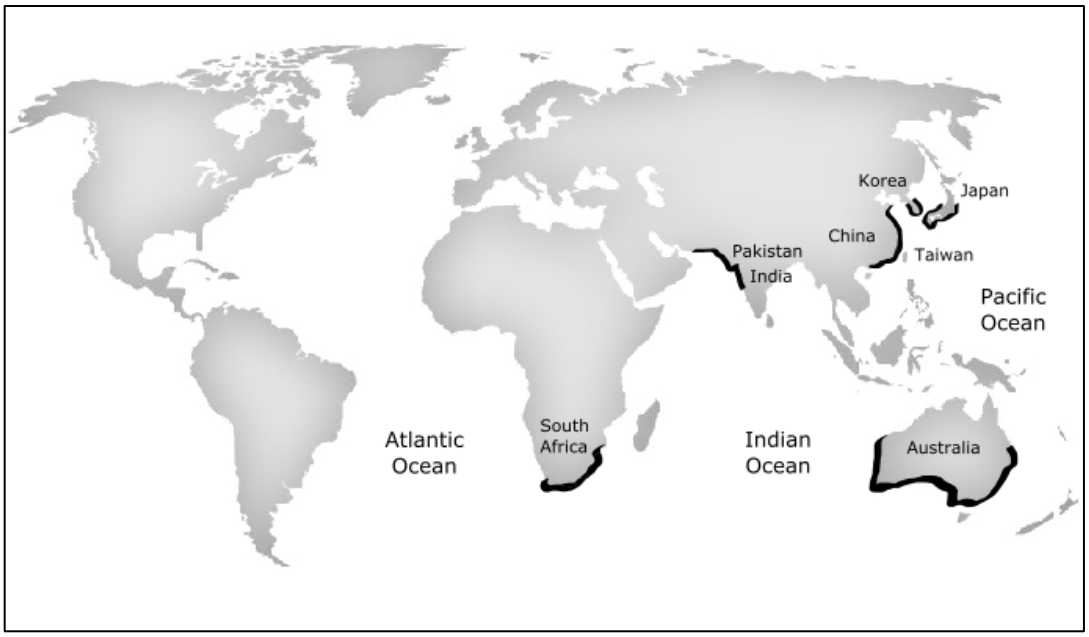



**Figure 1.1. The Indo-Pacific distribution of *Argyrosomus japonicus i.e.*, Australia, Africa, India, Pakistan, China, Korea and Japan. The figure was adapted from the original by Silberschneider and Gray (2007).**

### *1.1.2) Ecology, Distribution and Life-History in South Africa*

Dusky kob are known to be predatory fish that hunt using lateral line senses and smell instead of relying on their sight; this is a specialised adaption which is ideal for hunting in their muddy and murky environments (Griffiths, 1997). Adult fish have the ability to hunt throughout the water column, predominantly making use of an ambush strategy when feeding along the ocean floor. While the adults are mainly piscivorous, they are known to sometimes feed on squid and octopus when given the opportunity. The juveniles' diet however consists mainly of crustaceans and smaller fish (Bergamino et al., 2014; Griffiths, 1997). Over time this species has developed adaptive traits to fit their feeding style, such as a large mouth, sharp teeth for gripping, widely spaced gill rakers and a large rigid distensible stomach (Kailola et al., 1993). A notable trait of sciaenids is the ability to produce drumming sounds by vibrating their swim bladder. However, the pitch and use of croaking varies between species, with some males using it as a mating call (Ramcharitar et al., 2006). This phenomenon is linked to territorial display and spawning behaviour, and may reflect

adaptation to spawning at night and communication in habitats that are turbid, announcing hazards and location (Blaber, 2000; Roach et al., 2005). In some species the sonic muscle fibres are only present in males. These muscles which atrophy throughout the year, only strengthen during the mating season to assist in finding a mate. The croaker mechanism in other species such as dusky kob and *A. regius*, is found to be present in both sexes throughout the year (Griffiths and Heemstra, 1995; Lagardere and Mariani, 2006), with individuals able to produce up to several call variations (Parsons and McCauley, 2017). This ability allows for constant communication between individuals and populations, assisting in the survival reproductive success of the species; it can however be detrimental as constant acoustic communication allows for predators such as the bottlenose dolphin, to easily locate large groups of croaker and drum as they broadcast their position (Roach et al., 2005).

Dusky kob is the largest South African sciaenid reaching up to two meters in length and achieving a record weight of 75 kilograms (Griffiths 1997a; Griffiths and Hecht ,1995b). They are long-lived animals with some individuals being recorded to reach a maximum of 42 years of age. This longevity does however result in a late onset of sexual maturity, with silver and squaretail kob (*Argyromus thorpei*) maturing in less than half the time required for dusky kob. While silver and squaretail kob females attain sexual maturity at a length of 35cm, which is reached at approximately one and a half years of age. While dusky kob females only mature once reaching 1.1m in length or six years of age, male kob reach sexual maturity earlier at approximately 5 years of age or 900mm in length (Griffiths, 1997a). One of the main reasons for the species late onset of sexual maturity, is that unlike other kob species which show a consist growth rate post-maturity, dusky kob only divert their energy towards reproduction once the individual achieves a length greater than one meter, allowing the species to focus solely on growth. Dusky kob are migratory, spawning fish that are found to be abundant within South African waters. The primary distribution of the homogeneous genetic stock occurs between Cape Agulhas, located in the Western Cape and the southern Mozambique border (Griffiths, 1995b; Mirimin et al., 2015) with the species being particularly abundant between Cape Agulhas and KwaZulu-Natal as a result of warmer waters (Griffiths and Heemstra, 1995) (Figure 1.2). During the mating season the majority of the adult population migrate northward of the Cape to the warmer waters of KwaZulu-Natal where spawning activity coincides with the utilisation of predator-poor estuarine nurseries. This usually occurs between August and November, although dusky kob eggs have been observed in the coastal waters of KZN as early as July and as late as February (Connell et al., 2007). Due to differences in water temperature and oceanography along the coast, the

time of spawning varies, with spawning commencing in the northern regions above KwaZulu-Natal between winter and spring (August to November). While during the summer months (October to January), spawning commences in the southern and southern-eastern Cape Regions when adults return from KwaZulu-Natal (Griffiths, 1996). Some adult fish do not migrate to KwaZulu-Natal, but remain in the southern and southern-eastern Cape Regions to spawn in the summer months. Spawning occurs at night on shallow inshore reefs, pinnacles and wrecks at depths of 10-15m. The Sciaenidae family has adapted its spawning strategy to reduce predation on eggs by zooplanktivores whom primarily feed during the daylight as light intensity has been shown to directly affect successful foraging (Connell, 2007; Griffiths, 1996; Griffiths, 1997a; Skibinski, 2005). The dispersal of the eggs and larvae in and out of estuaries (<50m depth) along the South African coastline have been shown to be facilitated by the Agulhas Current which moves in a downward direction (Beckley 1993; Beckley, 1995; Harris et al., 1995). Dusky kob typically remain within their estuaries until reaching maturity but as they grow, they start to gradually move into deeper waters (5-120m) consisting mainly of soft substrata of sand or mud (Cowley et al., 2007).

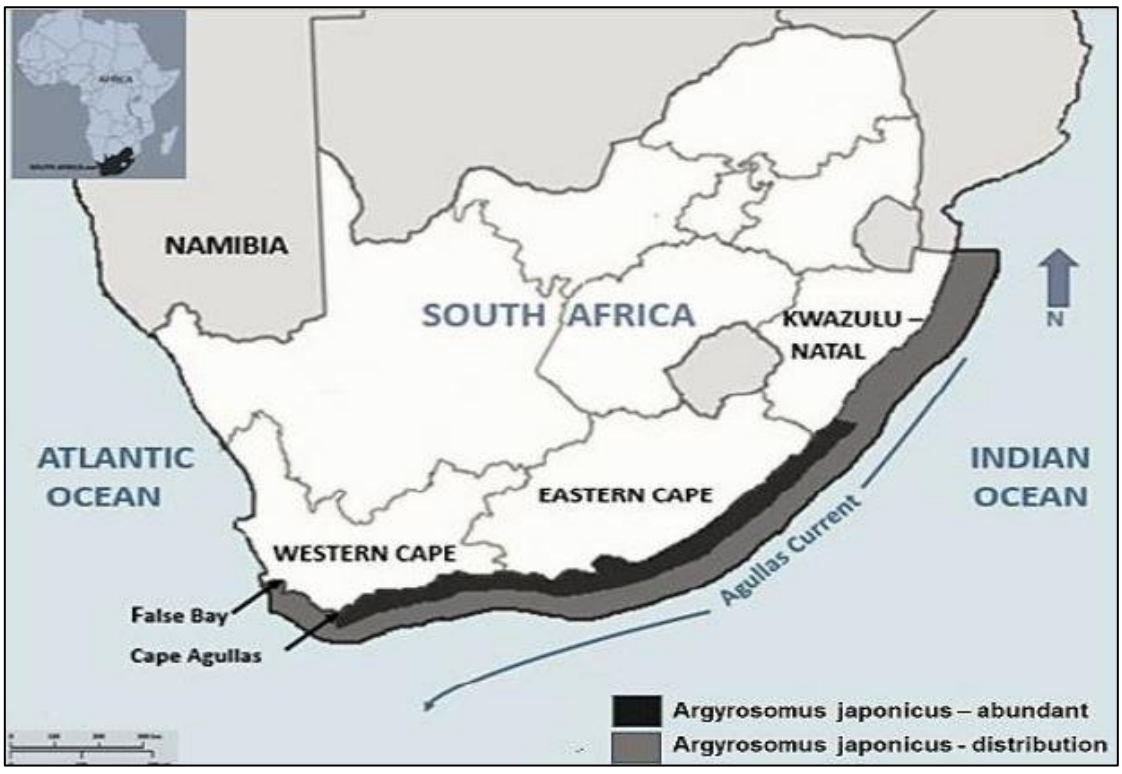



**Figure 1.2. Areas of distribution and abundance of dusky kob in South African waters. The figure was adapted and modified from Mirimin et al., (2015) in Jenkins (2018).**

## 1.2) Aquaculture of the Finfish, Dusky Kob

### *1.2.1) History and Development of the Industry*

The marine ecosystems along the South Africa coastline, support a well-established fishery sector that is responsible for exploiting a variety of indigenous living resources; however, this resource is continuously under threat from poaching, pollution, estuarine habitat degradation, inappropriate developments and poor management (Branch and Clark, 2006; Mead et al. 2013). As such, seafood production, *via* mariculture has been characterised as an emergent industry in South Africa (Bolton et al., 2013) and is developing at a faster rate than the freshwater aquaculture sector, with particular emphasis on *Mytilus galloprovincialis* and *Choromytilus meridionalis* (mussels), *Crassostrea gigas* (oysters), *Haliotis midae* (abalone), seaweeds and *Macrobrachium rosenbergii* (prawns) [Department of Agriculture, Forestry and Fisheries (DAFF), 2012]. The significant Increase in the production of farmed fish over the last few decades has resulted in marine species becoming of great economic importance. Aquaculture species have been targeted by commercial, recreational and subsistence fisheries for decades (Childs and Fennessy, 2013; Hutchings and Lamberth, 2003), which has resulted in collapse of the natural populations and exploitation far beyond optimal levels. Marine living resources such as dusky kob are particularly vulnerable, as it is currently one of the most commercially, ecologically and culturally important aquaculture species in South Africa. The wild stocks of dusky kob have come under extreme pressure as a result of having to sustain both commercial and recreational fisheries for decades (Brouwer et al., 1997; Childs and Fennessy, 2013; Pradervand et al., 2007). The spawner biomass is an estimate used to determine the total weight of the fish in a stock that are old enough to spawn, with populations considered unsustainable if they have an estimated value of 20% or less than pristine levels (Griffiths et al., 1997; Otgaar et al., 2012). The spawner biomass of dusky kob was estimated to be well below the threshold with the estimated value falling between 1.0 and 4.5%. This is the result of fishing efforts being shifted towards estuarine nursery areas in response to the predictable distribution patterns of the species (Cowley et al., 2013; Dunlop and Mann, 2012; Griffiths et al., 2000) as well as the late onset of sexual maturity, which has resulted in the majority of the populations being removed by anglers before having the opportunity to spawn. This was only further aggravated by the mismanagement of the species, caused by the taxonomic confusion within *Argyrosomus* (Griffiths and Heemstra, 1995), which was only rectified in 2004 when regulations for recreational fishers were changed (Sauer et al., 2003). Prior to this dusky and silver kob were managed as a single species "*A. holopidotus*" (with legal size set at 40

cm. In an effort to better manage the species and allow for the restock of wild populations, dusky kob is listed as red on the South African Sustainable Seafood Initiative's (SASSI) Customer Seafood List if caught from linefish or trawl and is considered to be a threatened species.

On-shore production of dusky kob in South Africa has been fairly well-established in response to the declining wild stocks and ever-growing demand for seafood (Saker and Griffiths, 2000). Since the commencement of dusky kob production in South Africa, a number of research efforts have been initiated to gain a better understanding of the biological determinants such as growth, disease resistance and fecundity, which influence this finfish. These traits are three of the main production limiting factors caused by a lack of understanding, *e.g.* traits such as growth and fecundity are often investigated individually however recent studies have shown that egg production increases exponentially with size (Barneche and Andrew, 2018). Therefore, understanding the biological role that genes play in the influence of commercially important traits of the species, the information can be utilised for stock assessment and improved management strategies to develop a sustainable fish-farming industry (Bernatzeder et al., 2007; Collett et al., 2007; Daniel et al., 2004; Kaiser et al., 2011; Musson and Kaiser, 2014). Fortunately, studies have shown that dusky kob compares well to *Sciaenops ocellatus (*red drum), an established Sciaenid species cultured in China (Hong and Zhang, 2003) and in the United States (Lee and Ostrowski, 2001). Thus, information obtained over the years through the establishment of this species can assist in the accelerated production of dusky kob. This comparison between the individuals assisted in accessing the candidacy of dusky kob for aquaculture with criteria such as, a fast initial growth rate, good feed conversion ratio, tolerance to low salinity and low oxygen levels, high crowding densities and disease resistance being assessed (Collett et al., 2008; Fielder and Heasman, 2011; Fitzgibbon et al., 2007, 2011; Griffiths et al., 1996; Whitfield, 1998).

The South African marine finfish industry, which is currently centred around dusky kob and yellowtail (*Seriola lalandii*), is still underdeveloped and will take a number of years before reaching its full potential. In 2011, a significant investment was made to establish the aquaculture of marine finfish within South Africa, (*i.e.* 42% of the total aquaculture investment; DAFF, 2012) and although this is still a developing field, the Food and Agriculture Organization of the United Nations (FAO) showed that while marine catch has been plateauing, marine production within South Africa has been experiencing a steady growth (6% per year) (DAFF, 2016). The total aquaculture production during 2015 and 2016 in South Africa was 7430 tons and 7819 tons, respectively. The total value of aquaculture

production during these years was USD 52 million for 2015 and USD 46 million for 2016. During 2016 the marine aquaculture production in South Africa amounted to 6160 tons with a value of USD 42 million. Of this, aquatic plants and mussels were the top contributors totalling 4300 tons. Other contributors to the total marine aquaculture production during 2016 consisted of abalone (*Haliotis midae,* 1500 t), oysters (*Crassostrea gigas*, 280 t) and finfish (various species, 80 t). Unfortunately, in 2017 there was a 23% decline in the production by aquaculture, with South Africa only producing a total of 6047 tons. This is likely the result of a number of factors which include but are not limited to the labour-intensive nature of Recirculation Aquaculture Systems (RAS), transport costs to major cities such as Cape Town, Durban and Johannesburg, the increasing cost of imported feed, and the increasing cost of electricity (Viljoen, 2019).

### *1.2.2) Current Perspectives and Practices*

Around the world there are multiple methods used to culture fish, cages, RAS and ponds. Of these the most recent method is cage culture which is used to culture fish in natural or artificial water bodies. With one of the main advantages being that they do not require land-ownership and can be moved around to the most suitable area for the target species (Viljoen, 2019). Another advantage is that cages allow for the fish to be kept in groups which facilitates the size-sorting and can prevent unwanted reproduction. However, there are a number of disadvantages to this method that should be considered. In cages the fish are unable to access the bedrock from which they can feed or seek refuge and there is also the ever-present risk of losing the entire group should the fish escape from the cage. It is also known that in certain waterbodies cages suffer from fouling of the mesh, thus preventing the free-flow of water through the cage, resulting in poor water quality at times. Some waterbodies could potentially be polluted by the accumulation of uneaten feed and fish waste gathered below the cage (Pearson and Black, 2000; Viljoen, 2019). Although there are various advantages to this method, the main reason for their exclusion from the South African aquaculture sector is the turbulent seas. The coastline of South Africa lacks protective locations such bays or deep lagoons meaning the cages would be exposed to harsh conditions which would result in the loss or damage of cages. Therefore, production practices in South Africa rely on the use of ponds or RAS.

The RAS systems are used worldwide for the commercial production of aquaculture species. These systems can be divided into categories based on their complexity and water use strategies. Systems can vary from flow-through, to partial flow-through, to a complete water recirculation system. Most systems incorporate the use of a water treatment plant that

recirculates and cleans the water to maintain a high level of water quality as sustaining a high density of aquatic animals in a confined space requires the exceptional maintenance of water quality, oxygen content, ammonia control and nitrate dilution (Rurangwa et al., 2011). This can be achieved using biological, mechanical and even sometimes chemical methods. In order to try and maximise profitability high value species are generally farmed using these systems to try and counteract the production cost. However, the high risk, complexity and energy consumption of RAS systems have led to other systems being implemented such as pond culture. These earthen ponds are used to produce fish and other aquatic organisms particularly in developing countries as it is often used for polyculture, the culture of more than one species. Although these ponds are cost-effective the design principles are essential, as large volumes of free-flowing water are required with a gravity supply being best. The fertilisation of these ponds has also been proven to increase the natural productivity of the water as the increased nutrient load supports the increased growth of algae which can be of benefit to species such as *Oreochromis aureus* (tilapia) which feed on the algae (Stoneham et al., 2018). This fertilisation can be achieved by the administration of animal manure into the water. The low running costs and simplicity make pond farms a very attractive alternative to current RAS.

Open reproduction systems utilise undomesticated broodstock (wild) to produce seed animals for culture; these cultured animals are not kept for breeding purposes as the system relies entirely on the use of wild individuals. To induce reproduction in fish, aquaculture farms generally rely on one of two methods. The first method is to provide an environment similar to that in which spawning naturally occurs. This is achieved by simulating the species preferred environment through the manipulation of photoperiod in the hatchery and an increase of the water temperature among various other things. The second method utilises one or more naturally occurring reproductive hormones, which is injected into the fish. However, this method is only effective in fish that are already in breeding condition, requiring these fish to have mature eggs where the germinal vesicle has already migrated. These two methods are often used sequentially: the first being used to manipulate maturation, while the second is used to induce ovulation (Griffiths, 1996). However, given the extreme sensitivity of adult fish when coerced into an artificial breeding and the high costs involved in the maintenance of a large number of fish, future expansion of this industry is likely to use a closed reproduction system that utilises cultured fish with favourable production characteristics to replace wild broodstock. Therefore, the use of a selective breeding programme would assist in increased production and more efficient resource utilisation. For

this specific reason a first-stage selective breeding programme has been implemented. During the initial phase of domestication, it is essential to maintain the genetic diversity within the breeding population. This can be achieved by maximizing the founder population and avoiding excessive inbreeding thus maximising the response to selection. It is possible to reduce chances of inbreeding in the initial generations of selection by establishing a large breeding population which has a low level of average relatedness but high levels of allelic variability (Hayes et al. 2006; Sekino et al. 2004). The stock performance of breeding programmes in aquaculture can be optimised as other economically important traits can be developed through artificial selection, such as traits relating to growth and disease resistance (Lillehammer et al., 2011). The ability to maintain a high quality broodstock population has greatly contributed to the improvement of biological efficiencies in many aquaculture species, including finfish species such as carp (*Cyprinus carpio L)* (Spasić et al. 2010; Dong et al. 2015) and Atlantic salmon (*Salmo salar L)* (Gjedrem et al., 1991; Skaalav et al., 2004).

Current reproduction practices of dusky kob rely on the mass spawning of broodstock (*i.e.* each male reproducing with many females and each female reproducing with many males in a single tank). The broodstock are housed underneath photoperiod control to ensure the continuous production of eggs throughout the year. Prior to the commencement of spawning, the female broodstock are sedated and cannulated for the collection of oocytes using a catheter. Generally, oocytes with a diameter of 0.5mm or more are considered optimal, increasing the chance of successful spawning (Jenkins, 2018). Following this test, the male and female broodstock individuals are hormonally induced and the water temperature raised (>22ºC) to initiate the spawning process. With a spawning female being able to produce anything between 2 million and 12 million eggs at a time. Upon completion of spawning, all the viable (floating) fertilised eggs are collected and placed into incubation tanks for hatching, with hatching taking place at approximately 24-30 hours after spawning. During the first 48 hours the larvae feed on the yellow yolk sac, after which they are transferred to a larval rearing system which consists of circular tanks that are on a recirculation aquaculture system. These recirculating systems filter and clean the water for recycling back through to the fish, recovering waste products that can supply nutrients for vegetable production in an aquaponics system reducing the amount of water required. After this period live feeds are introduced beginning with *Branchionus spp.* (rotifers), followed by *Artemia* (brine shrimp) until the larvae are fully weaned and then transferred to the nursing tanks (juvenile stage). Although there are various options, the best choice for first live larval

food in hatcheries is rotifers. This is due to a number of reasons, such as the organism's small size (130–320l), calorie value, relatively low mortality, slow swimming velocity and its ability to rise in high-density conditions (Lubzens et al., 2001; Yoshimatsu et al., 2014). Even at high densities, the rotifers reproduce rapidly, building up large quantities of live food in a very short period of time. It has also been suggested that marine fish larvae only have a partially developed digestive tract after hatching. These larvae therefore depend strongly on exogenous enzymes, provided by the live food which they consume, for digestion of their prey, meaning that the rotifers or brine shrimp are partly digested by their own enzymes, which are released as they reach the gut of the larvae (Kolkovski et al., 1993; Munilla-Moran et al., 1990; Walford and Lam, 1993). Between 30 to 35 days, the weaned larvae metamorphosize to become fully developed. Once reaching an average size of 1.6g, these fingerlings are moved to the grow-out section, where they are fed according to specific feeding charts, temperature calculations and growth rate indicators (Griffiths, 1996).

At approximately several months of age, the juveniles of similar age are pooled and divided into two or more independent size grades, depending on their body weight and length. With the slower-growing juveniles often being culled before and/or after grading, or alternatively when tank space is limited. These practices are necessary in order to maintain standard growth rates throughout harvest (which can range from 400g to 3kg), and subsequently minimising detrimental behavioural effects such as aggression (Jenkins, 2018). Aggressive behaviour in the aquaculture of dusky kob is a common occurrence, often resulting in cannibalism and can occur in as little as 18 days post hatching (O'Sullivan and Ryan, 2001). Cannibalism has also been reported for other aquaculture species, including *Lates calcarifer* (barramundi) (Loughnan et al., 2013), *Epinephelus lanceolatus* (giant grouper) (Hseu et al., 2007), *Clarias gariepinus* (sharp tooth catfish) (Baras et al., 2001), *Paralichthys olivaceus* Japanese flounder (Dou et al., 2004) and *Sciaenops ocellatus* (red drum) (Liao and Chang, 2002). Although aggression can arise regardless of the situation, the degree of cannibalism has been shown to be more pronounced in groups where offspring from multiple families are raised in a communal environment (Baras and Jobling, 2002; Liu et al., 2017). Factors such as inadequate food source, low feeding frequency, crowding density and light intensity have also been shown to increase the level of aggression (Collett et al., 2008; Fessehaye et al., 2006; Hecht and Pienaar, 1993; Kestemont et al., 2003; Timmer and Magellan, 2011; Qin et al., 2004).

## 1.3) Molecular Markers

Genetic variation is necessary in an ever-changing environment, where transformation and adaption are essential for the survival of the species (Bailey et al., 2010). Genetic variation arises between individuals when evolutionary forces such as mutation, selection and genetic drift causes differentiation at the level of population, and in extreme cases the creation of new species. Molecular markers are genetic polymorphisms that arise through mutation and is subject to demographic and/or functional effects population effects, and can be used to deduce population dynamics, familial relationships, or for studying the genetic mechanisms that underlie phenotypic traits. These markers are classified into two types, type I and type II. Type I markers are associated with genes of known function, while type II markers are associated with anonymous genomic regions (O'Brien, 1991). Type II markers can be converted to type I markers once a marker has been associated with genes of known function. The significance of type I markers is becoming extremely important for aquaculture genetics (Chauhan and Rajiv, 2010). During the early stages of aquaculture, all the molecular work was performed using allozymes (enzyme products of genes, type I marker) and despite the known limitations of allozymes it did have a profound effect on the management and research of fisheries, as this research demonstrated the usefulness of genetic markers in stock identification that has a direct functional link (Grant et al., 1999; May, 2003). These markers, do however, have a limited power in detecting genetic variability, and require large amounts of tissue from organs (*i.e.* liver and heart) for their assay, resulting in the death of the animal.

The use of allozymes were followed by the development of Type II DNA markers, which include amplified fragment length polymorphism (AFLP), random amplified polymorphic DNA (RAPD), and minisatellites (Carvalho and Pitcher, 1995; Clifford et al., 1998; Vos et al., 1995). These simple methods are rapid, cheap, and only require a small amount of DNA, with no prior knowledge regarding the genetic make-up of the organism being necessary (Hadrys et al., 1992). The weakness is that these are all dominant markers, making them difficult to analyse (Ignal and Ilan, 2002; Liu and Cordes, 2004). One of the main criticisms of minisatellites and AFLPs is that the allele frequencies for a given locus cannot be determined as multiple loci are assayed simultaneously (Magoulas et al., 1998). As a result of these limitations, molecular genetic studies performed on aquaculture species have expanded to include the use mitochondrial DNA (mtDNA) markers, microsatellites, and more recently SNPs. Markers using mtDNA, represent a single locus, is a very popular marker which has been prevalent in genetic studies looking at phylogeny and population structure

in fish for more than a decade (Billington, 2003) at inter-specific level, but it is still not the most effective for assessing genetic variability within commercial stocks (Hurst and Jiggins, 2005). This is because mtDNA was strictly a marker for historical processes in females, therefore should male and female history differ in a species (such as the interdiction of wild broodstock), then this marker would not reflect the history of the species as a whole, but only that of the maternal lineage.

The development of genetic markers has transformed molecular studies with microsatellites and single nucleotide polymorphisms (SNPs) playing a fundamental role in this transformation. Microsatellites are co-dominant markers that consist of short tandem repeats which are located mostly within the non-coding regions of DNA. Each of the repeat motifs generally consist of two to four base pairs, with the number of repeat regions varying between individuals and populations (Morin et al., 2004). On the other hand, SNPs are caused by, a base pair substitution resulting in two alleles differing at a particular position on a locus, by a single base pair, in otherwise identical sequences. Each of these markers have slightly different advantages and disadvantages that make them ideal for studying populations, however, microsatellites have been the marker of choice for aquaculture development as they are highly polymorphic, simple and cheap to score and exhibit cross-species utility in closely related species (Dawson et al., 2000; Dawson et al., 2005). Recently though, SNPs have emerged as a viable marker for use in non-model species as advances in technology have led to reduction in the time and cost involved in the location and genotyping of these markers (Hansson et al., 2005; Syvänen, 2005). As a result of these advances there has been an increased use of SNPs despite their predominately biallelic nature, which means that in comparison to the highly polymorphic microsatellites, SNPs provide relatively less information per locus. Thus, making linkage between markers more difficult to detect as SNPs are unable to identify as many informative meiosis as would be possible with microsatellites. Therefore, a larger number of evenly spaced makers can be utilised to cover a higher proportion of the genome in order to compensate for this reduction (Xing et al. 2005).

Although microsatellites are highly polymorphic in comparison to SNPs, they are known to be relatively prone to genotyping errors therefore generating potentially a lower ‘quality’ of data. The quality of data is only further affected by the use of semi-automated microsatellite-based methods of genotyping and allele-calling, which can introduce human-based errors. While modern SNP genotyping platforms are almost fully automated and error rates tend to be much lower resulting in data of a higher quality (Heaton et al., 2002; Lindblad-Toh et al.,

2000; Wang et al., 1998). This is an important factor to consider when selecting markers as these genotyping errors can have a large impact on parentage inference and population structure analyses (Bonin et al., 2004; Slate et al., 2008). Of the many benefits involved in utilising SNP markers, reproducibility is one of the most important. This reproducibility is only possible due to universal nucleotide calls and the flexibility of SNP detection protocols, which is not possible for microsatellites, which rely on the migration of microsatellite fragments during electrophoresis for comparison to known standards. This can be a very unreliable method for size-based allele determination as the migration rate can differ between electrophoresis methods, making it extremely difficult and time consuming for laboratories to compare the genotype data (Kim et al., 2008).

## 1.4) SNP development strategies and genotypic technologies

Approaches for the detection and development of SNP markers relies on the comparison of sequence data from multiple individuals and detecting sequence polymorphism in multiple alignments. Historically this was done by generating BAC - (bacterial artificial chromosome) (random genomic DNA fragments) or EST libraries (from cDNA) (Chauhan and Rajiv, 2010). However, there has been significant advances made in high throughput sequencing technology (HTS) over the last decade, which has resulted in the cost of sequencing being reduced while simultaneously improving the usability and accuracy of the sequence data. Some of the most significant innovations have been made in whole genome studies, which use a combination of *de novo* assembly, re-sequencing, and bioinformatic approaches to identify a large number of SNPs for many organisms with complex genomes (Bertioli et al., 2016; Lee et al. 2015; Yang et al. 2012). Along with this mass sequence data being produced there has also been significant development in SNP genotyping technology, with recent advances including PCR-based fluorescently-labelled high-throughput methods, high-resolution melting (HRM) curve analysis, TaqMan® and KASP™ assay (Martino et al., 2010), fixed array systems such as Illumina Infinium (Mason et al., 2017), Affymetrix Axiom (Allen et al., 2017), and high throughput sequencing (HTS) enabled approaches such as restriction-enzyme-based genotyping by sequencing (GBS) (Thomson, 2014). One of the most popular approaches that is currently used for the detection of SNPs is the use of HTS technologies in combination with genotyping arrays (Ganal et al., 2014). However, a requirement for commercial SNP-genotyping platforms is information regarding the target organism, resulting in an increased cost and duration required for sequencing, making this an ineffective approach for non-model organisms (Ekblom and Galindo, 2010).

Although identification through HTS in comparison to conventional SNP detection methods, does reduce the duration and simplify the scoring of data, there is still a significant amount of research required for the development of new markers in non-model organisms (Chung et al., 2017). Methods such as whole genome resequencing (WGR) and reduced-representation sequencing (RRS) are constantly being improved to try and overcome limitations. These approaches have been successfully used in several species to identify multiple loci, genome wide, which has been essential to understanding and answering a variety of molecular ecology questions (Hohenlohe et al., 2010; Foote et al., 2016; Lamichhaney et al., 2017). Whole-genome sequencing can be classified in two categories, *de novo* whole-genome sequencing (WGS); and whole genome resequencing (WGR). The aim of WGS is to determine the complete DNA sequence of an organism's genome for the first time, which can be challenging depending on the level of completeness which is desired, the complexity and size of the genome, computing resources and bioinformatics experience. However, the completeness and the accuracy of the genome assembly will determine whether the draft genome is suitable for further analyses and applications (Fuentes-Pardo and Ruzzante, 2017). Despite the usefulness of this approach in some applications, the general consensus is that incomplete draft genomes can create more problems than solutions, particularly for accurate SNP calling where high coverage and accurate alignments are essential (Li and Wren, 2014). Unlike WGS, the aim of WGR is to rather compare the genomic variability among individuals or populations than sequence the entire genome. However, for read mapping and variant identification this approach does require the availability of a reference genome. This is why many researchers have implemented the use of WGR using the genome sequences of a closely related species (Dennenmoser et al., 2017; Lamichhaney et al., 2012). Differences in the genomic organisation can occur (*e.g.* copy number variation, structural variants) even between closely related species, thus restricting this approach to conserved regions between the species (Ekblom and Wolf, 2014). There are three main techniques which are used for reduced-representation sequencing namely Restriction site Associated DNA sequencing (RAD-seq; Andrews et al., 2016), Sequencing of cDNA obtained from mRNA (RNAseq; Ozsolak and Milos, 2011) and Whole-exome sequencing (WES; Warr et al., 2015).

All these techniques have their strengths and weaknesses which make them better suited for specific applications. For RAD-Seq methods (*e.g.* traditional RAD, ddRAD, ezRAD) the marker density is limited by the selection of the restriction enzyme, which can be either be a frequent or rare cutter, as this method evaluates the genetic variation that is present

around restriction cut sites. However, this does make it a flexible and customisable method for examining thousands of low-density SNPs, genome wide in multiple individuals and populations. Although with RAD-seq, the marker density and levels of linkage disequilibrium (LD) are important considerations (Andrews et al., 2016). The RNA-seq technique is a transcriptome sequencing method which is not restricted by the target size; however, this technique is limited with regard to distinguishing nonsense mutations and in the discovery of genomic lesions that affect splice sites (Bowen et al., 2011; Leshchiner et al., 2012; Obholzer et al., 2012) as this technique focuses on the genetic variants that are being transcribed in specific regions of the genome at the time of sampling. Therefore, this approach is mostly used as a cost-effective approach for gene expression quantification and for the comparison of variants within genes being transcribed in a particular tissue or at a specific time (Ozsolak and Milos, 2010). Thus, causing the genome to have regions where there is little to no coverage as a result of gene expression at the time of sampling. This does not only affect the coverage, but introduces an ascertainment bias where highly expressed genes are given a greater chance of detection during sequencing, thus skewing downstream gene ontology (Costa et al., 2012; Ozsolak and Milos, 2010).

Thus, targeted sequencing of the genome using high throughput sequencing has become a powerful method for identifying variants (Albert et al., 2007; Hodges et al., 2007; Hodges et al., 2009; Okou, 2007). Exome sequencing also known as whole-exome sequencing (WES) is the most widely used targeted sequencing method. For the identification of causal variants this method has quickly become the strategy of choice, as it is rapid and cost-effective. This is due to the ability of this method to only sequence the coding regions of the genome, therefore focusing on the genes that are most likely to have a causative effect on the phenotype (Belkadi et al., 2016; Warr et al., 2015). Normally obtaining this information would require the genotyping of thousands of gene-targeted-loci across the genome. However, with the coding gene sequences (the exome) within the typical eukaryotic genome, only comprising of 2% and the advances made in the development of techniques for the isolation exome DNA, thousands of informative gene markers can be simply and cost-effectively located and identified within the genome (Luikart, 2003). WES is a powerful tool but it has been precluded in studies as a result of its non-uniform exon coverage across the genome. However, in recent years there has been a significant increase in the utilisation of this strategy with the release of commercial exon capture kits, which has enabled researchers to target exons from non-human organisms for resequencing. These kits are found to be easily adaptable to high-throughput workflows and do not require any sort of investment in

array-processing equipment, making them particularly useful and important (Parla et al., 2011).

In humans, approximately 85% of known phenotypically associated mutations can be found within the coding region or splice sites of protein-coding genes (Ng et al., 2010). Whilst this number is most likely the bias of studies which have only focused on protein-coding genes, exome sequencing has still become the standard tool for the identification of variants in humans (Bilguvar et al., 2010; Raffan, et al. 2011; Worthey et al., 2011). While exome capture was initially performed using microarrays (Albert et al., 2007; Hodges et al., 2007), newer methods, such as Agilent's SureSelect and Nimblegen's SeqCap EZ systems rely on solution-based capture (Bainbridge et al., 2010; Gnirke et al., 2009). Until recently, exon capture had only been tested almost entirely in model species (Raca et al. 2010; Wang et al., 2010), usually performed by baiting a single chromosomal or the entire exome region using the available genome sequences of the target organism. Probe design for exome-wide capture requires knowledge of thousands of exon sequences, as such, studies have not yet tested the potential of exon capture to a wider variety of organisms. This information is not available for many eukaryotic species as only a small portion of these species have had their genomes fully sequenced. Although there would still be tens of thousands of vertebrate species without genome sequences or any genomic resources, even if researchers were able to eventually sequence a large number of eukaryotic species. Hence the need to investigate the potential of cross-species exome capture. As such, studies have been performed using whole-exome sequencing in combination with solution-based exon-capture kits, which have been designed specifically for model organisms such as, cattle and humans (Cosart et al., 2011; Vallender, 2011). Using these kits in closely related species, the researchers were able to achieve a high number of quality on-target reads as well as providing a reliable set of SNPs. This allowed for the accurate determination of critical genomic intervals while reducing the number of candidate mutations requiring evaluation. Due to the high success of these kits in closely related species there is a large amount of potential in the utilisation of model organisms, capture kits in non-model organisms, as the functional elements tend to be highly conserved despite millions of years of divergence. The inclusion of the exome capture kits in WES strategies will enhance the ability of this method to identify genetic markers, with or without the availability of a reference genome thus aiding in the rapid development of genomic resources (Warr et al. 2015).

## 1.5) Application of SNPs in aquaculture

### *1.5.1) Individual identification, Pedigree inference and Population Assessments*

Fish are known to have some of the most complex mating systems within the animal kingdom. Meaning that effective methods are required for the traceability of these animals, methods which can also be utilised not only for research purposes but for controlling the trade and management of marine animals/products. Most marine species are accurately traced by inferring parentage, kinship and population structure, which are most effectively estimated using molecular markers such as SNPs and microsatellites (Liu and Cordes, 2004). Although there has been an exponential growth in the use of SNPs over the last decade for such analyses, (Guichoux et al., 2011) these markers are not yet widely used for parentage assignment. This is largely due to the fact that there are still many questions regarding ascertainment (SNP discovery and selection) methods (Aitken et al., 2004; Rosenblum and Novembre, 2007; Smith et al., 2007) and the large discrepancy observed between the statistical power of SNPs and microsatellites. Some studies have tried to address these questions in terms kinship (Krawczak, 1999), individual identification (Chakraborty et al., 1999) and parentage inference (Anderson and Garza, 2006), with a study performed by Glaublitz et al., (2003) showing that a single microsatellite appears to have the same resolving power of ~6 SNPs making SNP markers extremely costly for this application. This issue was also addressed in terms of population structure by Kalinowski (2002), which showed that the statistical power of genetic markers for detecting differentiation as a result of genetic drift is not related to the number of loci but rather primarily to the total number of independent alleles. Therefore, this can be used to provide a rough estimation as to how many SNP loci are required to obtain the same statistical power as a given set of microsatellite loci. This was determined to have quite a wide range, with the effects of ascertainment bias, allele frequency and linkage still needing to be taken into consideration when determining the statistical power of the loci (Smith and Seeb, 2008). In general, the statistical power of a certain marker set varies depending on the purpose and application, thus the markers should be tested in advance to assure sufficient power for the application (Vignal et al., 2002).

Due to the aforementioned advantage of microsatellites, this marker has been frequently used in population genetics. However, this is rapidly changing as an evaluation of these two markers for inferences such as hybrid detection (Väli et al., 2010), inbreeding (Santure et al., 2010), and parentage or kinship analyses (Hauser et al., 2011; Ross et al., 2014) has shown SNPs to be far superior to that of microsatellites. Although, when solely looking at a

per-locus basis, microsatellites do retain advantages over SNPs, advantages which include a lower ascertainment bias, higher allelic richness, and higher statistical power (Guichoux et al., 2011; Haasl and Payseur, 2010; Payseur and Jing, 2011; Sun et al., 2009). With studies having shown microsatellites to be better or relatively similar to SNPs in regards to population structure inference (Ciani et al., 2013; Granevitze et al., 2014; Livingstone et al., 2010; Ross et al., 2014). However, these studies only evaluated a modest number of markers, and it has been stated that the use of a large number of SNP loci, which can be obtained using high-throughput sequencing is likely to overcome many of the markers' weaknesses. A study by Haasl and Payseur (2010) evaluated the utility of microsatellites and SNPs for addressing various population genetics questions and what they determined is that SNPs were generally found to have a greater power in detecting population structure in comparison to microsatellites, as only a few SNP loci were needed to detect structure between populations with moderate divergence times. Although, this study did show that as divergence times decreased, significantly more SNP loci were required in comparison to microsatellites (Haasl and Payseur, 2010). The stability of SNPs is however, considered to be a major advantage in evolutionary, population biology and pedigree studies, as theses markers are not limited to the non-coding regions of the genome and are therefore likely to be subjected to evolutionary selective forces (Stoneking, 2001). However, microsatellites may still be more appropriate for studies on short temporal or spatial scales, where the applications require both cross-species range and good resolution, (Buschiazzo and Gemmell, 2010; Dawson et al., 2013; Seeb et al., 2011) in taxa that are highly clonal or slowly evolving (Stolle et al., 2013).

However, it is important to understand that the majority of the studies used for these comparisons, focused on breed/stock identification of well-studied systems (*e.g.* salmon), meaning that the set of loci used for these studies were developed prior to the genomic era. Therefore, studies need to utilise high-throughput methods to develop both SNP and microsatellite markers in order to accurately determine the marker choice for future population genetics studies of non-model species. Overall, molecular genetic markers such as microsatellites and SNPs have a wide range of potential applications in the long-term management of farmed and wild populations, and with the continuous improvements being made to high throughput sequencing methods, microsatellite data are already available as a by-product of methods obtaining genome wide SNPs. Thus, the simultaneous use of these markers will assist in the greatest accuracy and resolution.

*1.5.2) Loci Associated with Complex Traits in Aquaculture and Marker-Assisted Selection*

Aquaculture genetics programmes became prevalent in the 1900s due to the better understanding and knowledge obtained regarding breeding and inheritance. In the 1960s, the first selective breeding programmes for genetic enhancement of aquatic animals were implemented (Gjedrem and Baranski, 2009). At the same time there was a shift in animal breeding from classical quantitative genetics to more molecular approaches with an increase in molecular genetics research, particularly during the 1990's, with the focus of animal breeding shifting in 1990 from quantitative to molecular genetics (Misztal, 2006). Yet despite the rapid growth in aquaculture production and the advances made in genetic tools, the vast majority of aquaculture facilities across the world still maintain and propagate stock without the assistance of advanced selective breeding programmes (Gjedrem et al., 2012; Janssen et al., 2016). Meaning that genetic markers are not being utilised to their full potential, with genetic tools only being applied for general stock assessments (of diversity) and to the pedigree reconstruction of many aquaculture species (Chavanne et al., 2016). Considerable scope remains for the use of molecular markers as a diagnostic tool in the identification of genetically superior animals even before the phenotypic trait is expressed (*i.e.* marker assisted selection). This however does require that the marker is associated with a phenotype of a trait of importance. Classically this was done *via* the construction of genetics linkage maps and investigating the co-segregation of phenotypes with marker loci in pedigrees (Slate, 2008). Historically, most linkage maps suitable for crude QTL analysis was based on microsatellite genotypes, however most high-density maps are now constructed using SNP markers. SNP markers have also become the marker of choice in other genotype-phenotype correlation strategies, including candidate gene- and whole genome association studies (also referred to as LD mapping), with each method having its own advantages and disadvantages.

Genome-wide scanning usually proceeds without any assumptions regarding the specific functional features of molecular makers, which may be of importance to the traits of interest, making this a resource intensive approach. This intensive approach is much like QTL mapping in its ability to identify multiple regions within the genome that contain potential QTLs. The problem with this is that these regions are typically very large, containing thousands of putative genes, with a large number of candidate genes still remaining following fine scale mapping, making the investigation into all the genes unfeasible (Wayne and McIntyre, 2000). Therefore, due to the excessive data and cost involved with this approach, candidate gene approaches have been employed, with studies showing this to

be extremely powerful for studying the genetic architecture of complex traits (Hebert et al., 2013; Pacitti et al., 2013; Tao and Boulding, 2003). This method is far more effective and economical for direct gene discovery, as this approach is able to narrow down the large number candidate genes to only a few genes that have been identified in literature as having an effect on the particular trait related to the biological function of interest, thus increasing the likelihood of identifying QTL which are directly linked to the trait of interest. Thus, allowing researchers to study the genetic architecture of complex traits, by focusing on a select number of gene regions where association with the trait of interest is typically located. This method is often employed in non-model organisms, which have limited genomic information or when the cost of QTL mapping is excessive as a result of the organism's large genome. Although, this approach is largely limited by its reliance on existing knowledge about the known or presumed biology of the phenotype under investigation, which is generally limited in a large number of organisms (Korte et al., 2013).

This provides a major challenge in relation to discovery of candidate genes in non-model species thus, making the candidate gene approach labour intensive. However, some studies have taken advantage of the available sequence information for related species to identify genetic variation in a non-model species (*e.g.* Aitken et al., 2004; Cosart et al., 2012; Hemmer-Hansen et al., 2011; Primmer et al., 2002), thereby increasing the number of potential target genes in species with limited genomic information. Although the candidate gene identified as having association to the trait of interest within the related species may not have the same association within the target organism (Aguirre-Hernández and Sargan, 2005). Nevertheless, this approach allows for the rapid development of markers, which can aid in the development of genomic resources in non-model organisms which previously would have been labour intensive and costly. Thus, despite the known limitations of this method, it is a powerful tool that will greatly assist researchers in understanding the biological role that genes play in the expression of economically important complex traits. Knowledge pertaining to the functional role of genes in the expression of desired traits, is essential for the improvement of future selection methods. The traditional selection of animals based solely on phenotypic quality, has shown to improve livestock populations, reducing production costs and improving the quality of the products however it has actually only assisted in the improvement of a very limited number of traits. Therefore, in order to meet the public demand and develop a sustainable industry, it is necessary to address the limitations that are associated with traditional selection approaches by utilising new technology, genetic markers, for the selection of genetically superior animals. Traditionally,

selective breeding of local livestock was aimed at improving the genetics to ensure survival and success in their surrounding environment, thus providing a food source to the local communities (van Marle-Köster and Visser, 2018). This selective breeding led to distinct breeds of livestock being formed as a result of characteristic phenotypes. These diverse phenotypes are controlled by equally diverse genetic elements, therefore providing the opportunity for the selection of animals with superior performance in specific desirable traits, such as growth rate, fertility, hardiness, product yield and quality (*e.g.* milk, meat, egg, etc.), and disease resistance. With countries where the economic environment supports high input agriculture or aquaculture, there has been a significant increase in the level of productivity from the selective improvement of livestock in simple production traits (Allaire and Gibson, 1992; Broderick, 2003). This was achieved by utilising the latest technological advances such as artificial insemination (AI), to maximise selection for genetic gain (Hunt et al., 1974; McMahon et al., 1985; Wilcox et al.,1984). However, most of the economically important production traits are found to be more complex than these simple traits and have a very large range of variation within the observed phenotype. This could result in the improvements of one trait by selective breeding causing the loss or decreased performance of other traits (Williams, 2005). Therefore, to overcome this limitation, marker assisted selection can be employed where knowledge obtained regarding the genes involved in the underlying expression of the traits can be used for future improvements. This is possible by using the most beneficial loci to detect and identify a number of complex traits within the livestock, prior to breeding allowing for the most beneficial allelic combinations being passed on to the next generation.

## 1.6) Study rationale, aims and objectives

### *1.6.1) Problem Statement*

Dusky kob is a large estuarine-dependent sciaenid finfish that is found to be abundant within South African waters. For decades, this species has been targeted by commercial, recreational and subsistence fisheries, leading to the subsequent collapse of the natural populations. With the poor management and unsustainable harvesting of fisheries, a shift towards aquaculture as a sustainable alternative supply to the ever-increasing market, has been initiated. For many years, the primary focus of the South African marine finfish industry has been the improvement of dusky kob's growth rate through the implementation of a selective breeding programme. However, there are currently no genomic resources available for dusky kob, resources which are necessary to understand the various genetic determinants influencing complex traits, particularly growth rate, as this will facilitate in the

effective utilisation of the species through selective breeding. The development and implementation of these genomics tools in the selection process of broodstock will potentially enhance commercial productivity as well as value for this species.

A powerful method used to develop these genomic resources is targeted sequencing. This method is often used for the identification of variants associated with traits of interest using next-generation sequencing. This method is however limited by cost considerations hence the recent development of exon capture kits. These kits allow for studies to rapidly and cost-effectively focus on the coding regions of the genome where variation with causative effects on the phenotype are likely to occur. Although there has been a significant increase in the use of these capture kits, their ability to sequence a non-model organism still remains unknown. Therefore, making it an untapped resource which could assist in the rapid development of genomic resources for organisms where no genomic resources are available. These resources, however, cannot be created without answers to key knowledge gaps pertaining to the transferability of these kits. Theoretically the use of a model organism's capture kit in the exon capture of a non-model organism should be successful, as the functional elements of the genome tend to be highly conserved. However, there are a number of questions regarding this strategy that remain unanswered thus impeding the improvement of dusky kob's selective breeding programme.

### *1.6.2) Aims and Objectives*

This study aimed to describe the development of SNP markers to assess the genetic variation associated with growth rate for the species dusky kob (*Argyrosomus japonicus*), using an optimised whole-exome sequencing protocol. The transferability of the zebrafish's solution-based exon-capture kit to the non-model organism dusky kob was assessed using sixteen individuals (in Chapter 2). The modified protocol used for the sequencing of the exomes was be discussed in detail to enable its replication in future studies. Additionally, the quality of the sequenced reads and the generated *de novo* assemblies was assessed, as well as their similarity to the model organism. Concluding this section with an evaluation of the capture kits ability to allow for the location and identification of variants within the genome. The identified variants were then be analysed (in Chapter 3) to identify genetic variation occurring within 15 candidate growth genes of two cohorts, with one of the cohorts consisting of 8 phenotypically characterised large individuals and the second cohort consisting of 8 individuals characterised as small. The identified genetic variants were then analysed in a case-control study to determine whether the loci are associated with the economically important trait, growth, in a larger cohort consisting of individuals from various

families generated a single spawning event. The obtained results (chapters 2 and 3) were then interpreted, and synthesised with the context of the broader body of knowledge and (in Chapter 4) discussed in terms of broad managerial recommendations related to the development of genetic resources and of genetic improvement strategies for the South African dusky kob were made and the development of genomic resources. As genotyping by sequencing becomes a more common method, this study therefore provides an accurate genetic resource which can aid in future genomics research and the acceleration of molecular breeding programmes.

## References

Aguirre-Hernández, J., Sargan, D.R., 2005. Evaluation of candidate genes in the absence of positional information: a poor bet on a blind dog. Journal of Heredity, 96, 475–484.

Aitken, N., Smith, S., Schwarz, C., Morin, P.A., 2004. Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach. Molecular Ecology, 13, 1423–1431.

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X, Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., Weinstock, G.M., Gibbs, R.A., 2007. Direct selection of human genomic loci by microarray hybridization. Nat Methods. 4, 903-905.

Allaire, F., Gibson, J., 1992. Genetic Value of Herd Life Adjusted for Milk Production. Journal of Dairy Science, 75(5), 1349-1356. https://doi.org/10.3168/jds.S0022-0302(92)77886-2

Allen, A., Winfield, M., Burridge, A., Downie, R., Benbow, H., Barker, G., Wilkinson, P., Coghill, J., Waterfall, C., Davassi, A., Scopes, G., Pirani, A., Webster, T., Brew, F., Bloor, C., Griffiths, S., Bentley, A., Alda, M., Jack, P., Phillips, A., Edwards, K., 2016. Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). Plant Biotechnology Journal, 15(3), 390-401. https://doi.org/10.1111/pbi.12635

Anderson, E.C., Garza, J.C., 2006. The power of single-nucleotide polymorphisms for large-scale parentage inference. Genetics, 172,2567–2582.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., Hohenlohe, P. A., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics, 17, 81–92.

Bailey, J., Genung, M., O'Reilly-Wapstra, J., Potts, B., Rowntree, J., Schweitzer, J., Whitham, T. 2011. New frontiers in community and ecosystem genetics for theory, conservation, and management. New Phytologist, 193(1), 24-26. https://doi.org/10.1111/j.1469-8137.2011.03973.x

Bainbridge, M.N, Wang, M, Burgess, D.L., Kovar, C., Rodesch, M.J., D'Ascenzo, M., Kitzman, J., Wu, Y.Q., Newsham, I., Richmond, T.A., Jeddeloh, J.A., Muzny, D., Albert, T.J., Gibbs, R.A., 2010. Whole exome capture in solution with 3 Gbp of data. Genome Biol. 11 (6): 62.

Baras, E., 2001. Size heterogeneity prevails over kinship in shaping cannibalism among larvae of sharptooth catfish *Clarias gariepinus*. Aquat. Living Resour. 14, 251–256. https://doi.org/10.1016/S0990-7440(01)01118-4

Baras, E., Jobling, M., 2002. Dynamics of intracohort cannibalism in cultured fish. Aquac. Res. 33, 461–479. https://doi.org/10.1046/j.1365-2109.2002.00732.x

Barneche, D., Allen, A., 2018. The energetics of fish growth and how it constrains food-web trophic structure. Ecology Letters, 21(6), 836-844. https://doi.org/10.1111/ele.12947

Beckley, L. E., 1993. Linefish larvae and the Agulhas Current. In Fish, Fishers and Fisheries. Res. Inst. S. Afr. 2: 57-63.

Beckley, L.E., 1995. The Agulhas Current ecosystem with particular reference to dispersal of fish larvae, in: Status and Future of Large Marine Ecosystems of the Indian Ocean. Blackwell Science, 74–91.

Belkadi, A., Pedergnana, V., Cobat, A., Itan, Y., Vincent, Q.B., Abhyankar, A., Shang, L., El Baghdadi, J., Bousfiha, A., Alcais, A., Boisson, B., Casanova, J.L., Abel, L., 2016. Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. Proc. Natl. Acad. Sci. 113, 6713–6718.

Bergamino, L., Dalu, T., Whitfield, A., Carassou, L., Richoux, N., 2014. Stable isotope evidence of food web connectivity by a top predatory fish (*Argyrosomus japonicus* Sciaenidae: Teleostei) in the Kowie Estuary, South Africa. African J. Mar. Sci. 36, 207–213. https://doi.org/10.2989/1814232X.2014.923782

Bernatzeder, A., Britz, P., 2007. Temperature preference of juvenile dusky kob *Argyrosomus* japonicus (Pisces: Sciaenidae). African J. Mar. Sci. 29, 539–543. https://doi.org/10.2989/AJMS.2007.29.3.19.349

Bertioli, D., Cannon, S., Froenicke, L., Huang, G., Farmer, A., Cannon, E., Liu, X., Gao, D., Clevenger, J., Dash, S., Ren, L., Moretzsohn, M., Shirasawa, K., Huang, W., Vidigal, B., Abernathy, B., Chu, Y., Niederhuth, C., Umale, P., Araújo, A., Kozik, A., Do Kim, K., Burow, M., Varshney, R., Wang, X., Zhang, X., Barkley, N., Guimarães, P., Isobe, S., Guo, B., Liao, B., Stalker, H., Schmitz, R., Scheffler, B., Leal-Bertioli, S., Xun, X., Jackson, S., Michelmore, R., Ozias-Akins, P., 2016. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. Nature Genetics, 48(4), 438-446. https://doi.org/10.1038/ng.3517

Bilguvar, K., Ozturk, A.K., Louvi, A., Kwan, K.Y., Choi, M., Tatli, B., Yalnizoglu, D., Tuysuz, B., Caglayan, A.O., Gokben, S., Kaymakcalan, H., Barak, T., Bakircioglu, M., Yasuno, K., Ho, W., Sanders, S., Zhu, Y., Yilmaz, S., Dincer, A., Johnson, M.H., Bronen, R.A., Kocer, N., Per, H., Mane, S., Pamir, M.N., Yalcinkaya, C., Kumandas, S., Topcu, M., Ozmen, M., Sestan, N., 2010. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature. 467 (7312), 207-210.

Billington, N., 2003. Mitochondrial DNA. E. M. Hallerman (Ed.), Population genetics: principles and applications for fisheries scientists. American Fisheries Society, Bethesda, Maryland: 59-100.

Blaber, S.J.M., 2000. Tropical estuarine fishes: ecology, exploitation and conservation. Fish and Aquatic Resources. 372. https://doi.org/10.1002/9780470694985

Broderick, G., 2003. Effects of Varying Dietary Protein and Energy Levels on the Production of Lactating Dairy Cows. Journal of Dairy Science, 86(4), 1370-1381. https://doi.org/10.3168/jds.S0022-0302(03)73721-7

Bolton, J., Davies-Coleman M., Coyne V., 2013. Innovative processes and products involving marine organisms in South Africa. African Journal of Marine Science, 35(3), 449-464.

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

Bowen, M.E., 2011. Efficient Mapping and Cloning of Mutations in Zebrafish by Low-Coverage Whole-Genome Sequencing. Genetics. 190(3), 1017–1024, 10.1534/genetics.111.136069.

Branch, G.M., Griffiths, C.L., Branch, M.L., Beckley, L.E., 2007. Two Oceans, A guide to the marine life of Southern Africa., Struik Publishers, ISBN 978-1-77007-633-4bolt

Brouwer, S.L., Mann, B.Q., Lamberth, S.J., Sauer, W.H.H., Erasmus, C., 1997. A survey of the South African shore-angling fishery. South African J. Mar. Sci. 18, 165–177. https://doi.org/10.2989/025776197784161126

Buschiazzo, E., Gemmell, N.J., 2010. Conservation of human microsatellites across 450 million years of evolution. Genome Biol. Evol. 2, 153–165.

Carvalho, G.R., Pitcher, T.J. (Eds.), 1995. Molecular Genetics in Fisheries. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-94-011-1218-5

Chakraborty, R., Stivers, D.N., Su, B., Zhong, Y., Budowle, B., 1999. The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. Electrophoresis, 20, 1682–1696.

Chao, L.N., 1986. Sciaenidae. In Fishes of the North-Eastern Atlantic and the Mediterranean, Whitehead, P. J. P., Bauchot, M.L., Hureau, J.E., Nielson, J., E. Tortonese. UNESCO: 865-874.

Chao, N.L., Frédou, F.L., Haimovici, M., Peres, M.B., Polidoro, B., Raseira, M., Subirá, R., Carpenter, K., 2015. A popular and potentially sustainable fishery resource under pressure-extinction risk and conservation of Brazilian Sciaenidae (Teleostei: Perciformes). Glob. Ecol. Conserv. 4, 117–126. https://doi.org/10.1016/j.gecco.2015.06.002

Chauhan, T., Kumar, R., 2010. Molecular Markers and Their Applications in Fisheries and Aquaculture. Advances in Bioscience and Biotechnology, 1(04), 281–291.

Chavanne, H., Janssen, K., Hofherr, J., Contini, F., Haffray, P., Komen, H., Nielsen, E., Bargelloni, L., 2016. A comprehensive survey on selective breeding programs and seed market in the European aquaculture fish industry. Aquaculture International, 24(5), 1287-1307. https://doi.org/10.1007/s10499-016-9985-0

Childs, A.R., and Fennessy, S.T., 2013, Dusky kob (*Argyrosomus japonicus*). In 'Southern African Marine Linefish Species Profiles. Special Publication. 9, 154–156.

Chung, Y., Choi, S., Jun, T., Kim, C., 2017. Genotyping-by-sequencing: a promising tool for plant genetics research and breeding. Horticulture, Environment, and Biotechnology, 58(5), 425-431. https://doi.org/10.1007/s13580-017-0297-8

Ciani, E., Cecchi, F., Castellana, E., D'Andrea, M., Incoronato, C., D'Angelo, F., 2013. Poorer resolution of low-density SNP vs. STR markers in reconstructing genetic relationships among seven Italian sheep breeds. Large Anim. Rev. 19, 236–241.

Clifford, S.L., McGinnity, P., Ferguson, A., 1998. Genetic changes in an Atlantic salmon population resulting from escaped juvenile farm salmon. J. Fish Biol. 52, 118–127. https://doi.org/10.1111/j.1095-8649.1998.tb01557.x

Collett, P., 2007. Toward the development of a rearing protocol for juvenile dusky kob, *Argyrosomus japonicus* (Pisces: Sciaenidae). http://agris.fao.org/agris-search/search.do? recordID=AV20120141662

Collett, P.D., Vine, N.G., Kaiser, H., 2008. The effect of light intensity on growth of juvenile dusky kob *Argyrosomus japonicus*. Aquac. Res. 39, 526–532. https://doi.org/10.11

Collett, P., Kaiser, H., Vine, N., 2011. The effect of crowding density on growth, food conversion ratio and survival of juvenile dusky kob *Argyrosomus japonicus* (Teleostei: Sciaenidae). African J. Aquat. Sci. 36, 155–158. https://doi.org/10.2989/16085914.2011.589113

Connell, A., 2007. Marine fish eggs and larvae from the east coast of South Africa. http://www.dnabarcodes2011.org/documents/presentations/12-1/vertebrates/1000-Steinke.pdf

Costa, V., Aprile, M., Esposito, R., Ciccodicola, A., 2012. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. European Journal of Human Genetics, 21(2), pp.134-142.

Cosart, T., Beja-Pereira, A., Chen, S., Ng, S.B., Shendure, J., Luikart, G., 2011. Exome-wide DNA capture and high throughput sequencing in domestic and wild species. BMC Genomics 12, 347.

Cowley, P., Childs, A.R., Bennett, R., 2013. The trouble with estuarine fisheries in temperate South Africa, illustrated by a case study on the Sundays Estuary. African J. Mar. Sci. 35, 117–128. https://doi.org/10.2989/1814232X.2013.789079

Daniel, S.J., 2004. Investigations into the nutritional requirements of juvenile dusky kob, *Argyrosomus japonicus* (Pisces: Sciaenidae), under ambient culture conditions. http://agris.fao.org/agris-search/search.do?recordID=AV2012060742

Dawson, D.A., Hanotte, O., Greig, C., Stewart, I.R.K., Burke, T., 2000. Polymorphic microsatellites in the blue tit *Parus caeruleus* and their cross-species utility in 20 songbird families. Mol. Ecol. 9(11):1941-1944.

Dawson, D.A., Hunter, F.M., Pandhal, J., Buckland, R., Parham, A., Jones, I.L., Bradshaw, M., Jehle, R., Burke, T., 2005. Assessment of 17 new whiskered auklet (*Aethia pygmaea*) microsatellite loci in 42 seabirds identifies 5-15 polymorphic markers for each of nine Alcinae species. Mol. Ecol. Notes, 5(2):289-297.

Dawson, D.A., Ball, A.D., Spurgin, L.G., Martín-Gálvez, D., Stewart, I.R.K., Horsburgh, G.J., 2013. High-utility conserved avian microsatellite markers enable parentage and population studies across a wide range of species. BMC Genom. 14, 176.

Department of Agriculture, Forestry and Fisheries (DAFF), 2012. Aquaculture Annual Report 2011, South Africa.

Department of Agriculture, Forestry and Fisheries (DAFF), 2016. Aquaculture Annual Report 2015, South Africa.

Dennenmoser, S., Vamosi, S.M., Nolte, A. W., Rogers, S.M., 2017. Adaptive genomic divergence under high gene flow between freshwater and brackish-water ecotypes of prickly sculpin (*Cottus asper*) revealed by Pool-Seq. Molecular Ecology, 26, 25–42.

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

Dou, S.Z., Masuda, R., Tanaka, M., Tsukamoto, K., 2004. Size hierarchies affecting the social interactions and growth of juvenile Japanese flounder, *Paralichthys olivaceus*. Aquaculture 233, 237–249. https://doi.org/10.1016/j.aquaculture.2003.09.054

Dunlop, S., Mann, B., 2012. An assessment of participation, catch and effort in the KwaZulu-Natal shore-based marine linefishery, with comments on management effectiveness. African J. Mar. Sci. 34, 479–496. https://doi.org/10.2989/1814232X.2012.725526

Ekblom, R., Wolf, J.B.W., 2014. A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications, 7, 1026–1042.

FAO, 2017. (Food and Agriculture Organization of the United Nations). Fisheries and Aquaculture Department, Global Aquaculture Production Statistics (online). http://www.fao.org/fishery/statistics/global-aquaculture-production/en.

Farmer, B.M., 2008. Comparisons of the biological and genetic characteristics of the Mulloway *Argyrosomus japonicus* (Sciaenida) in different regions of Western Australia 1–217. http://fish.gov.au/reports/Documents/2014_refs/1. Farmer 2008.pdf

Ferguson, G.J., Ward, T.M., Ivey, A., Barnes, T., 2014. Life history of *Argyrosomus japonicus*, a large sciaenid at the southern part of its global distribution: Implications for fisheries management. Fish. Res. 151, 148–157. https://doi.org/10.1016/j.fishres.2013.11.002

Fessehaye, Y., Kabir, A., Bovenhuis, H., Komen, H., 2006. Prediction of cannibalism in juvenile *Oreochromis niloticus* based on predator to prey weight ratio, and effects of age and stocking density. Aquaculture 255, 314–322. https://doi.org/10.1016/J.AQUACULTURE.2005.11.033

Fessehaye, Y., Bovenhuis, H., Rezk, M.A., Crooijmans, R., van Arendonk, J.A.M., Komen, H., 2009. Effects of relatedness and inbreeding on reproductive success of Nile tilapia (*Oreochromis niloticus*). Aquaculture 294, 180–186. https://doi.org/10.1016/j.aquaculture.2009.06.001

Fielder, D.S., Heasman, M.P., 2011. Hatchery Manual for the production of Australian Bass, Mulloway and Yellowtail Kingfish. Industry and Investment NSW. ISBN: 978 1 74256 058 8.

Fitzgibbon, Q.P., Strawbridge, A., Seymour, R.S., 2007. Metabolic scope, swimming performance and the effects of hypoxia in the mulloway, *Argyrosomus japonicus* (Pisces: Sciaenidae). Aquaculture, 270, 358–368. http://www.sciencedirect.com/science/article/pii/S0044848607003523

Foote, A.D., Vijay, N., Avila-Arcos, M.C., Baird, R.W., Durban, J. W., Fumagalli, M., Wolf, J.B.W., 2016. Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. Nature Communications, 7, 11693.

Fuentes-Pardo, A., Ruzzante, D., 2017. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Molecular Ecology, 26(20), 5369-5406. https://doi.org/10.1111/mec.14264

Ganal, M.W., Wieseke, R., Luerssen, H., Durstewitz, G., Graner, E., Plieske, J., Polley, A., 2014. High-throughput SNP Profiling of Genetic Resources in Crop Plants Using Genotyping Arrays. Genomics of Plant Genetic Resources. 113-130.

Gemayel, R., Cho J., Boeynaems, S., Verstrepen, K.J., 2012. Beyond Junk-Variable Tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. Genes. 3:461–480.

Gjedrem, T., Baranski, M., 2009. The Success of Selective Breeding in Aquaculture. Selective Breeding in Aquaculture: An Introduction. 13-23.

Gjedrem, T., Robinson, N., Rye, M., 2012. The importance of selective breeding in aquaculture to meet future demands for animal protein: A review. Aquaculture. https://doi.org/10.1016/j.aquaculture.2012.04.008

Gjedrem, T., Salte, R. and Gjøen, H. 1991. Genetic variation in susceptibility of Atlantic salmon to furunculosis. Aquaculture, 97(1), 1-6.

Glaubitz, J.C., Murrell, J.C., Moran, G.F., 2003. Effects of native forest regeneration practices on genetic diversity in *Eucalyptus consideniana*. Theor. Appl. Genet. 107:422–431.

Granevitze, Z., David, L., Twito, T., Weigend, S., Feldman, M., Hillel, J., 2014. Phylogenetic resolution power of microsatellites and various single-nucleotide polymorphism types assessed in 10 divergent chicken populations. Anim. Genet. 45, 87–95.

Grant, W.S., Garcia-Marin, J.L., Utter, F.M., 1999. Defining population boundaries for fishery management. S. Mustafa (Ed.), Genetics in Sustainable Fisheries Management. Blackwell Scientific Publications, Oxford: 27-72

Griffiths, M.H., 1996. Life history of the dusky kob *Argyrosomus japonicus* (Sciaenidae) off the east coast of South Africa. South African J. Mar. Sci. 17, 135–154. https://doi.org/10.2989/025776196784158653

Griffiths, M.H., 1997. Feeding ecology of South African *Argyrosomus japonicus* (Pisces: Sciaenidae), with emphasis on the Eastern Cape surf zone. South African J. Mar. Sci. 18, 249–264. https://doi.org/10.2989/025776197784160965

Griffiths, M.H., 1997a. Management of South African dusky kob *Argyrosomus japonicus* (Sciaenidae) based on per-recruit models. South African J. Mar. Sci. 18, 213–228. https://doi.org/10.2989/025776197784160938

Griffiths, M.H., 2000. Long-term trends in catch and effort of commercial line fish off South Africa's Cape Province: snapshots of the 20th century. South African J. Mar. Sci. 22, 81–110. https://doi.org/10.2989/025776100784125663

Griffiths, M.H., Heemstra, P.C., 1995. A contribution to the taxonomy of the marine fish genus *Argyrosomus* (Perciformes: Sciaenidae), with descriptions of two new species from southern Africa. Ichthyol. Bull. 65. http://fish.gov.au/reports/Documents/2014_refs/2.

Griffiths, M.H., Hecht, T., 1995b. Age and growth of South African dusky kob *Argyrosomus japonicus* (Sciaenidae) based on otoliths. South African J. Mar. Sci. 16, 119–128. https://doi.org/10.2989/025776197784160938

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D.B., Lander, E.S., Nusbaum., 2009 Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 27 (2): 182-189

Guichoux, E., Lagache, L., Wagner, S., Chaumeil, P., Léger, P., Lepais, O., 2011. Current trends in microsatellite genotyping. Mol. Ecol. Resour. 11, 591–611.

Hadrys, H., Balick, M., Schierwater, B.,1992. Application of random amplified polymorphic DNA (RAPD) in molecular ecology. Mol. Ecol. 1: 55-63.

Hansson, B. and Kawabe, A., 2005. A simple method to score single nucleotide polymorphisms based on allele-specific PCR and primer-induced fragment-length variation. Molecular Ecology Notes, 5(3), 692-696. https://doi.org/10.1111/j.1471-8286.2005.01033.x

Harris, S. A., Cyrus, D. P, Forbes, A.T., 1995. The larva fish assemblage at the mouth of the Kosi Estuary, KwaZulu-Natal, South Africa. S. Afr. J. mar. Sci. 16: 351-364.

Haasl, R., Payseur, B., 2010. Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. Heredity, 106(1), 158-171. https://doi.org/10.1038/hdy.2010.21

Hauser, L., Baird, M., Hilborn, R., Seeb, L.W., Seeb, J.E., 2011. An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. Mol. Ecol. Resour. 11,150–161.

Hayes, B., He, J., Moen, T., Bennewitz, J., 2006. Use of molecular markers to maximise diversity of founder populations for aquaculture breeding programs. Aquaculture 255, 573–578. https://doi.org/10.1016/j.aquaculture.2005.11.038

Heaton, M., Harhay, G., Bennett, G., Stone, R., Grosse, W., Casas, E., Keele, J., Smith, T., Chitko-McKown, C., Laegreid, W., 2002. Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. Mammalian Genome, 13(5), 272-281. https://doi.org/10.1007/s00335-001-2146-3

Hecht, T., Pienaar, A.G., 1993. A Review of Cannibalism and its Implications in Fish Larviculture. J. World Aquac. Soc. 24, 246–261. https://doi.org/10.1111/j.1749-7345.1993.tb00014.x

Hemmer-Hansen, J., Nielsen, E., Meldrup, D., Mittleholzer, C. 2011. Identification of single nucleotide polymorphisms in candidate genes for growth and reproduction in a nonmodel organism; the Atlantic cod, *Gadus morhua*. Molecular Ecology Resources, 11, 71-80. https://doi.org/10.1111/j.1755-0998.2010.02940.x

Herbert, N.A., 2013. Practical Aspects of Induced Exercise in Finfish Aquaculture. Swimming Physiology of Fish. 377-405

Hodges, E., Rooks, M., Xuan, Z., Bhattacharjee, A., Benjamin, Gordon, D., Brizuela, L.R., McCombie, W., Hannon, G.J., 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nat Protoc, 4:960-974.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., McCombie, W.R., 2007. Genome-wide in situ exon capture for selective resequencing. Nat Genet. 39 (12), 1522-1527.

Hong, W., Zhang, Q., 2003. Review of captive bred species and fry production of marine fish in China, in: Aquaculture. Elsevier, 305–318. https://doi.org/10.1016/S0044-8486(03)00511-8

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

Hseu, J.R., Shen, P.S., Huang, W. Bin-Hwang, P.P., 2007. Logistic regression analysis applied to cannibalism in the giant grouper *Epinephelus lanceolatus* fry: Short paper. Fish. Sci. 73, 472–474. https://doi.org/10.1111/j.1444-2906.2007.01358.x

Hunt, M., Burnside, E., Freeman, M., Wilton, J., 1974. Genetic Gain When Sire Sampling and Proving Programs Vary in Different Artificial Insemination Population Sizes. Journal of Dairy Science, 57(2), 251-257.

Hurst, G. and Jiggins, F., 2005. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proceedings of the Royal Society B: Biological Sciences, 272(1572), 1525-1534.

Hutchings, K., Lamberth, S.J., 2003. Likely impacts of an eastward expansion of the inshore gill-net fishery in the Western Cape, South Africa: Implications for management. Mar. Freshw. Res. 54, 39–56

Ignal, A.V., Ilan, D.M., 2002. A review on SNP and other types of molecular markers and their use in animal genetics 34, 275–305. https://doi.org/10.1051/gse

Janssen, K., Chavanne, H., Berentsen, P., Komen, H., 2017. Impact of selective breeding on European aquaculture. Aquaculture, 472, 8-16. https://doi.org/ 10.1016/j.aquaculture.2016.03.012

Jenkins, S.F., 2018. Genetic and phenotypic characterisation of commercial dusky kob (*Argyrosomus japonicus*) cohorts, MSc thesis. Stellenbosch University, Stellenbosch.

Kailola, P.J., Williams, P.C., Stewart, R.E., Reichelt, A., McNee, C., 1993. Australian Fisheries Resources. Canberra, Australia. Brisbane, Bureau of Resource Sciences, Fisheries Research and Development Corporation. Australian Fisheries Resources: 318-320.

Kaiser, H., Collett, P.D., Vine, N.G., 2011. The effect of feeding regimen on growth, food conversion ratio and size variation in juvenile dusky kob *Argyrosomus japonicus* (Teleostei: Sciaenidae). African J. Aquat. Sci. 36, 83–88. https://doi.org/10.2989/16085914.2011.559712

Kalinowski, S.T., 2002. How many alleles per locus should be used to estimate genetic distances? Heredity, 88, 62–65.

Kestemont, P., Jourdan, S., Houbart, M., Mélard, C., Paspatis, M., Fontaine, P., Cuvier, A., Kentouri, M., Baras, E., 2003. Size heterogeneity, cannibalism and competition in

cultured predatory fish larvae: Biotic and abiotic influences, in: Aquaculture. Elsevier, 333–356. https://doi.org/10.1016/S0044-8486(03)00513-1

Kim, J., Stewart, R., Kim, S., Yang, S., Shin, I., Kim, Y., Yoon, J., 2008. BDNF genotype potentially modifying the association between incident stroke and depression. Neurobiology of Aging, 29(5), 789-792. https://doi.org.10.1016/j.neurobiolaging.2006.11.021

Kolkovski, S., Tandler, A., Kissil, G., Gertler, A., 1993. The effect of dietary exogenous digestive enzymes on ingestion, assimilation, growth and survival of gilthead seabream (*Sparus aurata, Sparidae, Linnaeus*) larvae. Fish Physiology and Biochemistry, 12(3), 203-209. https://doi.org/10.1007/bf00004368

Krawczak, M., 1999. Informativity assessment for biallelic single nucleotide polymorphisms. Electrophoresis, 20, 1676–1681.

Lagardère, J.P., Mariani, A., 2006. Spawning sounds in meagre *Argyrosomus regius* recorded in the Gironde estuary, France. J. Fish Biol. 69, 1697–1708. https://doi.org/10.1111/j.1095-8649.2006.01237.x

Lamichhaney, S., Martinez Barrio, A., Rafati, N., Sundstrom, G., Rubin, C.J., Gilbert, E. R., Andersson, L., 2012. Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proceedings of the National Academy of Sciences, 109, 19345–19350.

Lamichhaney, S., Fuentes-Pardo, A. P., Rafati, N., Ryman, N., McCracken, G. R., Bourne, C., Andersson, L., 2017. Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean. Proceedings of the National Academy of Sciences, 114(17), E3452–E3461.

Lee, C.S., Ostrowski, A.C., 2001. Current status of marine finfish larviculture in the United States, in: Aquaculture. 89–109. https://doi.org/10.1016/S0044-8486(01)00695-0

Lee, J., Izzah, N.K., Jayakodi, M., Perumal, S., Joh HJ., Lee HJ., 2015. Genome-wide SNP identification and QTL mapping for black rot resistance in cabbage. BMC Plant Biol. 15:32. https://doi.org/10.1186/s12870-015-0424-6. PMID: 25644124. PMCID: 4323122.

Leshchiner, I., 2012. Mutation Mapping and Identification by Whole-Genome Sequencing. Genome Research, 22(8), 1541–1548

Li, H., Wren, J., 2014. Toward better understanding of artefacts in variant calling from high-coverage samples. Bioinformatics, 30, 2843–2851. https://doi.org/10.1093/bioinformatics/btu356

Li, L., Lin H., Tang, W., Liu, D., Bao, B., Yang, J., 2017. Population genetic structure in wild and aquaculture populations of *Hemibarbus maculates* inferred from microsatellites markers. Aquac. Fish. 2, 78–83. https://doi.org/10.1016/j.aaf.2017.03.004

Liao, I.C., and Chang, E.Y., 2002. Timing and factors affecting cannibalism in red drum, *Sciaenops ocellatus*, larvae in captivity. Environ. Biol. Fish. 63, 229–233. https://doi.org/10.1023/A:1014244102276

Lillehammer, M., Meuwissen, T., Sonesson, A., 2011. Genomic selection for maternal traits in pigs. Journal of Animal Science, 89(12), 3908-3916. https://doi.org/10.2527/jas.2011-4044

Lindblad-Toh, K., Winchester, E., Daly, M., Wang, D., Hirschhorn, J., Laviolette, J., Ardlie, K., Reich, D., Robinson, E., Sklar, P., Shah, N., Thomas, D., Fan, J., Gingeras, T., Warrington, J., Patil, N., Hudson, T., Lander, E., 2000. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nature Genetics, 24(4), 381-386.

Liu, P., Xia, J.H., Lin, G., Sun, F., Liu, F., Lim, H.S., Yue G.H., 2012. Molecular Parentage Analysis Is Essential in Breeding Asian Seabass. PLoS ONE 7, e51142. https://doi.org/10.1371/journal.pone.0051142

Liu, X., Xia, J., Pang, H., Yue, G., 2017. Who eats whom, when and why? Juvenile cannibalism in fish Asian seabass. Aquac. Fish. 2, 1–9. https://doi.org/10.1016/j.aaf.2016.12.001

Liu, Z.J., Cordes, J.F., 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture. https://doi.org/10.1016/j.aquaculture.2004.05.027

Livingstone, D.S., Motamayor, J.C., Schnell, R.J., Cariaga, K., Freeman, B., Meerow, A.W., 2010. Development of single nucleotide polymorphism markers in *Theobroma cacao* and comparison to simple sequence repeat markers for genotyping of Cameroon clones. Mol. Breeding. 27, 93–106.

Loughnan, S.R., Domingos, J.A., Smith-Keune, C., Forrester, J.P., Jerry, D.R., Beheregaray, L.B., Robinson, N.A., 2013. Broodstock contribution after mass

spawning and size grading in barramundi (*Lates calcarifer*, Bloch). Aquaculture 404–405, 139–149. https://doi.org/10.1016/j.aquaculture.2013.04.014

Lubzens, E., Zmora, O., Barr, Y., 2001 Biotechnology and aquaculture of rotifers. In: Sanoamuang, L., Segers, H., Shiel, R.J., Gulati, R.D. (eds) Rotifera IX. Developments in Hydrobiology, 153. Springer, Dordrecht

Luikart, G., England, P.R., Tallmon, D., Jordan, S., Taberlet, P., 2003 The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet, 4:981-994

Magoulas, A., Gjetvaj, B., Terzoglou, V., Zouros, E., 1998. Three polymorphic microsatellites in the Japanese oyster *Crassostrea gigas* (Thunberg). Anim. Genet. 29, 69–70. http://www.marbigen.org/content/three-polymorphic-microsatellites-japanese-oyster-icrassostrea-gigasi-thunberg

Martino, A., Mancuso, T., Rossi, A.M., 2010. Application of high-resolution melting to large-scale, high-throughput SNP genotyping a comparison with the TaqMan® method. J Biomol Screen. 15: 623–629. https://doi.org/10.1177/1087057110365900

Mason, A., Higgins, E., Snowdon, R., Batley, J., Stein, A., Werner, C., Parkin, I., 2017. A user guide to the Brassica 60K Illumina Infinium™ SNP genotyping array. Theoretical and Applied Genetics, 130(4), 621-633.https://doi.org/10.1007/s00122-016-2849-1

May, B., 2003. Allozyme variation. E. M. Hallerman (Ed.), Population genetics: principles and applications for fisheries scientists. American Fisheries Society, Bethesda, Maryland: 23-36.

McMahon, R., Blake, R., Richard Shumway, C., Tomaszewski, M., 1985. Selection of Young and Proven Holstein Artificial Insemination Sires to Maximize Profits from Milk. Journal of Dairy Science, 68(9), 2303-2308.

Mead, A., Griffiths, C., Branch, G., McQuaid, C., Blamey, L., Bolton, J., Anderson, R., Dufois, F., Rouault, M., Froneman, P., Whitfield, A., Harris, L., Nel, R., Pillay, D. and Adams, J. 2013. Human-mediated drivers of change — impacts on coastal ecosystems and marine biota of South Africa. African Journal of Marine Science, 35(3), 403-425.

Mirimin, L., Kerwath, S.E., Macey, B.M., Bester-van der Merwe, A.E., Lamberth, S.J., Bloomer, P., Roodt-Wilding, R., 2014. Identification of naturally occurring hybrids

between two overexploited sciaenid species along the South African coast. Mol. Phylogenet. Evol. 76, 30–33. https://doi.org/10.1016/j.ympev.2014.02.010

Mirimin, L., Macey, B., Kerwath, S., Lamberth, S., Bester-Van Der Merwe, A., Cowley, P., Roodt-Wilding, R., 2015. Genetic analyses reveal declining trends and low effective population size in an overfished South African sciaenid species, the dusky kob (*Argyrosomus japonicus*). Marine and Freshwater Research 67, 266–276. https://doi.org/10.1071/MF14345

Misztal, I., 2006. Properties of random regression models using linear splines. Journal of Animal Breeding and Genetics, 123(2), 74-80. https://doi.org/10.1111/j.1439-0388.2006.00582.x

Morin, P., Luikart, G., Wayne, R., 2004. SNPs in ecology, evolution and conservation. Trends in Ecology and Evolution, 19(4), 208-216. https://doi.org/10.1016/j.tree.2004.01.009

Munilla-Moran, R., Stark, J., Barbour, A., 1990. The role of exogenous enzymes in digestion in cultured turbot larvae (*Scophthalmus maximus L*.). Aquaculture, 88(3-4), 337-350.https://doi.org/10.1016/0044-8486(90)90159-k

Musson, M., Kaiser, H., 2014. Development of the digestive system in dusky kob, Argyrosomus japonicus, larvae. Aquac. Int. 22, 783–796. https://doi.org/10.1007/s10499-013-9706-x

Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., Shendure, J., Bamshad, M.J., 2010. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 42 (1): 30-35.

Obholzer, N., 2012. “Rapid Positional Cloning of Zebrafish Mutations by Linkage and Homozygosity Mapping Using Whole-Genome Sequencing.” Development, vol. 139(22), 4280–4290.

O’Brien, S.J., 1991. Molecular genome mapping: lessons and prospects. Current Opinion in Genetic Development,1(1), 105-111.

Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J., Zwick, M.E., 2007. Microarray-based genomic selection for high-throughput resequencing. Nat Methods, 4:907-909

O'Sullivan, D. and Ryan, M., 2001. Mulloway Aquaculture in Southern Australia. Primary Industries and Resources Factsheet. Australian Government Publ. Service, Canberra

Otgaar, M., 2012. Kob – Kabeljous. Fish SA, .1–5. Available at: http://fishsa.co.za/species/kob-argyrosomus-japonicu s/ Fish-farming in South Africa: A study of the market environment and the suitable species

Ozsolak, F., Milos, P., 2010. RNA sequencing: advances, challenges and opportunities. Nature Reviews Genetics, 12(2), 87-98.

Ozsolak, F., Milos, P. M., 2011. RNA sequencing: Advances, challenges and opportunities. Nature Reviews Genetics, 12, 87–98.

Pacitti, D., Wang, T., Page, M., Martin, S., Sweetman, J., Feldmann, J., Secombes, C., 2013. Characterization of cytosolic glutathione peroxidase and phospholipid-hydroperoxide glutathione peroxidase genes in rainbow trout (*Oncorhynchus mykiss*) and their modulation by in vitro selenium exposure. Aquatic Toxicology, 130-131, 97-111. https://doi.org/10.1016/j.aquatox.2012.12.020

Parla, J.S., Iossifov, I., Grabill, I., Spector, M.S., Kramer, M., McCombie, W.R., 2011. A comparative analysis of exome capture. Genome Biol. 12, R97. https://doi.org/10.1186/gb-2011-12-9-r97

Parsons, M.J.G., McCauley, R.D., 2017. Sound Production by Mulloway (*Argyrosomus japonicus*) and Variation Within Individual Calls. Acoust. Aust. 45, 1–12. https://doi.org/10.1007/s40857-017-0112-9

Payseur, B.A., Cutter, A.D., 2006. Integrating patterns of polymorphism at SNPs and STRs. Trends Genet. 22, 424–429.

Payseur, B.A., Jing, P., 2011. A genome wide comparison of population structure at STRPs and nearby SNPs in humans. Mol. Biol. Evol. 26, 1369–1377.

Pearson, T. H., Black, K. D., 2000. The environmental impacts of marine fish cage culture. Environmental impacts of aquaculture. 1-31.

Pradervand, P., Mann, B.Q., Bellis, M.F., 2007. Long-term trends in the competitive shore fishery along the KwaZulu-Natal coast, South Africa. African Zool. 42, 216–236. https://doi.org/10.3377/1562-7020(2007)42

Primmer, C.R., Borge, T., Lindell, J., Saetre, G.P., 2002. Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. Molecular Ecology, 11, 603–612.

Qin, J.G., Mittiga, L., Ottolenghi, F., 2004. Cannibalism Reduction in Juvenile Barramundi *Lates calcarifer* by Providing Refuges and Low Light. J. World Aquac. Soc. 35, 113–118. https://doi.org/10.1111/j.1749-7345.2004.tb01067.x

Raca, G., Jackson, C., Warman, B., Bair, T., Schimmenti, L.A., 2010. High throughput sequencing in research and diagnostics of ocular birth defects. Molecular Genetics and Metabolism. 100, 184-192.

Raffan, E., Hurst, L.A., Turki, S.A., Carpenter, G., Scott, C., Daly, A., Coffey, A., Bhaskar, S., Howard, E., Khan, N., Kingston, H., Palotie, A., Savage, D.B., O'Driscoll, M., Smith, C., O'Rahilly, S., Barroso, I., Semple, R.K.M., 2011. Early Diagnosis of Werner's Syndrome Using Exome-Wide Sequencing in a Single, Atypical Patient. Front Endocrinol. 2, 8.

Ramcharitar, J., Gannon, D., Popper, A., 2006. Bioacoustics of Fishes of the Family Sciaenidae (Croakers and Drums). Transactions of the American Fisheries Society, 135(5),1409-1431.

Roach, J.C., Glusman, G., Rowen, L., Kaur, A., Purcell, M.K., Smith, K.D., 2005. The evolution of vertebrate Toll-like receptors. Proc. Natl. Acad. Sci. 102, 9577e9582.

Rosenblum, E.B., Novembre, J., 2007. Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard. Journal of Heredity, 98, 331–336.

Ross, C.T., Weise, J.A., Bonnar, S., Nolin, D., Satkoski, T.J., Smith D.G., 2014. An empirical comparison of short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs) for relatedness estimation in Chinese rhesus macaques (*Macaca mulatta*) Am. J. Primatol. 76, 313–324.

Rurangwa, E., Verdegem, M., 2014. Microorganisms in recirculating aquaculture systems and their management. Reviews in Aquaculture, 7(2), 117-130.

Ryan, S., Willer, J., Marjoram, L., Bagwell, J., Mankiewicz, J., Leshchiner, I., Goessling, W., Bagnat, M., Katsanis, N., 2013. Rapid identification of kidney cyst mutations by whole

exome sequencing in zebrafish. Development 140, 4445–4451. https://doi.org/10.1242/dev.101170Seafood

Saker, M., Griffiths, D., 2000. The effect of temperature on growth and cylindrospermopsin content of seven isolates of Cylindrospermopsis raciborskii (*Nostocales, Cyanophyceae*) from water bodies in northern Australia. Phycologia, 39(4), 349-354. https://doi.org/10.2216/i0031-8884-39-4-349.1

Sauer, W. H. H., Hecht, T., Britz, P. J., Mather, D., 2003. An economic and sectoral study of the South. African fishing industry. Volume 2. Fishery profiles. Report prepared for Marine and Coastal Management by Rhodes University. South Africa.

Santure, A.W., Stapley, J., Ball, A.D., Birkhead, T.R., Burke, T., Slate, J., 2010. On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. Mol. Ecol. 19, 1439–1451.

Seeb, J.E., Carvalho, G., Hauser, L., Naish, K., Roberts, S. Seeb, L.W., 2011. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Mol. Ecol. Resour. 11, 1–8.

Sekino, M., Sugaya, T., Hara, M., Taniguchi, N., 2004. Relatedness inferred from microsatellite genotypes as a tool for broodstock management of Japanese flounder Paralichthys olivaceus. Aquaculture 233, 163–172. https://doi.org/10.1016/j.aquaculture.2003.11.008

Silberschneider, V., Gray, C.A., 2007. Synopsis of biological, fisheries and aquaculture-related information on mulloway *Argyrosomus japonicus* (Pisces: Sciaenidae), with particular reference to Australia. J. Appl. Ichthyol. 24, 7–17. https://doi.org/10.1111/j.1439-0426.2007.00913.x

Silberschneider, V., Gray, C.A., Stewart, J., 2009. Age, growth, maturity and the overfishing of the iconic sciaenid, *Argyrosomus japonicus*, in south-eastern, Australia. Fish. Res. 95, 220–229. https://doi.org/10.1016/j.fishres.2008.09.002

Skaalav, Ø., 2004. Microsatellite Analysis in Domesticated and Wild Atlantic Salmon (*Salmo Salar L.*): Allelic Diversity and Identification of Individuals. Aquaculture, 240 (1–4), 131–143. https://doi.org/10.1016/j.aquaculture.2004.07.009.

Skibinski, H., 2005. Light positively affects foraging success of the night feeding zooplanktivorous fish hardyhead silverside (*Atherinomorus lacunosus*). Uppsala University.

Slate, J., 2008. Robustness of linkage maps in natural populations: a simulation study. Proceedings of the Royal Society B: Biological Sciences, 275(1635), 695-702.

Smith, C.T., Antonovich, A., Templin, W.D., 2007. Impacts of marker class bias relative to locus-specific variability on population inferences in chinook salmon: a comparison of single-nucleotide polymorphisms with short tandem repeats and allozymes. Transactions of the American Fisheries Society, 136, 1674–1687.

Smith, C.T., Seeb, L.W., 2008. Number of alleles as a predictor of the relative assignment accuracy of STR and SNP baselines for chum salmon. Transactions of the American Fisheries Society, 137,751–762.

Spasic, M., Poleksic, V., Stankovic, M., Dulic, Z., Raskovic, B., Zivic, I., Ciric, M., Relic, R., Vukojevic, D., Boskovic, D., Markovic, Z., 2010. Selective breeding programme of common carp (*Cyprinus carpio L.*) in Serbia: Preliminary results. Journal of Agricultural Sciences, Belgrade, 55(3), 243-251. https://doi.org/ 10.2298/JAS1003243S

Stolle, E., Kidner, J., Moritz, R., 2013. Patterns of Evolutionary Conservation of Microsatellites (SSRs) Suggest a Faster Rate of Genome Evolution in Hymenoptera Than in Diptera. Genome Biology and Evolution, 5(1), 151-162. https://doi.org/10.1093/gbe/evs133

Stoneham, T., Kuhn, D., Taylor, D., Neilson, A., Smith, S., Gatlin, D., Chu, H., O'Keefe, S., 2018. Production of omega-3 enriched tilapia through the dietary use of algae meal or fish oil: Improved nutrient value of fillet and offal. PLOS ONE, 13(4), p.e0194241. https://doi.org/10.1371/journal.pone.0194241

Stoneking, M., 2001. Single nucleotide polymorphisms. From the evolutionary past. Nature. 409, 821–822.

Sun, J.X., Mullikin, J.C., Patterson, N., Reich, D.E., 2009 Microsatellites are molecular clocks that support accurate inferences about history. Mol. Biol. Evol. 26, 1017–1027.

Syvänen, A., 2005. Toward genome-wide SNP genotyping. Nature Genetics, 37(S6), S5-S10.

Tao, W., Boulding, E., 2003. Associations between single nucleotide polymorphisms in candidate genes and growth rate in Arctic charr (*Salvelinus alpinus L.).* Heredity, 91(1), 60-69. https://doi.org/10.1038/sj.hdy.6800281

Taylor, M., Laffan, S., Fielder, S., Suthers, I., 2006. Key habitat and home range of mulloway *Argyrosomus japonicus* in a south-east Australian estuary: finding the estuarine niche to optimise stocking. Marine Ecology Progress Series, 328, .237-247.

Thomson, M., 2014. High-Throughput SNP Genotyping to Accelerate Crop Improvement. Plant Breeding and Biotechnology, 2(3), 195-212. https://doi.org/10.9787/PBB.2014.2.3.195

Timmer, R., Magellan, K., 2011. The effects of light intensity and color on aggressive interactions in the dusky kob, *Argyrosomus Japonicus*. Isr. J. Aquac. - Bamidgeh 63. http://www.academia.edu/8073127

Trewavas, E., 1977. The sciaenid fishes (croakers or drums) of the Indo-West-Pacific. Trans. Zool. Soc. Lond. 33, 253–541. https://doi.org/10.1111/j.1096-3642.1977.tb00052.x

Väli, Ü., Saag, P., Dombrovski, V., Meyburg, B.U., Maciorowski, G., Mizera, T., 2010. Microsatellites and single nucleotide polymorphisms in avian hybrid identification: a comparative case study. J. Avian Biol. 41, 34–49.

Vallender, E.J., 2011. Expanding whole exome resequencing into non- human primates. Genome Biol. 12, R87

van Marle-Köster, E., Visser, C., 2018. Genetic Improvement in South African Livestock: Can Genomics Bridge the Gap Between the Developed and Developing Sectors?. Frontiers in Genetics, 9, 331. https://dx.doi.org/10.3389%2Ffgene.2018.00331

Vignal, A., Milan, D., SanCristobal, M., Eggen, A., 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution, 34(3), 275-305. https://doi.org/10.1051/gse:2002009

Viljoen, M.J., 2019. Determining production characteristics of dusky kob*, Argyrosomus japonicus*, grown in sea cages under commercial conditions in Richards Bay, South Africa, MSc thesis. Stellenbosch University, Stellenbosch.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T., Hornes, M., Zabeau, M., 1995. AFLP: A new technique for DNA fingerprinting. Nucl. Ac Res., 23(21), 4407–4414. https://doi.org/10.1093/nar/23.21.4407

Walford, J., Lam, T.J., 1993. Development of Digestive Tract and Proteolytic Enzyme Activity in Seabass (*Lates Calcarifer*) Larvae and Juveniles. Aquaculture, 109(2), 187–205

Wang, D., 1998. Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. Science, 280(5366), 1077-1082. https:// doi.org/10.1126/science.280.5366.1077

Wang, H., Chattopadhyay, A., Li, Z., Daines, B., Li, Y., Gao, C., Gibbs, R., Zhang, K., Chen, R., 2010. Rapid identification of heterozygous mutations in Drosophila melanogaster using genomic capture sequencing. Genome Research.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., Watson, M., 2015. Exome Sequencing: Current and Future Perspectives. Genes|Genomes|Genetics 5, 1543–1550.

Wayne, M., McIntyre, L., 2002. Combining mapping and arraying: An approach to candidate gene identification. Proceedings of the National Academy of Sciences, 99(23), 14903-14906. https://doi.org/10.1073/pnas.222549199

Whitfield, A., 1998. Biology and ecology of fishes in Southern African estuaries, in: Icthyological Monographs; 2, 183–203. https://www.biodiversitylibrary.org/bibliography/141872

Wilcox, M., Shumway, C., Blake, R., Tomaszewski, M., 1984. Selection of Artificial Insemination Sires to Maximize Profits. Journal of Dairy Science, 67(10), 2407-2419.

Williams, J.L., 2005. The use of marker-assisted selection in animal breeding and biotechnology. Rev. sci. tech. 24 (1), 379-391.

Worthey, E.A., Mayer, A.N., Syverson, G.D., Helbling, D., Bonacci, B.B., Decker, B., Serpe J.M., Dasu, T., Tschannen, M.R., Veith, R.L., Basehore, M.J., Broeckel, U., Tomita-Mitchell, A., Arca, M.J., Casper, J.T., Margolis, D.A., Bick D.P., Hessner, M.J., Routes J.M., Verbsky, J.W., Jacob, HJ, Dimmock, D.P., 2011. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med. 13 (3), 255-262.

Xing, C., Schumacher, F., Xing, G., Lu, Q., Wang, T., Elston, R., 2005. Comparison of microsatellites, single-nucleotide polymorphisms (SNPs) and composite markers derived from SNPs in linkage analysis. BMC Genetics, 6(1), S29. https://doi.org/10.1186/1471-2156-6-s1-s29

Yang, W., Tempelman, R., 2011. A Bayesian Antedependence Model for Whole Genome Prediction. Genetics, 190(4), 1491-1501. https://doi.org/10.1534/genetics.111.131540

Yoshimatsu, T., and Hossain, M., 2014. “Recent Advances in the High-Density Rotifer Culture in Japan.” Aquaculture International, 22(5), 1587–1603.

# CHAPTER 2

## Transferability of a model organism's solution-based exon-capture kit, in the non-model organism the dusky kob

---


### Abstract

The use of gene-targeted, genome-wide markers are essential for the advancement of evolutionary biology, animal production, and biodiversity conservation as it increases our understanding of genetic processes underlying complex traits, adaptation and speciation. Transcriptome sequencing, although gene targeted does have limitations, and thus a more flexible, gene-targeted method is required for the identification of genome-wide SNPs over a number of individuals and genes. This study demonstrates the usefulness of a recent technology, exome capture, in the discovery of genome-wide markers in a species with limited available genomic resources. Exome sequencing was performed in *Argyrosomus japonicus* using a model organism's exome capture kit, (*Danio rerio*), zebrafish on the IonTorrent$^{TM}$ platform. By applying this method, the capture kit was able to successfully sequence, 6,623 of the 346,263 exons found within zebrafish as well as a large number of exons that could potentially be species-specific. Characterisation of the exon regions determined that the exons were distributed among the various functional classes of GO and KOG databases indicating how the exome data, even though not covering the entire genome, encompasses a broad gene functional diversity. The annotated exons were used to identify genomic regions involved in development and metabolic processes, gene expression, as well as regions involved in processes of stress response. Overall, the exome data proved to be a valuable resource for the identification of variants, with variant detection identifying 4.5 million potential molecular markers with a total of 2.8 million putative SNPs and 3,276 tandem repeats. These variants were spread across the exome regions with a SNP occurring approximately every 1000 nt. This study demonstrated the ability of exon capture to be customised for cross-species capture to assist in molecular marker discovery for non-model organisms with limited or no genomic resources.

# Background

## 2.1) Introduction

Single nucleotide polymorphisms (SNPs) have become the marker of choice in numerous genomic studies focusing on population diversity, conservation genetics and functional gene identification for biological traits and selective breeding (Seeb et al., 2011). These variants are found to be the most frequent type of variation to occur within the genome, as their mostly biallelic nature increases their abundance and accurate high-throughput scoring, making these markers ideal for studying the genomic patterns of inheritance of biological traits (Schlötterer, 2004; Stickney, 2002). Due to the popularity of these markers, a variety of technologies were developed, which led to the development and application of SNP microarrays, which enabled researchers to study genome-wide SNPs in a high-throughput manner. However, without reference sequences, microarray approaches are unable to discover novel SNP loci thus, hindering the wider usage of this application in non-model species, particularly in endangered and emerging organisms, of economic importance for fisheries and aquaculture (De Donato et al., 2013; Xiao et al., 2016). Unfortunately, for many eukaryotic species with large and complex genomes, it remains costly and technically challenging to obtain and develop genome resources, such as whole genome assemblies or genome-wide SNP panels (Cosart et al., 2011). Thus, targeted sequencing of the genome using next-generation sequencing has become a powerful method for identifying DNA variation that is associated with traits of interest. Despite the continued advances made in sequencing technologies, there is still value in a gene-targeted, next-generation sequencing method, that can efficiently sequence a large number of genes from many individuals targeting functional regions of the genome (Schott et al., 2017). Conventionally, transcriptome sequencing is used (*e.g.* Wang et al., 2009), however this normally relies on fresh tissue and could also create an ascertainment bias, due to differential gene expression, in subsequent population genetic analyses (Ozsolak and Milos, 2010). A more flexible reduced complexity, targeted sequencing method might thus be warranted, and a variety of methods that can be utilised to target, enrich and capture specific sections of the genome are available.

Targeted capture is a method that is able to selectively enrich genomic libraries for particular regions of interest, this is performed by using a set of DNA or RNA probes as bait (Gnirke et al., 2009). The utilisation of probes can be performed on microarray chips or in solution, although the principal remains the same (Albert et al., 2007; Gnirke et al., 2009; Hodges et

al., 2007; Okou et al. 2007). Once the regions of interest have been identified, probes can be designed, whether it is only for a small portion of the genome with specific features or for all the protein coding regions. The most widely used method for enrichment currently is the use of hybrid enrichment, which was originally proposed to capture and re-sequence the human exome (Albert et al., 2007; Gnirke et al., 2009; Hodges et al., 2007; Porreca et al., 2007; Okou et al., 2007). Although it has since been applied to other model species for applications, such as variant discovery and population genetics (Jones and Good, 2015; Warr et al. 2015), it uses in non-model organism remains limited due to the limited ability of genome sequence for these species.

A recent whole-exome sequencing study, utilising a solution-based exon-capture kit (designed specifically for the model zebrafish (*Danio rerio*)), demonstrated the utility of the approach in the finfish model species, generating a high number of quality reads as well as a reliable SNP-markers (Ryan et al., 2013). Against this background, the objective of this study was to evaluate the transferability of the zebrafish exon-capture kit to non-model finfish species, such as dusky kob (*Argyrosomus japonicus*). Theoretically, it is assumed to be possible, as despite the divergence between species within the same taxonomic grouping, the functional elements of the genome tend to be highly conserved (Dickmeis and Füller, 2005). This has been successfully demonstrated in a study performed by Cosart et al. (2011), where SNPs were successfully identified by capturing the exomes of zebu cattle (*Bos indicus*) and American Bison (*Bison bison*) using a commercial cattle kit for *Bos taurus*. A similar approach involving the sequencing of Neanderthals and non-human primates using human capture kits (Burbano et al., 2010; Vallender, 2011). These studies have demonstrated the enhanced ability of WES strategies to identify genetic markers, without the availability of a reference genome when utilising the exome capture kits of model organisms therefore aiding in the accelerated development of genomic resources for non-model species (Warr et al., 2015).

## 2.2) Methods and Materials

### *2.2.1) Study populations and DNA extraction*

Sixteen individual fish samples were collected, from a single commercially produced F1-generation, at approximately two years of age (*i.e.* at marketable-size). All cultured animals were produced from a single spawning event through the mating of two broodstock individuals and were reared in a single tank. Fin clip tissue from all cultured individuals was preserved in 70% ethanol and stored at -20°C, after which genomic DNA extraction was

performed on each specimen as a single extraction using a standard CTAB DNA extraction protocol (Saghai-Maroof et al., 1984).

*2.2.2) Library construction and sequencing*

*Sample preparation*

Library preparation and DNA sequencing was performed at the Sequencing Unit of the Central Analytical Facility of Stellenbosch University, Stellenbosch, South Africa. For library preparation, 1µg of gDNA was diluted in a low-TE buffer (10 mM Tris·Cl, pH 8.0. 1 mM EDTA) to a concentration of 7.7ng/µl in a final volume of 130µl. The diluted gDNA (130µl) was fragmented using the Covaris S2 in frequency sweeping mode, with the bath temperature between 4°C and 6°C, at 5% intensity, 20% duty cycle, 200 cycles/burst, for a 60 sec treatment time across seven cycles. The volume of sheared gDNA was adjusted to 158µl with a low TE buffer. Forty microlitres of end-repair buffer and 2µl end-repair enzyme from the IonXpress™ Fragment Library Kit (ThermoFisher™ Scientific) was added to the gDNA, pipette mixed and briefly centrifuged, before incubating at room temperature (~21°C) for 20min. To purify the sheared, end-repaired gDNA, 220ul Ampure™ XP reagent was added to each sample, incubated at room temperature for 5min and allowed to separate on a magnetic stand for a further 5min, before collecting 410µl of the eluate. A further 143.5µl Ampure™ XP reagent was added to the 410µl eluate, homogenised and incubated at room temperature for 5 min. The Ampure™ beads were allowed to separate on a magnetic stand and were washed with two steps of 500µl 70% ethanol. The end-repaired, sheared gDNA was eluted from the beads in 27µl nuclease-free water. One microliter of this DNA was assessed on the Perkin Elmer™ 3K Labchip to determine the fragment size distribution of each sample. Lastly, platform-specific adaptors were ligated and samples barcoded using the SureSelect™ Target Enrichment System Protocol for Sequencing on Ion Proton, following the manufacturer's instructions and purified with Agencourt AMPure™ XP beads (Beckman-Coulter) as described in SureSelect™ Target Enrichment System Protocol for Sequencing on Ion Proton (Version C0, December 2016). The amplification reaction was set up in two sets of eight reactions. The reaction was performed across eight cycles of amplification on the SimplyAmp™ thermal (ThermoFisher™ Scientific) cycler. The amplified library was assessed on the Perkin Elmer™ 3K Labchip to determine the fragment size distribution and concentration of each sample.

*Hybridisation and Capture*

Exon capture was then performed using the Agilent SureSelect™ solution-based zebrafish exon-capture kit on the Ion PI™ Chip, following the manufacturer's protocol as described on p.39-42 of the SureSelect™ Target Enrichment System Protocol for Sequencing on Ion Proton (Version C0, December 2016). The entire volume of the amplified library for each of the 16 samples was evaporated at 37°C and reconstituted in 3.4µl nuclease-free water. The libraries were hybridised to the capture library at 65°C for 24hrs. The streptavidin-coated magnetic beads were then prepared and the hybridised DNA was captured using the streptavidin-coated beads.

*Post-Hybridisation Amplification and Sample Processing*

Post-capture amplification was performed using the SureSelect™ Target Enrichment System Protocol for Sequencing on IonProton™ (Version C0, December 2016). Specifically, 9 cycles of amplification were performed during the Post-capture Polymerase Chain Reaction (PCR). The libraries were purified and primer dimers reduced by repeating this part of the protocol for a total of two AMPure™ XP bead purifications. This was done using the method described on p. 47 of the SureSelect™ Target Enrichment System Protocol for Sequencing on Ion Proton (Version C0, December 2016). The purified, amplified libraries were assessed and quantified on the Bioanalyser DNA HS Chip and the templates prepared for sequencing using the Ion OneTouch 2 System with the Ion PI™ Template OT2 200 Kit v2 and Ion PI™ Sequencing 200 Kit v2.

*2.2.3) Assembly and analysis pipeline*

The raw reads of *A. japonicus* were aligned to the *Danio rerio* reference genome as a preliminary quality control (QC) step performed by the IonProton™ software post sequencing. This step was necessary to assess the level of similarity between the two species and the ability of the capture kit to sequence a non-target organism. Upon completing the preliminary alignment, the raw reads were obtained and their quality assessed using two different platforms: FASTQC v. 0.11.4 (Andrews, 2010) and CLC GWB Genomics Workbench® v7.0.3 (CLC GWB, Aarhus, Denmark). During assessment the reads were quality-filtered and trimmed to remove all adapters and artificial duplications as well as the first 9nt of the reads, which showed a nucleotide composition imbalance. The reads were then trimmed from the 5'-end for a minimum average Phred score of Q12 over a window of 3nt and only sequences with a minimum length of 50bp were retained. These alterations were performed using both TRIMMOMATIC v. 0.33 (Bolger et al., 2014) and CLC

Genomics Workbench®. The altered reads were visualised using FASTQC to ensure that the primers, barcodes, and adapter sequences had been sufficiently trimmed.

*De novo* assembly and analysis of the trimmed reads was done, using two programs, to ensure that the assembly was not biased by the use of a single algorithm: Velvet v. 0.7.62 (Zerbino and Birney., 2008) and CLC Genomics Workbench®. In order to accurately compare the assemblies, the parameters for each program were set as consistently as possible with the minimum contig length set at 1000bp, as to simplify mapping and downstream analyses of the contigs and a minimum of 8x coverage. Using VelvetOptimiser (Gladman and Seemannn, 2012) an optimal hash length of 69 was determined for the Velvet assembly, this step was not necessary for CLC GWB as the program determines the optimal hash length during assembly (Rahman and Pachter, 2013). All other parameters were kept at default for both programs. Upon completion of the assemblies, Bandage (Wick et al., 2015) was used to visualise the *de novo* assembly graphs of both CLC GWB and Velvet to identify and manually resolve ambiguities within the assemblies. Bandage was also used to compare the assemblies by assessing each of the assembly's individual quality scores which include the number of contigs, N50, average contig length, maximum contig length, minimum contig length, total length and median depth.

After curating the assembled scaffolds, the assemblies were compared to the National Centre for Biotechnology Information (NCBI) DNA database using the BLAST algorithm (Basic Local Alignment Search Tool) (Altschul et al., 1997). The BLASTn function was used to classify contigs according to the entry that had the highest hit score on GenBank. Default parameters were used and an expectation value (E-value) of less than 10e-5. These results were filtered to determine the most significant result using the BLASTn option, max target sequences, which assigns the most significant result to each contig; this was set to one, with the result determined as most significant being selected. The quality of the assembled contigs were assessed by aligning them against the Genbank non-redundant (NR) protein database using the BLASTx algorithm.

Functional annotation in the form of gene ontology (GO) was extracted from the NR database using Blast2GO v2.4.4 (Conesa et al., 2005) with an E-value threshold of 10e-10, as the matching of contigs to known proteins gives an indication of the quality of assembly, in assemblies where no reference sequences were available (Parchman et al., 2010). Blast2GO was then used to compare the consensus sequences of dusky kob against the reference genome of *Danio rerio* and the draft genome of *Larimichthys crocea* (yellow croaker), a species which had been identified in the initial BLASTn results and a known

sciaenid species. This was done to assess the similarity of the consensus sequence to the model organism, and to evaluate the method for assessing the accuracy and success of the exome kit's transferability. Potential exons and genes on the final scaffolds were predicted using CLC GWB, while gene prediction, for zebrafish and kob were performed using ENSEMBL Zv9 (www.ensembl.org).

*2.2.4) Putative Variant detection*

To investigate the utility of the exome sequences for potential type I molecular marker development, the sequences were mined for two classes of variants, single nucleotide polymorphisms (SNPs) and microsatellite repeats (also known as short sequence repeats, SSRs or short tandem repeats, STRs). For SNPs, the contigs were assessed to determine whether coverage was sufficient for accurate variant detection, this was achieved by filtering the contigs by excluding any contig that had a median depth of less than 8x, after mapping individual reads back to the contig consensus, used as reference sequence. Variant detection was performed on the filtered contigs across the genome by mapping the trimmed reads of sixteen samples back to the *de novo* assembly (used as the consensus, reference sequence) in CLC GWB. Upon completion of mapping, the fixed ploidy variant detection tool was run in CLC GWB to identify variants observed among the samples, with the required variant probability value being set at 0.9% and the ploidy at 2. The required variant probability is the minimum probability value required for the variant to be called, thus being stringent with this value should assist in the validity of the results obtained. This detection method relies on an error model estimation to identify and remove variants that are most likely the result of sequencing errors. Using a program called tandem repeat finder (Benson, 1999), the contigs were screed for repeats motifs (di- to hexanucleotide repeats) with a minimum of four contiguous repeat units.

## 2.3) Results

*2.3.1) Sequencing and capture efficiency*

The number of final library Ion Sphere Particles (ISP's) amounted to a total of 56% of the total reads. DNA enrichment and resequencing yielded more than 567 million sequence reads (Table 2.1), with an average of 35.5 million sequence reads per sample (range: 29.1M – 41.6M), with the amplified library inserts having an average size distribution of 240bp, with a concentration varying between 277 and 790ng per µl. The percentage of low-quality reads varied between 15% and 27% across eight PI v3 chips with a mean read length of 137bp.

**Table 2.1. Summary statistics of the reads and quality of the bases generated for dusky kob on the ion-torrent platform using the P1 chip in combination with the zebrafish exon capture kit**

| | |
|---|---|
| Total Reads | 567,409,399 |
| Total Usable Reads [Final Library] | 317,749,263 (56%) |
| Total ≥ Q20 [bp] | 63,504,893,735 (85%) |
| Total Bases [bp] | 74,667,689,870 |
| Mean Read Length [bp] | 137 (124-142) |
| Average Number of Reads Per Sample | 35,463,087 (29,1M – 41,6M) |
| Average Number of Bases Per Sample | 4,666,730,617 (4,1Bn - 5,8Bn) |
| Average ≥ Q20 Per Sample | 3,969,055,858 (3,1Bn– 5,0Bn) |

The preliminary mapping of *A. japonicus* reads to *D. rerio* genome indicated that approximately 67.58% of the sequenced reads were able to be aligned to the reference, with approximately 32.42% of the generated reads being unable to map (Figure 2.1).

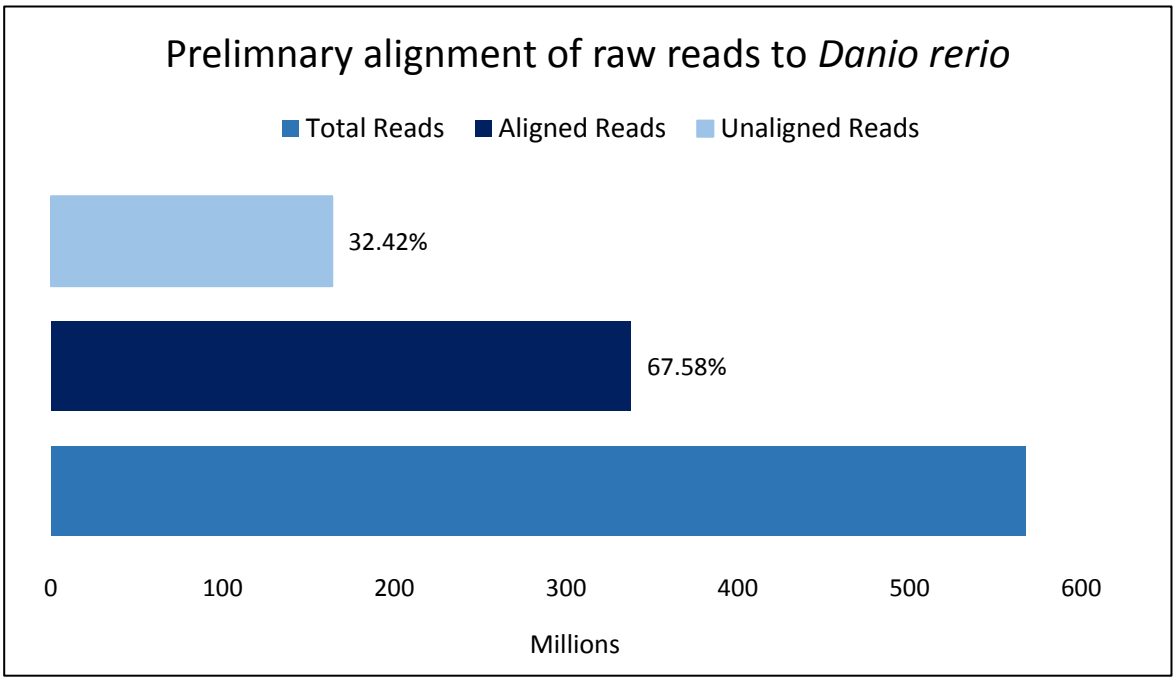



**Figure 2.1. Preliminary alignment of the raw reads of *A. japonicus* to the reference genome of *D. rerio* which was performed using the ion-torrent platform**

*2.3.2) Assembly and analyses*

After trimming the raw sequences, which included the removal of low quality and N-containing reads, a total of 112GB of clean reads were retained and used for *de novo* assembly. (Figure 2.2). For the Velvet assembly, 419,793,145 reads were retained following the quality control performed using both TRIMMOMATIC and FASTQC. Upon completion of QC, these reads were assembled into scaffolds, all of which were larger than 1,000 nt, with the largest scaffold being 3,169 nt in length (0.07% of the total length). For the second assembly performed by CLC GWB, a total of 499,014,555 reads were retained following the program's quality control step and then assembled into scaffolds, which all were larger than 1,000 nt, with the largest scaffold being 3,298 nt in length (0.23% of the total length) (Figure 2.2).

After filtering CLC GWB was found to have a higher number of contigs with a coverage of 8x or greater, with 3,831 and 7,940 contigs identified in Velvet and CLC GWB, respectively (Table 2.2). The BLASTn results showed that from the 7,940 contigs produced using CLC GWB a total of 25,893 hits were obtained while the 3,959 contigs produced using Velvet obtained a total of 17,018 hits (Table 2.2). After filtering these results in BLASTn by applying a stringent e-value of <10e-10 and the max targeted sequences option being set to 1, a total of 3,925 and 7,911, significant hits were identified in Velvet and CLC GWB, respectively. The overall median depth for both assemblies was low, Velvet achieving a depth of 5.74x and CLC GWB a depth of 6.23x. Of the 346,263 expected exons found within zebrafish only a small portion of these exon-homologues were recovered by each assembler; CLC GWB was able to recover 1.3% of the exons while the Velvet assembly only contained 0.6% of the expected zebrafish exon-homologues.

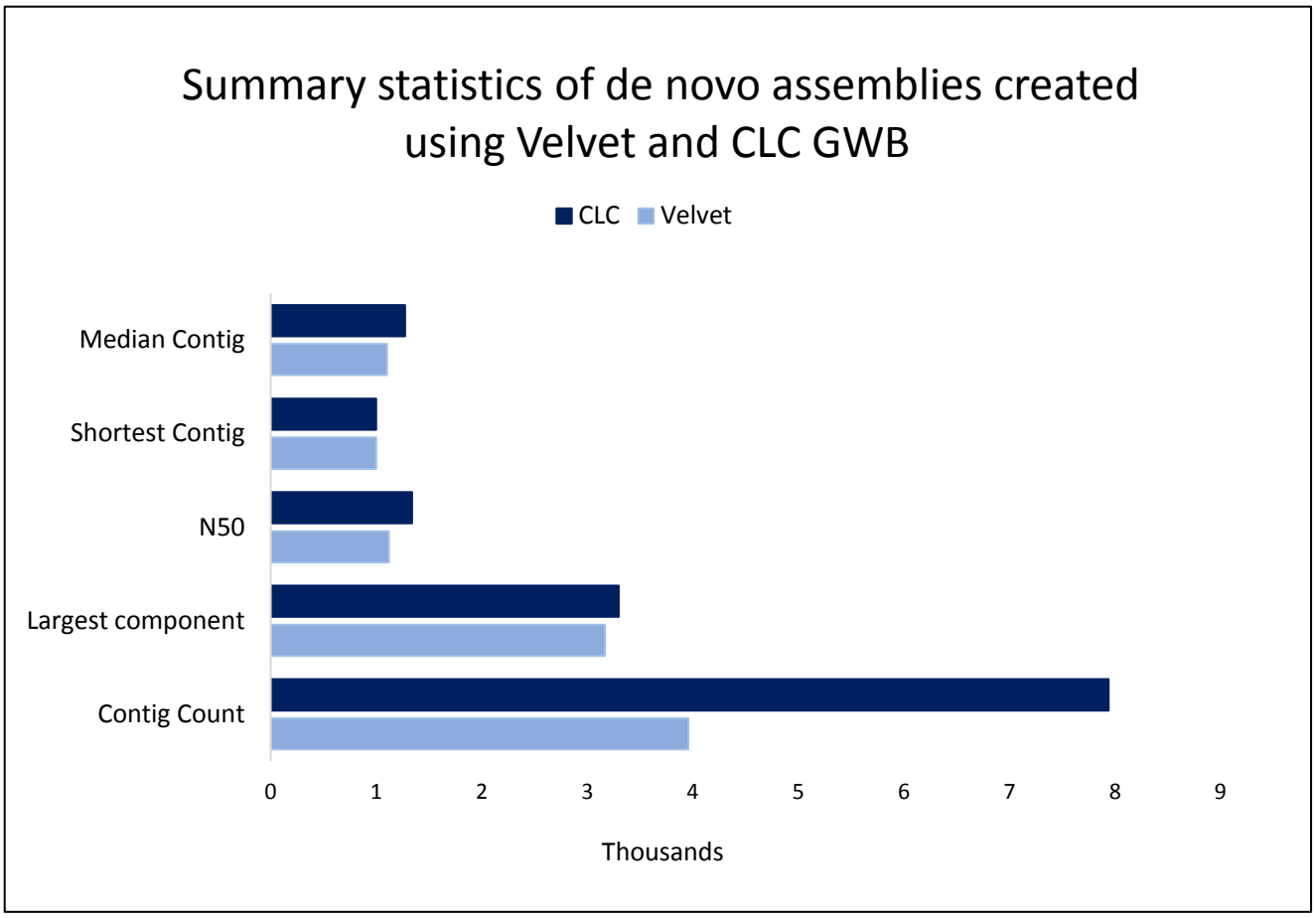



**Figure 2. 2. Comparison of results obtained from the *de novo* assemblies performed by CLC GWB and Velvet with main criteria: number of contigs, N50, average contig length, maximum contig length, minimum contig length and total length.**

**Table 2.2. BLASTn results for the contigs produced using CLC GWB and Velvet as well as the number of significant hits predicted to be *Larimichthys crocea*. Hits were regarded as significant when the E-value was <10e-10. The median depth of each assembly as well as the number of contigs determined as having a depth of ≥8x are included.**

| | Velvet | CLC |
|---|---|---|
| Number of contigs | 3,959 | 7,996 |
| Number of hits | 17,018 | 25,955 |
| Number of contigs having a depth of ≥8x | 3,959 | 7,940 |
| Number of hits for contigs having a depth of ≥8x | 17,018 | 25,893 |
| Number of significant hits | 3,925 | 7,911 |
| Significant hits to *Larimichthys crocea* | 3,494 (89%) | 7,120 (90%) |
| Median depth of the assembly (contig coverage) | 5.74x | 6.23x |
| Homologous Zebrafish exons detected (346,263) | 2,177 (0.6%) | 4,446 (1.3%) |

Of the total 7,911 significant hits obtained from the CLC GWB data, 7,120 were to *Larimichthys crocea*, which is ~90% of all significant hits. Similar results were obtained using the contigs generated by Velvet with ~89% of the hits being to *L. crocea*. Using Blast2GO the consensus sequence generated for dusky kob was compared to the reference genome of the zebrafish, GCF_000002035.6, and the draft genome of the large yellow croaker, GCF_000972845.2. Of the 7,940 contigs belonging to dusky kob, a total of 2,560 and 5,553 contigs had significant hits aligning to the zebrafish and the yellow croaker, respectively. (Table 2.3). The yellow croaker was found to have the highest overall similarity of 48.61%, while the zebrafish only had a 19.98% similarity to the dusky kob reference.

**Table 2.3. Results from Blast2GO assessing the similarity between *A. japonicus* contigs to the reference genome of *D. rerio* and the draft genome of *L. crocea***

| | Zebrafish | Yellow Croaker |
|---|---|---|
| Number of Significant Hits | 2,560 | 5,553 |
| Similarity to *A. japonicus* | 19.98% | 48.61% |

Using Blast2GO in combination with a stringent E-value threshold of 10e-10, 3,116 out of 7,940 contigs had a BLAST homologous match against the NR protein database. Of the 3,115 sequences with BLAST matches, 886 sequences were successfully annotated associating them with 4,168 GO terms. Of these 2,711 were assigned to the functional category 'Molecular Function' (65.04%), 1,085 to 'Cellular Component' (26.02%), and 189 to 'Biological Process' (4.53%) (Figure 2.3). The distribution of contigs in various functional classes of Gene Ontology (GO) and EuKaryotic Orthologous Groups (KOG) databases are shown in Figure 2.4

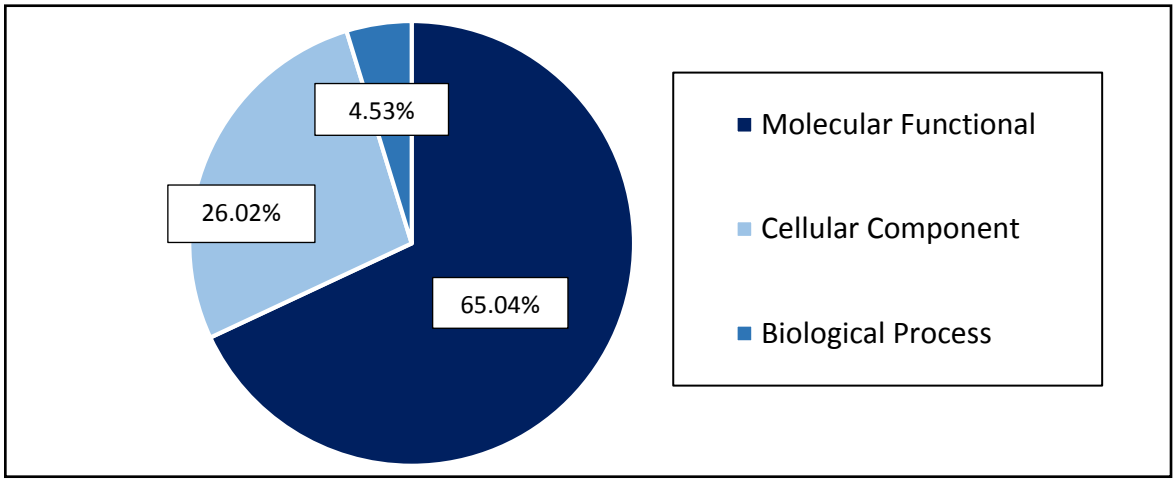



**Figure 2.3. The assignment of the *A. japonicus* contigs to the three subcategories (Molecular function, Cellular component, and Biological process) of the GO database. The main GO categories are represented with different colours**

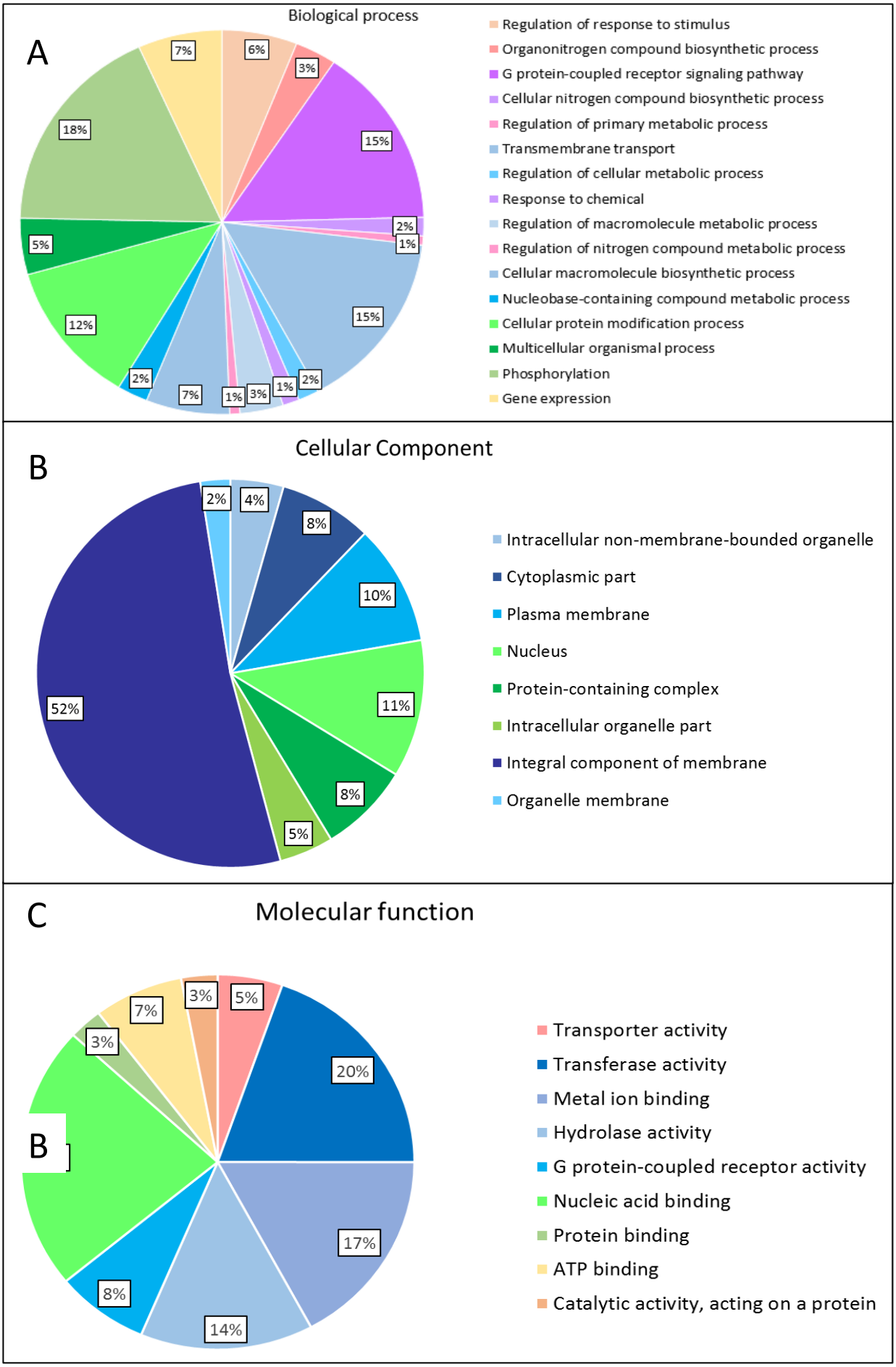



**Figure 2.4. The percentage distribution of *Argyrosomus japonicus* contigs to the 31 terms on the GO database within the 3 main subcategories (A) Biological Process, (B) Cellular Function and (C) Molecular Function**

*2.3.3) Variant detection*

A total of 4.5 million variants, of which, 2,840,198 were single nucleotide polymorphisms (SNPs) (Figure 2.5), were detected. Of the total number of putative SNPs, 386,266 were identified as being non-synonymous. A total of 31,525 transitions (A/G and C/T) and 22,415 transversions (A/C, A/T, C/G and G/T) were detected, with C/T being the most common (16,077) and A/C the least common (2,914) substitutions observed (Figure 3.3). This represented a transition/transversion (ts/tv) ratio of 1.407 (Table 2.4).

A total of 3,276 repeats with a minimum of four contiguous repeat units (motifs range from two to six) were identified using the 7,940 contigs generated by CLC GWB (Table 2.5). The most abundant tandem repeats were dinucleotide motifs (2,434), followed by tetranucleotide (560), trinucleotide (231), hexanucleotide (27) and pentanucleotide (22) motifs (Table 2.5). Among the dinucleotide motifs, the main repeats were the types AC (46.28%), AT (38.65%), AG (13.25%), and TG (1.82%) (Figure 2.6). Over all putative microsatellite loci, CA repeats had the highest frequency (~34%) of all sequence motifs in the *A. japonicus genome*, with the second most abundant microsatellite motif being TA, and in general, the dinucleotides were found to be the most abundant length motifs throughout the genome (Figure 2.7). It was determined that an informative marker was found at approximately every 500 nt, with a confirmed SNP at 1 in every 1000 nt. All polymorphic SNP loci showed two alleles and all of them agreed with those expected from the database information.

**Table 2.4. Summary statistics for the variants found in the exome data of *A. japonicus*.**

| | |
|---|---|
| Total number of variants | 4,546,626 |
| Putative SNPs | 2,840,198 |
| Non-synonymous SNPs | 386,266 |
| Identified tandem repeats | 3,276 |
| Transitions in feasible SNPs | 31,525 |
| Transversions in feasible SNPs | 22,415 |
| Transition-transversion ratio | 1.407 |

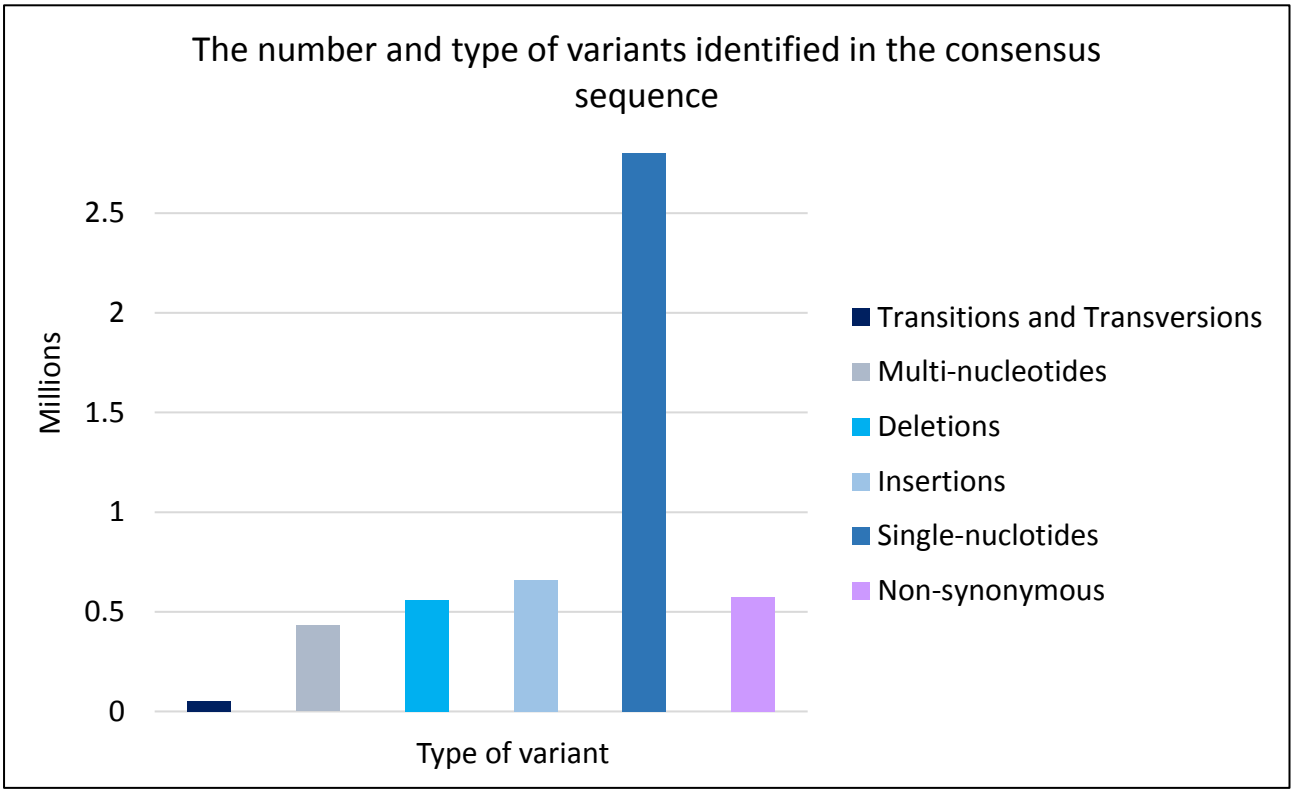



**Figure 2.5. The number and type of variants discovered in the consensus sequence of dusky kob using the fixed ploidy variant detection tool available in CLC Genomics Workbench. Variants included are: replacements, multi-nucleotides, deletions, insertions, single nucleotides and the number of these variants found to be non-synonymous**

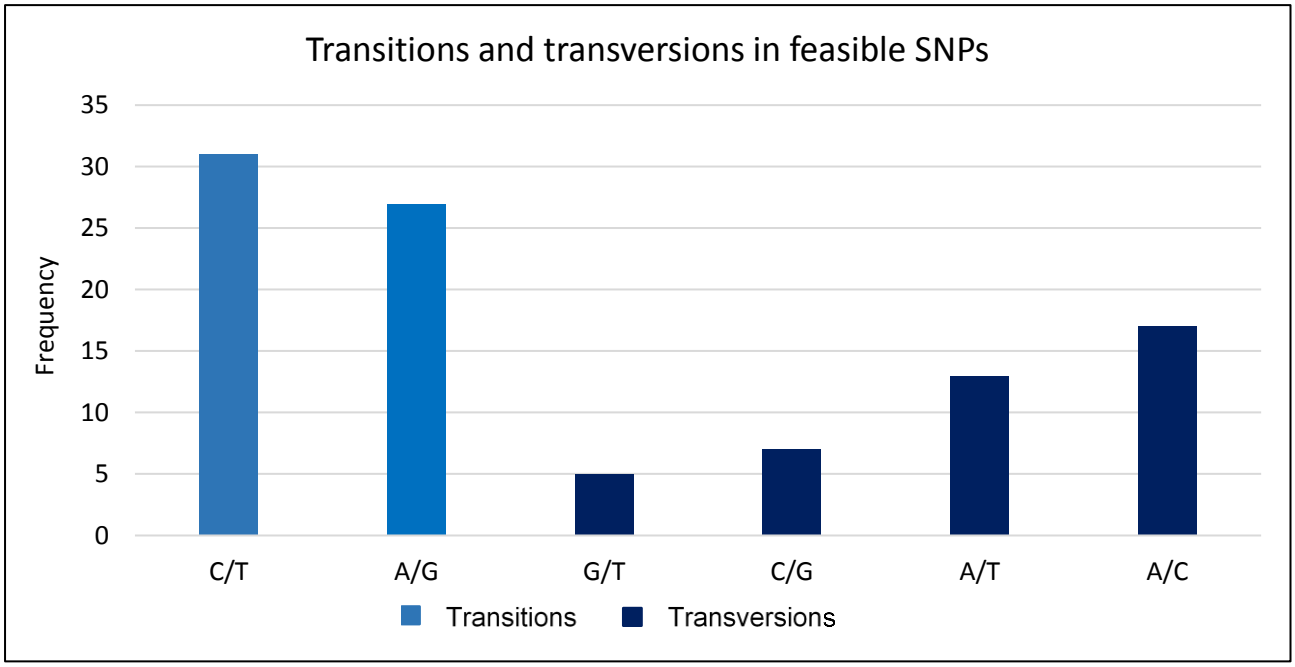



**Figure 2.6. Distribution of SNP variants analysed in this study using only feasible SNPs. Transitions (ts) and transversions (tv) are indicated in different colours with the frequency of each transition and transversion within the exome data indicated.**

**Table 2.5. Summary of the tandem repeats found in *A. japonicus* as well as the percentage of each repeat type found within the exome data**

| Tandem repeat | Count/percentage |
|---|---|
| Dinucleotides | 2,434 (74.3%) |
| Trinucleotides | 231 (7.1%) |
| Tetranucleotides | 560 (17.1%) |
| Pentanucleotide | 22 (0.67%) |
| Hexanucleotide | 27 (0.83%) |
| Total | 3,276 |

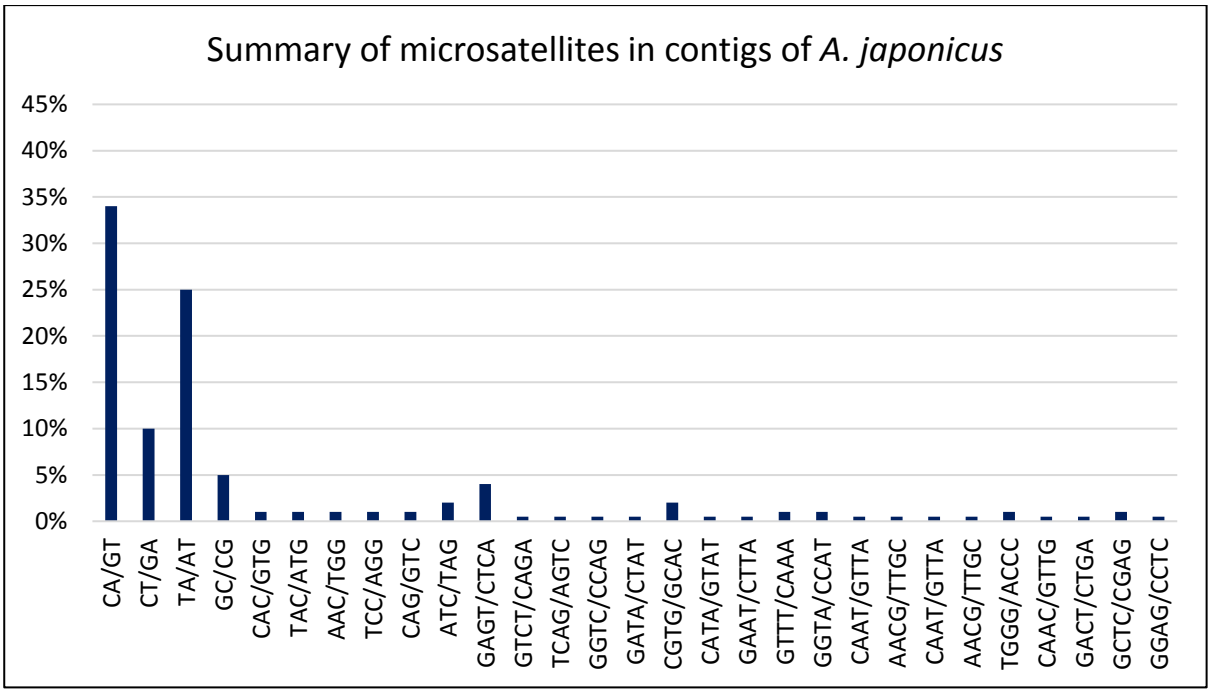



**Figure 2.7. The distribution of tandem repeat sequence motifs across the identified repeat regions in the contigs of A. japonicus from di- to tetra-nucleotides**

## 2.4) Discussion

Zebrafish and dusky kob both form part of a broad infraclass known as the teleosts, which is the largest infraclass in the class Actinopterygii, making up 96% of all existing species of fish. This infraclass originated in the Early Cretaceous period which occurred ~290 million years ago (mya). Throughout the years this class has diverged and evolved, resulting in the diverse number of species seen today. The model organism zebrafish is part of the order Cypriniformes, which was determined to originate approximately 250 mya, while dusky kob is a member of an order that originated much later in comparison, Perciformes, which first appeared and diversified in the Late Cretaceous period around 146 mya, more than 100 mya after zebrafish. Despite the divergence of the orders, there are still a number of genes

and functional clusters that have been observed as being highly conserved throughout the Teleosts (Dickmeis and Muller, 2004; Henzy et al., 2017; Paibomesai et al., 2010; Suzuki et al., 1999). Studies have shown that whole exome sequencing of related species has a decline in performance when faced with even limited divergence between species (Jones and Good, 2016; Vallender, 2011). However, success has also been reported in closely (*e.g*. Jia et al., 2013; Ryan et al., 2013) and even for more divergent species (*e.g.* Faircloth et al., 2012; Lemmon et al., 2012; McCormack et al., 2012), at a cost of losing comparative data, which could have been used for evolutionary and functional studies (Schott et al., 2017). Therefore, the use of the zebrafish kit in kob was tested for efficiency and as a method for developing genomic resources for resource scarce species.

Library construction was performed utilising the exome capture kit, however during construction it was determined that the size selection estimations performed using the E-gel system or PippenPrep for purification, would result in the loss of up to 90% of the library yield. Library inserts used for exome sequencing are usually fragmented to a size of between 200-700bp, however due to the potential yield loss and the fact that the libraries yielded fragments that were within the recommended sequencing range (~240bp) of the IonProton™ platform, it was decided to sequence the 16 exome libraries without a size selection step. This decision, theoretically should not have negatively impacted the sequencing of the exome as a study was performed by Krasnenko et al. (2018) which showed that DNA fragments ranging between 250–330 bp demonstrated the highest enrichment efficiency. Showing that size selection is an important step for effective enrichment and subsequent sequencing but is greatly dependant on the species. The study performed by Krasnenko et al. (2018) was however based on human exome data and the size of the exome regions between fish species is highly variable (Henkel et al., 2012). Without this step, DNA enrichment and resequencing yielded sequences within the expected range for an average ion-torrent run which produces a maximum of 60 million reads per sample and a maximum read length of 200bps (Table 2.1).

Preliminary mapping of the raw kob reads to the zebrafish reference genome was successful, however there were a significant amount of sequences that could not be mapped, which could have occurred due to a number of reasons. Mapping could have failed as a result of poor-quality sequences being generated, alternatively the sequences failed to map as they were specific to the target species, dusky kob. The generation of sequences that are not found to occur within the model organism could be indicative to the randomness of the capture kit as a result of divergence between the species. As the generation of non-

targeted sequences would have been a result of the region-specific probes binding at non-specified locations of the genome, which is most likely a consequence of divergence between the species, which inhibited the ability of the probes to bind to the specified location. Despite the apparent divergence between the species, the kit was still able to successfully produce a number of sequences, with the BLASTn results of the contigs produced by the assembly programs, shown to have approximately 89.5% of their significant hits belonging to *Larimichthys crocea* (Table 2.2). This was not an unexpected result as the *Larimichthys crocea*, commonly known as the large yellow croaker is a marine fish in the Sciaenidae family, making this finfish a closer relative to kob than that of zebrafish (Gui, 2018). However, what was unexpected, is that despite the lack of similarity between zebrafish and kob, the exome capture kit was still able to capture a large amount of gene sequences that appear to be specific to kob, thus explaining the high level of similarity seen with the yellow croaker, demonstrating the kits lack of bias towards genes that are found to be highly conserved between kob and zebrafish (Table 2.3). This was further validated by the low coverage seen throughout the contigs as it is most likely the result of probes being able to bind at non-targeted regions across the genome, causing the probe placement to differ between each individual, thus resulting in a large number of sequences being generated in one sample that has not been generated in another and *vice versa.* However, non-uniform coverage across the exome is not unprecedented; this is one of the main limitations of standard whole-exome sequencing methods, which can be predominantly overcome with the development of species-specific capture kits (Wang et al., 2017). The fact that this problem is seen in this study when using the model organism's capture kit indicates that the level of divergence between the species is a more prominent limitation than initially anticipated.

Inspection of the *de novo* assemblies revealed individual strengths and weaknesses for each assembly program (Figure 2.2). The results obtained for the assemblies were comparable with one another, with CLC GWB aligning a higher number of contigs with a greater median depth. Although CLC GWB was able to achieve a higher median depth than that of the Velvet assembly, it was still considerably low in comparison to what was expected. Using 16 samples to generate the exome data a minimum coverage of 16x was to be expected, as theoretically the capture kit should be binding to the same locations within each sample, thus resulting in at least 16x coverage of every coding region sequenced. A study performed by Nielsen et al. (2011) determined that despite previous sequencing methods requiring target regions to have a coverage of greater than 20x for reducing uncertainty associated with genotype and SNP calling, newer technology and the demand for larger

samples have suggested that medium (5–20x) or low-coverage is still able to accurately detect at least 70% of the true variants (Spence et al., 2019). For this reason, the contigs were assessed to determine which of the assemblies were able to achieve a higher number of contigs with a median depth of 8x or greater. The Velvet assembly was assessed and although this assembly did have a number of its contigs having a higher coverage of 10x or greater, there were not enough to reliably continue with variant calling and therefore the *de novo* assembly generated using CLC GWB, was used for downstream analyses. Similar results have been seen in previous studies, where assemblies generated using both CLC GWB and Velvet were found to differ in reference to number of contigs and median depth, which is most likely the result of the more stringent assembly parameters of the Velvet program (Ghangal et al., 2013; Kotwal et al., 2016). In general, it has been stated that a better assembler would be expected to produce a large number of contigs with a high coverage that are able to attain significant BLAST hits (Ashrafi et al., 2012; Jiang et al., 2016). Even though the contigs from each assembler were able to return a large number of significant hits, the sheer number of CLC GWB produced contigs in combination with a slightly higher significant hit rate resulted in CLC GWB BLAST results being far superior to that of Velvet, making it a better assembler for this specific dataset (Table 2.2). This poor performance by Velvet for the Ion Torrent assembly, is a result of Velvet's assembly methodology, which was designed to be used for Illumina sequences, (Zerbino, 2008) unlike CLC GWB which has been designed to run using the output data produced by multiple platforms.

The coding genes of *D. rerio* has been described (26,247 protein-coding genes and 346,263 exons) as making up 3.8% of the genome (Howe et al., 2013), but genomic studies on *A. japonicus*, are still in the early stages, and the percentage of its coding genes is still unknown. Since only a small portion of the zebrafish exomes were able to be sequenced (1.8%), this result could indicate to the fact that dusky kob has a significantly lower number of noncoding sequences than zebrafish, such as is the case with common carp, whom has a noncoding region that is approximately 1.7 times smaller than that of zebrafish (Henkel et al., 2012). This could be an indication that the genome size of dusky kob may be smaller than predicted, not being much larger than that of zebrafish. However, this low number could also be the result of divergence between the species, which resulted in a number of non-exon regions being sequenced such as introns and untranslated regions (UTRs). However, with the kob genome still not sequenced, the gene predictions are purely based on the zebrafish data and cannot serve as a primary guide as it may not be accurate and needs to

be further investigated. This, together with the lack of genomic information in other kob species (*e.g. Argyrosomus inodorus*), it is difficult to predict how deep the *de novo* assembled exome sequences cover the dusky kob's genome. Therefore, by characterising the consensus sequence, it will give an indication of how well the exome regions were presented by this sequence. Characterisation determined that the contigs were distributed among the various functional classes of GO and KOG databases indicating how the exome data, even though not covering the entire genome, encompasses a broad gene diversity (Figure 2.3 and 2.4), even though only a small number of the contigs were able to be annotated. The limited number of annotated contigs is most likely due to the fact that a number of sequences which were generated using the capture kit were not part of the protein-coding regions and as a result could not be annotated. Annotation of the contigs was then limited even further by the limited genomic resources available for the genus *Argyrosomus*, which could have led to a number of proteins remaining unidentified as a result of being species-specific. This result, however, was not directly comparable to other studies where the *de novo* assembly of non-model species was performed using high-throughput sequencing (Carruthers et al., 2018; Yasuike et al., 2018; Lin et al., 2019). Previous studies, were able to detect a far higher number of GO annotations (~20,000 – 40,000). This is largely due to the fact that these studies although not performed in the specific species under investigation, had genomic information regarding its close relatives and as a result most of the genes for the species had already been characterised allowing for a more informative result. A large number of the annotations were assigned to molecular function rather than biological processes which is generally seen as the most well represented KOG category in other studies including that of zebrafish, which could be a potential indictor to the lack of exome regions sequenced (Figure 2.3) (Harney et al., 2016). Functional annotation against the GO database rendered 889 sequences that had BLAST matches with an E-value less than $10e^{-10}$ (Figure 2.4). These were uniquely assigned to the three main GO classes and each of their sub-categories were populated by at least eight different classes. The main represented sub-category was biological process with a total of 15 different classes. The results of functional annotation showed that the sequences of dusky kob had a low proportion of annotated genes when compared to the database of zebrafish. Yet despite the low number of annotated contigs, the contigs that were able to be annotated did allowed identification of genomic regions responsible for development and metabolic processes, gene expression, and regions involved in processes of stress response. Subsequently, this data can be used as the basis for further biological studies in other areas of aquaculture or future breeding programmes in dusky kob.

In addition, tandem repeats were identified (3,276) through exome sequencing, although the number identified was significantly lower in comparison to SNPs, this is not unexpected as these repeat regions are normally found to be associated with non-coding DNA and their discovery in the coding regions of dusky kob is quite interesting. Although this is most likely evidence to support the fact that a number of the generated sequences were not exome regions but rather a result of the kits ability to sequence random regions of the genome as a consequence of divergence, whole genome sequencing has shown evidence for the prevalence of trinucleotides repeats in coding regions. A study performed the human genome determined that 17% of the genes contained in open reading frames, with similar results being seen in studies performed by Jansen et al. (2012) and Gemayel et al. (2010). Microsatellites, are found to be enriched in regulatory genes that encode for transcription factors, DNA-RNA binding proteins, and chromatin modifiers (Young et al., 2000). The use of gene-associated markers has become an important part of constructing genetic maps (Shin et al., 2012) because, based on comparative genomics, the use of a fish genome that has already been sequenced, it is possible to predicted the location of studied loci. However, given the lack of trinucleotide repeats identified within the sequences of *A. japonicus*, it is more likely that the majority of the tandem repeats identified are a result of non-coding regions being sequenced. This is because trinucleotides are found to be the most common repeat motif located within the coding regions of organisms as these repeats are often associated with inherited disorders, which result in changes to the DNA (Tan et al., 2012; Almedia et al., 2013; Li et al., 2017). Thus, their limited presence within the identified repeat motifs only further supports the notion that the sequences generated using the capture kit did not entirely consist of exome sequences, as would be expected. Besides the lack of trinucleotides, the distribution of the repeat motifs was found to be consistent with what is to be expected from most higher eukaryotes (Brooker et al., 1994; Li et al., 2017; Srivastava et al., 2019), such as the prevalence of dinucleotides. Of the dinucleotides, the CA repeats were identified as having the highest frequency of all the identified tandem repeats, as they demonstrate an overrepresentation of this repeat motif. The TA motifs were found to be the second most abundant dinucleotide following CA which complies with previous studies performed in teleost's (Almedia et al., 2013 Li et al., 2017; Tan et al., 2012; Yang et al., 2008) however this was not true for the low frequency of CG and apparent absence of CCG/CGG repeats, which significantly differed from these studies. The CpG-like motifs are usually frequently found in vertebrates as a means to regulate gene activity, these motifs are found to be located in the 5'-UTR where they serve as protein binding sites, that are regulated by DNA-methylation. This lack of CpG-like motifs is more commonly seen in

invertebrates, as they do not regulate gene activity by means of CpG islands (Rhode and Roodt-Wilding, 2011; Toth et al., 2000). This low frequency which is observed is therefore more likely to have occurred as a result of the limited sequence data created by the bias during the hybridisation step of sequencing, rather than being a true representative of the dusky kob sequence data and therefore requires further investigation.

A total of ~2.8 million SNPs were identified across the gene regions with a total of 386,266 variants determined to be non-synonymous meaning these variants are found to have a direct effect on the amino acid, which can alter a phenotype or affect a trait of interest making these variants extremely useful in a variety of applications. The total number of SNPs identified across the genome from exome data alone is comparable to that of model organisms where whole genome sequencing detects roughly two to five million SNPs across the genome with a SNP occurring once in every 500 nt (Xiao et al., 2016). Following quality control SNPs were confirmed by setting the MAF to greater than 0.5, which resulted in the identification of 1,931,334 putative SNPs. Using these SNPs, the transition to transversion (ts/tv) ratio was calculated, which is typically reported as 2 in other fish species, with protein coding regions often even higher as the transversions in the protein coding regions are most likely to change the encoded amino acid. The dusky kob exome data represented a ts/tv ratio of 1.407 (Table 3.3; Figure 3.3). This ratio was lower than that observed by Vera et al., (2011) (1.885) and *in silico* (1.456) by Pardo et al., (2018), but it was very similar to that described for common carp (*Cyprinus carpio*) (1.310) (Zhu et al., 2012), gilthead seabream (*Sparus aurata*) (1.375) (Cenadelli et al., 2011) and turbot (*Scophthalmus maximus*) (1.354) (Vera et al., 2013). However, in other fishes, such as chum salmon (0.95), and sockeye salmon (0.98) (Smith et al., 2005), as well as zebrafish (1.20) (Stickney et al., 2002) the ts/tv ratios were significantly lower. The discrepancy seen in ts/tv ratios among the different species may suggest a biased codon usage or substitution rate as a result of fishes from different phylogenetic units being subjected to different selection pressures (Zhu et al., 2012). However, the ts/tv ratio could have also been influenced by fact that the sequences obtained from the ion-torrent platform most likely did not consist entirely of exome regions, thus skewing the results making any comparisons drawn between this result and previous studies a futile exercise.

Overall, the ion-torrent data proved to be a huge resource for discovering variants across the genome despite the low coverage, with studies demonstrating the use of large variant data in the identification of genetic markers, which can be correlated with production traits of interest, particularly for growth. In the fish *Sparus aurata*, a dinucleotide microsatellite in

the growth hormone gene (GH) is linked with faster growth rate, which can be used for breeding management and genetic selection for this trait (Almuly et al., 2005). In other fish species, such as rainbow trout and *Salvelinus fontinalis* (brook charr) (Salem et al., 2012; Sauvage et al., 2012), SNPs and QTL have also been reported to have association with traits of interest (growth and stress response). In general, the lack of knowledge regarding the genetic variation of stocks can cause inbreeding, leading to the fixation of deleterious genes, reduced growth rates, disease resistance problems and hinder the ability of the fish to adapt to new environments (Arkush et al., 2002; Gallardo et al., 2004; Hillen et al., 2017; Neira et al., 2006). Therefore, in addition to identification of QTL, which can assist in the selection of genotypes linked to traits of interest, by marker assisted selection (MAS), studies of microsatellites and SNPs are important for genetic monitoring, supporting dusky kob aquaculture and increasing its productivity.

## 2.5) Conclusions

This study was able to produce data that was more than sufficient, showing that exon capture can be used for genome-wide SNP discovery in non-model organisms, where there is limited genomic resources available. This method, unlike transcriptome sequencing, is able to sequence a large number of individuals while still being able to avoid ascertainment bias in subsequent population genetics analyses due to differential gene expression. Demonstrating it to be an affordable and reliable method, which can aid in the discovery of thousands of SNPs and in some cases a few thousand tandem repeats, which will greatly assist in marker development for the species. Overall, the cross-species targeted capture method used in this study was proven to be successful despite the non-uniform coverage across the genome. However, the unintended sequencing of non-protein coding regions, will have a significant impact on the reproducibility of this study, making the consolidation of data between projects extremely difficult. These singleton and non-coding sequences do require further investigation into their nature as these sequences could contain valuable information that is specific to dusky kob and therefore information obtained regarding these sequences' can be applied to aid not only in marker development but also in future genomics research for the species. This study was able to successfully capture exome regions which assisted in the identification of ~4.5 million of variants, showing the benefits of a cross-species sequencing approach in the development of markers which in turn could result in the acceleration of molecular breeding programmes in species where genomic resources are limited or not available.

## References

Albert, T., Molla, M., Muzny, D., Nazareth, L., Wheeler, D., Song, X., Richmond, T., Middle, C., Rodesch, M., Packard, C., Weinstock, G., Gibbs, R., 2007. Direct selection of human genomic loci by microarray hybridization. Nature Methods, 4(11), 903-905. https://doi.org/10.1038/NMETH1111

Almeida, B., Fernandes, S., Abreu, I., Macedo-Ribeiro, S., 2013. Trinucleotide Repeats: A Structural Perspective. Frontiers in Neurology, 4. https://dx.doi.org/10.3389%2Ffneur.2013.00076

Almuly R., Poleg-Danin Y., Gorshkov S., Gorshkova G., Rapoport B., Soller M., 2005. Characterization of the 5′ flanking region of the growth hormone gene of the marine teleost, gilthead sea bream *Sparus aurata*: analysis of a polymorphic microsatellite in the proximal promoter. Fish. Sci. 71, 479–490. https://doi.org/10.1111/j.1444-2906.2005.00991.x

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Andrews, S., 2010. FASTQC: a quality control tool for high throughput sequence data. Retrieved from http://www.bioinformatics.babraham. ac.uk/projects/fastqc

Arkush, K. D., Giese, A. R., Mendonca, H. L., McBride, A. M., Marty, G. D., Hedrick, P. W., 2002. Resistance to three pathogens in the endangered winter-run chinook salmon (*Oncorhynchus tshawytscha*): effects of inbreeding and major histocompatibility complex genotypes. Can. J. Fish. Aquat. Sci. 59,966–975. https://doi.org/10.1139/f02-066

Ashrafi, H., Hill, T., Stoffel, K., Kozik, A., Yao, J., Chin-wo, S.R., Van Deynze, A., 2012. *De novo* assembly of the pepper transcriptome (*Capsicum annuum*): a benchmark for in silico discovery of SNPs, SSRs and candidate genes 1–15.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Genome analysis Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Brooker, A. L., Cook, D., Bentzen, P., Wright, J. M., Doyle, R. W., 1994. Organization of Microsatellites Differs between Mammals and Cold-water Teleost Fishes. Canadian

Journal of Fisheries and Aquatic Sciences, 51(9), 1959–1966. https://doi.org/10.1139/f94-198

Burbano, H.A., Hodges, E., Green, R.E., Briggs, A.W., Krause, J., Meyer, M., Good, J.M., Maricic, T., Johnson, P.L.F., Xuan, Z., Rooks, M., Bhattacharjee, A., Brizuela, L., Albert, F.W., De La Rasilla, M., Fortea, J., Rosas, A., Lachmann, M., Hannon, G.J., Pääbo, S., 2010. Targeted investigation of the neandertal genome by array-based sequence capture. Science 328, 723–725.

Carruthers, M., Yurchenko, A., Augley, J., Adams, C., Herzyk, P., Elmer, K., 2018. Correction to: De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species. BMC Genomics, 19(1), 32. https://dx.doi.org/10.1186%2Fs12864-017-4379-x

Conesa, A., Götz, S., García-gómez, J.M., Terol, J., Talón, M., Genómica, D., Valenciano, I., Agrarias, D.I., Valencia, U.P. De, 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.

Cosart, T., Beja-Pereira, A., Chen, S., Ng, S.B., Shendure, J., Luikart, G., 2011. Exome-wide DNA capture and high throughput sequencing in domestic and wild species. BMC Genomics 12, 347.

De Donato, M., Peters, S.O., Mitchell, S.E., Hussain, T., Imumorin, I.G., 2013. Genotyping-by-Sequencing (GBS): A Novel, Efficient and Cost-Effective Genotyping Method for Cattle Using Next-Generation Sequencing. PLoS One 8(5): e62137.

Dickmeis, T., Müller, F., 2004. The identification and functional characterisation of conserved regulatory elements in developmental genes. Briefing in Fun. Geno. and Prot. 3(4), 332-350.

Faircloth, B., McCormack, J., Crawford, N., Harvey, M., Brumfield, R., Glenn, T., 2012. Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. Systematic Biology, 61(5), 717-726. https://doi.org/10.1093/sysbio/sys004

Gallardo, J. A., Garcia, X., Lhorente, J. P., Neira, R., 2004. Inbreeding and inbreeding depression of female reproductive traits in two populations of Coho salmon selected using BLUP predictors of breeding values. Aquaculture 234, 111–122. https://doi.org/10.1016/j.aquaculture.2004.01.009

Gemayel, R., Vinces, M.D., Legendre, M., Verstrepen, K.J., 2010. Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. Annu. Rev. Genet. 44:445–477. doi: 10.1146/annurev-genet-072610-155046

Gladman, S. and Seemann T., VelvetOptimiser. Retrieved from http://bioinformatics.net.au/software.velvetoptimiser.shtml

Ghangal, R., Chaudhary, S., Jain, M., Purty, R.S., Sharma, P.C., 2013. Optimization of *De Novo* Short Read Assembly of Seabuckthorn (*Hippophae rhamnoides L.*) Transcriptome 8, 1–7. https://doi.org/10.1371/journal.pone.0072516

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell,T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D.B., Lander, E.S., Nusbaum, C., 2009. Solution Hybrid Selection with Ultra-long Oligonucleotides for Massively Parallel Targeted Sequencing. Nat Biotechnol. 27(2): 182–189. https://dx.doi.org/10.1038%2Fnbt.1523

Gui, J., 2018. Aquaculture of the Large Yellow Croaker Aquaculture of the Large Yellow Croaker. https://doi.org/10.1002/9781119120759.ch3

Han, Z., Xiao, S., Li, W., Ye, K., 2018. The identification of growth, immune related genes and marker discovery through transcriptome in the yellow drum (*Nibea albiflora*) 1–27.

Harney, E., Dubief, B., Boudry, P., Basuyaux, O., Schilhabel, M., Huchette, S., Paillard, C., Nunes F., 2019. *De novo* assembly and annotation of the European abalone *Haliotis tuberculata* transcriptome.

Henkel, C.V, Dirks, R.P., Jansen, H.J., Forlenza, M., Wiegertjes, G.F., Howe, k., van den Thillart, G.E.E.J.M., Spaink, H.P., 2012. Comparison of the Exomes of Common Carp (*Cyprinus carpio*) and Zebrafish (*Danio rerio*). Zebrafish. 9(2), 59-67. https://doi.org/10.1089/zeb.2012.0773

Henzy, J., Gifford, R., Kenaley, C., Johnson, W., 2016. An Intact Retroviral Gene Conserved in Spiny-Rayed Fishes for over 100 My. Mol Biol Evol.1;34(3):634-639. https://doi.org/10.1093/molbev/msw262.

Hillen, J. E. J., Coscia, I., Vandeputte, M., Herten, K., Hellemans, B., Maroso, F., 2017. Estimates of genetic variability and inbreeding in experimentally selected

populations of European sea bass. Aquaculture 479, 742–749. https//.doi.org/10.1016/j.aquaculture.2017.07.012

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., McCombie, W.R., 2007. Genome-wide in situ exon capture for selective resequencing. Nat Genet 39:1522–1527.

Howe, K., Clark, M., Torroja, C., Torrance, J., Berthelot, C., Muffato, M., Collins, J., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J., Koch, R. et al., 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature, 496(7446), 498-503. https://doi.org/10.1038/nature12111

Jansen, A., Gemayel, R., Verstrepen, K.J., 2012. Unstable microsatellite repeats facilitate rapid evolution of coding and regulatory sequences. Genome Dyn. 7:108–125. doi: 10.1159/000337121.

Jia, X., Zhang, F., Bai, J., Gao, L., Zhang, X., Sun, H., Sun, D., Guan, R., Sun, W., Xu, L., Yue, Z., Yu, Y., Fu, S., 2013. Combinational analysis of linkage and exome sequencing identifies the causative mutation in a Chinese family with congenital cataract. BMC Med. Genet. 14, 107. https://doi.org/10.1186/1471-2350-14-107

Jiang, Z., Wang, H., Michal, J.J., Zhou, X., Liu, B., Woods, L.C.S., Fuchs, R.A., 2016. Genome wide sampling sequencing for SNP genotyping: Methods, challenges and future development. Int. J. Biol. Sci. 12, 100–108.

Jones, M., Good, J., 2015. Targeted capture in evolutionary and ecological genomics. Molecular Ecology, 25(1), 185-202. https://doi.org/10.1111/mec.13304

Krasnenko, A., Tsukanov, K., Stetsenko, I., Klimchuk, O., Plotnikov, N., Surkova, E., Ilinsky, V., 2018. Effect of DNA insert length on whole-exome sequencing enrichment efficiency: an observational study. Advances in Genomics and Genetics. 8, 13-15.

Korte, A., Farlow, A., 2013. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods, 9(1), 29. https://doi.org/10.1186/1746-4811-9-29

Kotwal, S., Kaul, S., Sharma, P., Gupta, M., Shankar, R., Jain, M., Dhar, M.K., 2016. De novo transcriptome analysis of medicinally important *plantago ovata* using RNA-seq. PLoS One 11. https://doi.org/10.1371/journal.pone.0150273

Lemmon, A., Emme, S., Lemmon, E., 2012. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Systematic Biology, 61(5),727-744. https://doi.org/10.1093/sysbio/sys049

Li, Z., Chen, F., Huang, C., Zheng, W., Yu, C., Cheng, H, Zhou, R., 2017. Genome-wide mapping and characterization of microsatellites in the swamp eel genome. Scientific Reports, 7(1). https://doi.org/10.1038/s41598-017-03330-7

Lin, G., Thevasagayam, N., Wan, Z., Ye, B., Yue, G., 2019. Transcriptome Analysis Identified Genes for Growth and Omega-3/-6 Ratio in Saline Tilapia. Frontiers in Genetics, 10. https://doi.org/10.3389/fgene.2019.00244

McCormick, M., Delaney, J., Tsuchiya, M., Tsuchiyama, S., Shemorry, A., Sim, S., Chou, A., Ahmed, U., Carr, D., Murakami, C., Schleit, J., Sutphin, G., Wasko, B., Bennett, C., Wang, A., Olsen, B., Beyer, R., Bammler, T., Prunkard, D., Johnson, S., Pennypacker, J., An, E., Anies, A., Castanza, A., Choi, E., et al., 2015. A Comprehensive Analysis of Replicative Lifespan in 4,698 Single-Gene Deletion Strains Uncovers Conserved Mechanisms of Aging. Cell Metabolism, 22(5), 895-906. https://doi.org/10.1016/j.cmet.2015.09.008

Neira, R., Díaz, N., Gall, G., Gallardo, J., Lhorente, J., Manterola, R., 2006. Genetic improvement in Coho salmon (*Oncorhynchus kisutch*). I: Selection response and inbreeding depression on harvest weight. Aquaculture, 257(1-4), 9-17. https://doi.org/10.1016/j.aquaculture.2006.03.002

Nielsen, R., Paul, J., Albrechtsen, A., Song, Y., 2011. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics. 12(6), 443-451.

Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J., Zwick, M.E., 2007. Microarray-based genomic selection for high-throughput resequencing. Nat Methods 4:907–909

Ozsolak, F., Milos, P., 2010. RNA sequencing: advances, challenges and opportunities. Nature Reviews Genetics, 12(2), 87-98. https://doi: 10.1038/nrg2934

Paibomesai, M., Moghadam, H., Ferguson, M., Danzmann, R., 2010. Clock genes and their genomic distributions in three species of salmonid fishes: Associations with genes regulating sexual maturation and cell cycling. BMC Research Notes, 3(1), 215. https://doi.org/10.1186/1756-0500-3-215

Parchman, T.L., Geist, K.S., Grahnen, J.A., Benkman, C.W., Buerkle., 2010. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics. 11:180.

Porreca G. J., Zhang K., Li J. B., Xie B., Austin D., Vassallo S. L., 2007. Multiplex amplification of large sets of human exons. Nat. Methods 4, 931–936. https://doi.org/10.1038/nmeth1110

Prado, F., Vera, M., Hermida, M., Blanco, A., Bouza, C., Maes, G., Volckaert, F., Aquatrace, C., Martínez, P., 2018. Tracing the genetic impact of farmed turbot *Scophthalmus maximus* on wild populations. Aquaculture Environment Interactions, 10, 447-463. https://doi.org/10.3354/aei00282

Rahman, A., Pachter, L., 2013. CGAL: computing genome assembly likelihoods. Genome Biology, 14(1), 8. https://doi.org/10.1186/gb-2013-14-1-r8

Rhode, C. and Roodt-Wilding, R., 2011. Bioinformatic Survey of Haliotis midae Microsatellites Reveals a Non-Random Distribution of Repeat Motifs. The Biological Bulletin, 221(2),147-154.https://doi.org/10.1086/bblv221n2p147

Ryan, S., Willer, J., Marjoram, L., Bagwell, J., Mankiewicz, J., Leshchiner, I., Goessling, W., Bagnat, M., Katsanis, N., 2013. Rapid identification of kidney cyst mutations by whole exome sequencing in zebrafish. Development 140, 4445–4451.

Saghai-Maroof, M. A., Soliman, K.M., Jorgensen, R. A., Allard, R.W., 1984. Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. U. S. A. 81, 8014–8018.

Saker, M.L., Griffiths, D.J., 2000. The effect of temperature on growth and cylindrospermopsin content of seven isolates of *Cylindrospermopsis raciborskii* (*Nostocales, Cyanophyceae*) from water bodies in northern Australia. Phycologia 39, 349–354.

Salem, M., Vallejo, R., Leeds, T., Palti, Y., Liu, S., Sabbagh, A., Rexroad, C., Yao, J., 2012. RNA-Seq Identifies SNP Markers for Growth Traits in Rainbow Trout. PLoS ONE, 7(5), p.e36264. https://doi.org/10.1371/journal.pone.0036264

Sauvage, C., Vagner, M., Derôme, N., Audet, C., Bernatchez, L., 2012. Coding Gene SNP Mapping Reveals QTL Linked to Growth and Stress Response in Brook Charr

(*Salvelinus fontinalis*). Genes|Genomes|Genetics, 2(6), 707-720. https://doi.org/10.1534/g3.112.001990

Schlötterer, C., 2004. The evolution of molecular markers — just a matter of fashion? Nat. Rev. Genet. 5, 63–69.

Schott, R.K., Panesar, B., Card, D.C., Preston, M., Castoe, T.A., Chang, B.S.W., 2017. Targeted capture of complete coding regions across divergent species. Genome Biol. Evol. 9, 398–414.

Seeb, J.E., Carvalho, G., Hauser, L., Naish, K., Roberts, S., Seeb, L.W., 2011. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in non-model organisms. Mol. Eco. Res. 11, 1–8.

Shin S. C., Kim S. J., Lee J. K., Ahn D. H., Kim M. G., Lee H., 2012. Transcriptomics and comparative analysis of three Antarctic notothenioid fishes. PLoS ONE. 7:e43762. https://doi.org/10.1371/journal.pone.0043762

Smith, C., Elfstrom, C., Seeb, L., Seeb, J., 2005. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. Molecular Ecology, 14(13), 4193-4203. https://doi.org/10.1111/j.1365-294X.2005.02731.x

Spence, M., Banuelos, M., Marcia, R., Sindi, S., 2019. Detecting inherited and novel structural variants in low-coverage parent-child sequencing data. Methods. https://doi.org/10.1016/j.ymeth.2019.06.025

Stickney, H.L., 2002. Rapid Mapping of Zebrafish Mutations with SNPs and Oligonucleotide Microarrays. Genome Res. 12(12), 1929–1934. https://doi.org/10.1101/gr.777302

Srivastava, S., Avvaru, A., Sowpati, D., Mishra, R., 2019. Patterns of microsatellite distribution across eukaryotic genomes. BMC Genomics, 20(1). http://doi.org/10.1186/s12864-019-5516-5

Suzuki, N., Ueda, K., Sakamoto, H., Sasayama, Y., 1999. Fish Calcitonin Genes: Primitive Bony Fish Genes Have Been Conserved in Some Lower Vertebrates. General and Comparative Endocrinology, 113(3), 369-373.

Tan, H., Xu, Z., Jin, P., 2012. Role of noncoding RNAs in trinucleotide repeat neurodegenerative disorders. Experimental Neurology, 235(2), 469-475. https://doi.org/10.1016/j.expneurol.2012.01.019

Toth, G., 2000. Microsatellites in Different Eukaryotic Genomes: Survey and Analysis. Genome Research, 10(7), 967-981. https://doi.org/10.1101/gr.10.7.967

Vallender, E.J., 2011. Expanding whole exome resequencing into non-human primates. Genome Biol. 12, R87

Vera, M., Álvarez-Dios, J., Millán, A., Pardo, B., Bouza, C., Hermida, M., Fernández, C., de la Herrán, R., Molina-Luzón, M., Martínez, P., 2011. Validation of single nucleotide polymorphism (SNP) markers from an immune Expressed Sequence Tag (EST) turbot, *Scophthalmus maximus*, database. Aquaculture, 313(1-4), 31-41. https://doi.org/10.1016/j.aquaculture.2011.01.038

Vera, M., Alvarez-Dios, J., Fernandez, C., Bouza, C., Vilas, R., Martinez, P., 2013. Development and Validation of Single Nucleotide Polymorphisms (SNPs) Markers from Two Transcriptome 454-Runs of Turbot (*Scophthalmus maximus*) Using High-Throughput Genotyping. International Journal of Molecular Sciences, 14(3),5694-5711.

Wang, Q., Shashikant, C., Jensen, M., Altman, N., Girirajan, S., 2017. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. Sci. Rep. 7(1), 885. https://doi.org/10.1038/s41598-017-01005-x

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), 57-63. doi: 10.1038/nrg2484.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., Watson, M., 2015. Exome Sequencing: Current and Future Perspectives. Genes|Genomes|Genetics 5, 1543–1550.

Xiao, S., Wang, P., Dong, L., Zhang, Y., Han, Z., Wang, Q., Wang, Z., 2016. Whole-genome single-nucleotide polymorphism (SNP) marker discovery and association analysis with the eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA). PeerJ. 4:e2664. https://doi.org/10.7717/peerj.2664

Yang, C., Zhu, X., Sun, X., 2008. Development of microsatellite markers and their utilization in genetic diversity analysis of cultivated and wild populations of the mud carp (*Cirrhina molitorella*). Journal of Genetics and Genomics, 35(4), 201-206. https://doi.org/10.1016/S1673-8527(08)60028-4

Yasuike, M., Iwasaki, Y., Nishiki, I., Nakamura, Y., Matsuura, A., Yoshida, K., Noda, T., Andoh, T., Fujiwara, A., 2018. The yellowtail (*Seriola quinqueradiata*) genome and transcriptome atlas of the digestive tract. DNA Research, 25(5), 547-560. https://doi.org/10.1093/dnares/dsy024

Young, E.T., Sloan J.S., van Riper, K., 2000. Trinucleotide repeats are clustered in regulatory genes in Saccharomyces cerevisiae. Genetics. 54:1053–1068. https://doi.org/10.7717/peerj.2664

Zerbino, D.R., Birney, E., Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs. Genome Res. 2008; 18:821.

Zhu, Y., Xue, W., Wang, J., Wan, Y., Wang, S., Xu, P., Zhang, Y., Li, J., Sun, X., 2012. Identification of common carp (*Cyprinus carpio*) microRNAs and microRNA-related SNPs. BMC Genomics, 13(1), 413.

# CHAPTER 3

# The development and analysis of SNP markers associated with growth rate in dusky kob using exome data


## Abstract

Growth rate is one of the most economically important traits in aquaculture, but the molecular mechanisms involved in the growth of dusky kob (*Argyrosomus japonicus*) is poorly understood. Therefore, the purpose of this study was to assess candidate genes in *A. japonicus*, to try and identify single-nucleotide polymorphism (SNP) markers associated with growth. The exomes of fast- and slow-growing individuals were previously obtained using a whole-exome sequencing approach, which identified a total of 4.5 million variants of which 2.8 million were SNPs. Using the candidate gene approach and a selection of 15 gene regions, 263 putative SNPs were identified, of which 38 SNPs in nine genes were confirmed and identified as having a potential association to the trait of interest. Association of these markers was analysed by performing both case-control and quantitative analyses using 80 individuals (classified as large and small) of dusky kob. These analyses were able to identify eight SNPs in three key genes (Bone morphogenetic protein 2a, Myogenic differentiation 1 and Tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase a) to be associated with growth rate. This research was therefore able to provide important information regarding novel SNPs markers associated with growth and their key genes. Thus, providing essential information, which is necessary for understanding the molecular basis of growth in dusky kob and assisting in future studies


## 3.1) Introduction

In traditional selective breeding, individuals showing a desirable phenotype are selected as breeding candidates for the production of the next generation. This method, called individual- or phenotypic selection, has been used as the standard in terrestrial animal breeding, but also in aquaculture, and it has shown to rapidly improve economically important traits, such as growth rate and disease resistance (Gjedrem et al., 2012). However, phenotypic-based breeding schemes are often time consuming and based on trial and error resulting in increased costs, because it requires a large number of individuals, which have extensive phenotypic variation. Another disadvantage is that these phenotypically selected traits often require multiple generations before a significant response to selection in the trait mean is observed. This is particularly costly and time-

consuming for species with long generation intervals, as is the case for many aquaculture species, including abalone (*Haliotis midae*) and dusky kob (*Argyrosomus japonicus*), which might take up to six years to reach sexual maturity (Griffiths, 1996). This is further complicated by the fact that the majority of commercially important traits in aquaculture are complex traits, meaning that these traits are normally governed by multiple genes with complex interactions occurring between multiple pathways, making the selection of multiple complex traits particularly difficult. Although studies have been conducted in certain aquaculture species such as salmonids, the genetic factors affecting complex traits of economic importance are still relatively unknown (Davidson et al., 2010). This is largely due to studies being hindered by the lack of genomic resources as well as information pertaining to the nature of the trait being investigated. Therefore, a fundamental goal of biological research is to gain understanding regarding the genetic basis of phenotypic variation (Tsai et al, 2015), as this will assist in determining how the underlying gene regulatory networks function and the way in which the environment impacts these traits.

Growth rate is of particular economic importance within aquaculture as it directly contributes to fish production efficiency and outputs. Improvements in growth-related traits could allow for fish to reach a marketable size much earlier, spending less time within the production cycle, subsequently lowering production input costs (Elliot et al., 2002). Alternatively, the faster growing animals could be kept within the production cycle for the same period of time as the slower growers, generating a larger net weight at harvest thereby increasing the profitability of the end product (Slabbert et al., 2010). Consequently, this has resulted in multiple studies being performed in order to understand the genetic basis of growth, with genes in the somatotropic axis and transforming growth factor superfamily being targeted as candidate genes in finfish (De-Santis and Jerry, 2007; Li et al., 2017). The somatotropic axis refers to the hormonal signalling from the hypothalamus to the anterior pituitary gland resulting in the release of growth hormone (GH), which in turn stimulates the production of insulin-like growth factor-1 and 2 (IGF-1, IGF-2), growth hormone-releasing hormone (GHRH) and growth hormone inhibiting hormone (GHIH) (De-Santis and Jerry 2007; Renaville et al. 2002). This axis is responsive to the external environments, such as nutrient intake and culture conditions, which results in the accumulation of protein and adipose to be associated with growth rate and size (De-Santis and Jerry 2007; Richmond et al. 2010). Most of those genes have shown to be associated with growth enhancement in different fish species (Du et al. 1992; Hu et al. 2013; Kang et al. 2002; Tao and Boulding 2003; Tsai et al. 2015; Wargelius et al. 2005). In addition to these hormones and peptides, other genes

have been found to significantly affect growth through physiological networks modulating energy metabolism and muscle growth. Although it is not practical, feasible or cost-effective to detect every single gene and its function (Li et al., 2017), this knowledge can be utilised for the implementation of a candidate gene approach. This approach allows for the prioritisation genes considered to be highly relevant to the trait of interest as the polymorphisms within the targeted genes are believed to be functional (in that it may alter gene product functionality) and can be tested for association with phenotype in question, therefore assisting in the identification of markers associated with a specific trait of interest. Fulton's conditioning factor (K), a trait derived from weight and length, reflects the body shape of the fish and could be of interest to dusky kob aquaculture if found to be significantly associated with the single nucleotide polymorphism (SNP) markers. Estimates of moderate to high heritability for K have frequently been observed in other fish (Nilsson, 1994; Kause et al., 2003; Kause et al., 2007) as well as the strong genetic correlations between K at different growth intervals (Saillant et al., 2007; Wringe et al., 2010). This factor is important and should be taken into consideration when testing for association between loci and growth rate as this factor has often been used for the estimation of growth patterns (Caldarone et al., 2012; Mozsár et al., 2014; Muchlisin et al., 2010).

The aim of this study chapter was therefore to confirm putative SNPs and identify markers associated with growth rate in dusky kob through a candidate gene approach using the annotated exome sequence data from Chapter 2 as the underlying genetics of growth in dusky kob have not been studied with the majority of the South African aquaculture industry relying on quantitative phenotypic data for the selection process of broodstock. Associations between variants and growth were determined by assessing the genotypic association to the quantitative traits: wet weight, length, and conditioning factor. The resulting allele frequencies and corresponding phenotypic data were subsequently analysed to investigate genotype-phenotype associations.

## 3.2) Methods and Materials

### *3.2.1) Experimental study populations*

Fin clip tissue from five families each containing a number of F1 animals (~18 months of age) were derived from two mass spawning events, which occurred at two separate aquaculture facilities using wild individuals. The families from each facility, were reared communally and pedigrees were inferred (ten broodstock individuals, five sires and five dams) based on the msat data generated by Jenkins (2018). All the individuals were

phenotyped for wet weight (W), standard body length (Ls), and Fulton's conditioning factor (K) was derived ($K = 10^6 W/Ls^3$, W = Weight of the fish in grams and Ls = Length of the fish in millimetres). For the association analysis the top and bottom eight individuals from each family were selected based on the length (mm) and live weight (g) to derive a family bias corrected (FBC) cohort, which consisted of large, fast-growing individuals (n=40) and small, slow-growing individuals (n=40) (Figure 3.1). The eight large and eight small animals used to generate the exome data in chapter 2 were from family C. DNA was extracted from the fin clip tissue using a standard CTAB method (Saghai-Maroof et al., 1984).

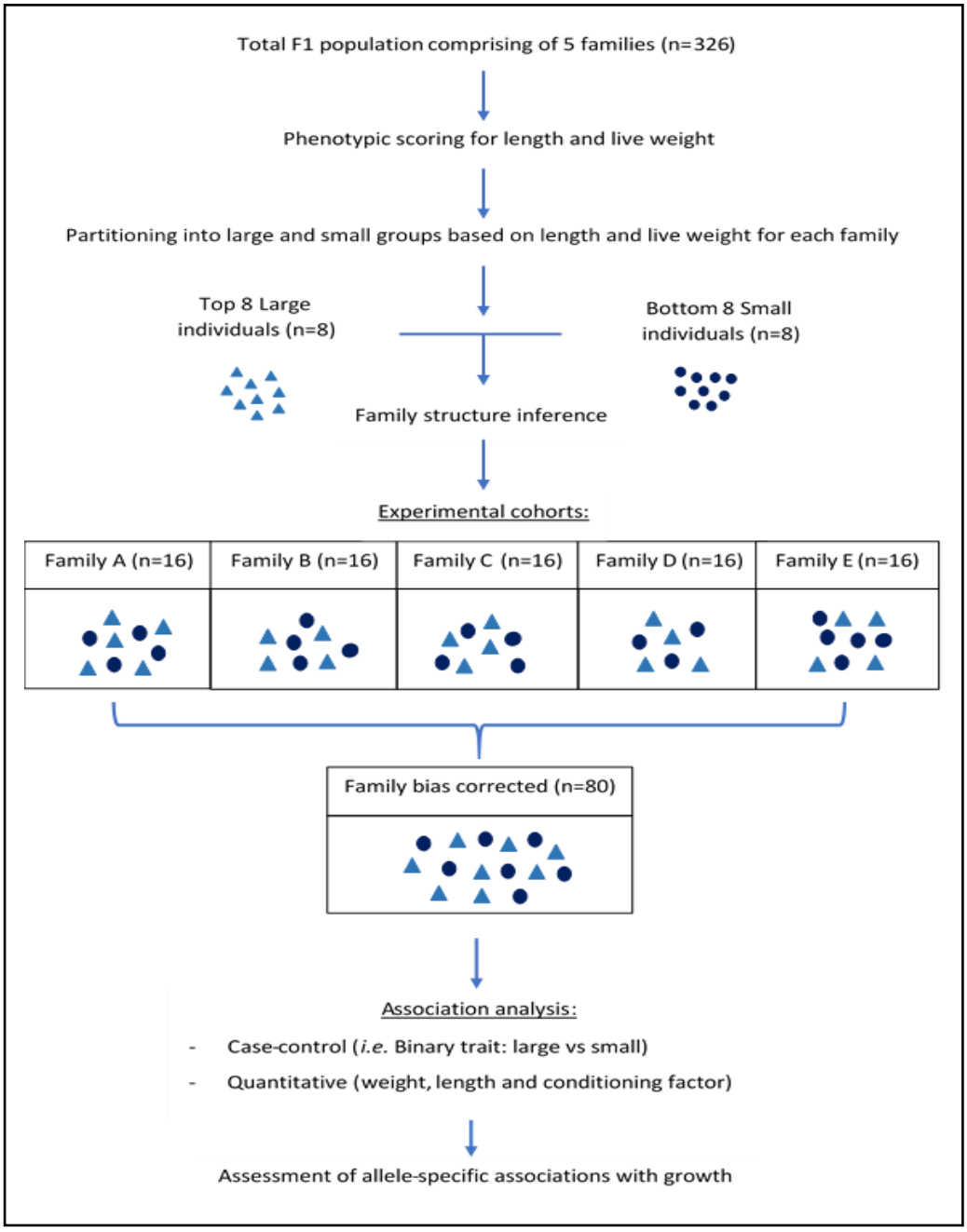



**Figure 3.1. Graphical summary of the methodological approach, detailing the construction of the study populations, the association analyses performed for the various cohorts and the assessment of allele-specific associations with size for significantly associated markers**

### *3.2.2) Variant detection in exome data and primer design*

Using the variant data generated for the 16 individuals in chapter 2, unique variants were identified in each cohort (large and small) using a fixed ploidy variant detection tool in CLC GWB that compares variants within and between groups. The frequency threshold for this analysis was set at 100%, ensuring that a variant is only considered identical in both cohorts when all 16 of the individuals contain the same variant. Due to the large number of variants a candidate gene approach was taken to locate specific genes of interest within the dusky kob sequence. This was made possible using CLC GWB to generate a consensus sequence with a chromosomal outline by aligning the *de novo* assembly of dusky kob to the reference genome of zebrafish. Using knowledge obtained from previous literature where genes were found to be associated with growth in aquaculture species, 15 gene regions were selected and their location determined using the known gene positions of zebrafish that are publicly available on NCBI (Table S3.1) (Han et al., 2018; Kamenskaya et al., 2015; Liu et al., 2014; Li et al., 2018; Opazo et al., 2017). These gene regions were located and extracted using CLC GWB, upon extraction the gene regions were BLAST against NCBI database using the BLASTn function to ensure correct identification. Once correctly identified these extracted gene regions served as a consensus for comparison in downstream analyses. Using Primer3Plus (Untergasser et al., 2007), primer pairs for the 15 gene regions were designed and assessed using the Primer Check function available on Primer3Plus. The specificity of each primer was then assessed using the NCBI's Primer-Blast function and once the minimum requirements were met (Table 3.1), the primers were sent for development (Table S3.2).

**Table 3.1. The requirements for primer design of the 15 gene regions, with the major aspects of primer properties including: specificity (3' stability), GC content, primer length, maximum temperature difference (between forward and reverse primers) and the melting temperature (Tm).**

| Properties | Requirement |
|---|---|
| 3' stability | Below 4 |
| GC content | Between 45-65% |
| Primer length [bp] | 18 to 28 |
| Maximum temperature difference | 2$^{o}$C |
| Tm | 55-60$^{o}$C |

*3.2.3) Putative SNP validation and Genotypic analysis*

The primers designed for the putative SNP markers identified *in silico* for the candidate gene regions, were tested by employing PCR and sanger sequencing. Using PCR, the 15 primer pairs were optimised using the genomic DNA of dusky kob. The stock DNA, which had previously been extracted, using a standard CTAB method (Saghai-Maroof et al., 1984), was evaluated for quantity and quality with the NanoDropTM ND 1,000 spectrophotometer (Thermo Fisher Scientific) and normalised to a final concentration of 20ng/µl. The 15 gene regions were amplified using the KAPA™ Taq PCR Kit, with all reactions being performed to a final volume of 10µl. Each reaction mixture contained 10X KAPA™ Taq buffer (containing 1.5µl of 25mM $MgCl_2$), 0.5µl of 10mM dNTPs, 0.2µl of 10µM stock solution of each primer, 0.1µl of 5U/µL KAPA™ Taq DNA Polymerase3 and 20ng of DNA. The cycling parameters for PCR were identical for all the primer pairs with the exception of the annealing temperature (Ta), as optimisation was necessary for each primer pair, starting at a Ta of 54ºC. A touch down PCR was used for the optimisation process, as to simultaneously assess the effects of different Ta's under the standard PCR conditions. The Ta was appropriately altered each round: increasing incrementally if non-specific products persisted or decreasing incrementally if no product could be observed. The PCR conditions were applied at an initial denaturing step at 95°C for 3 minutes, 35 cycles of denaturation at 95°C for 15 seconds, annealing ranging between 54-60°C for 30 seconds, extension at 72°C for 20 seconds and a final extension step at 72°C for 7 minutes.

After each successful round the PCR products were verified using agarose gel electrophoresis (1.5% w/v; 1 x TBE). Following successful amplification, bi-directional sequencing, was conducted *via* standard Sanger sequencing chemistry (BigDye® terminator V3.1 cycle sequencing kit, Applied Biosystems) and sent to the Stellenbosch University Central Analytical Facility (DNA sequencing unit) for capillary electrophoresis using the 3730xl DNA Analyzer. Upon receiving the sequences, the quality was assessed and the sequences edited in FinchTV 1.4.0 (Geospiza, Inc.; Seattle, WA, USA; http://www.geospiza.com) using the exome data as a reference to discriminate between variants and sequencing errors. Following quality control, the forward and reverse sequences (reverse compliment) were aligned in MEGA7 v7.0 (Kumar et al., 2016), using the ClustalW (Thompson et al., 1994) function for multiple alignments, to the consensus sequence of each gene to identify SNPs. False positive SNPs are found to commonly occur in HTS data, particularly when only partial sections of the genome are sequenced such as the exome. This is because partial sequences, which are often short segments, have an

increased chance of being misaligned. Meaning these sequences are able to align to areas of the reference sequence similar to that of their target sequence, thus potentially showing variation in positions where no true variation exists, hence the need for SNP validation. To validate the putative SNP markers, as identified *in silico*, an initial panel of four animals (two individuals phenotypically characterised as large and two as small, were obtained from family C from which the exome data were derived) were sequenced for each gene region, demonstrating *in silico* nucleotide variation. Visual identification of sequence variation in multiple alignments [MEGA7 (Kumar et al., 2016); ClustalW (Thompson et al., 1994)] was done and confirmed by investigating individual chromatograms. A SNP was confirmed if clear double peaks, could be identified in heterozygous individuals beyond any potential noise, and at least one of the homozygotes were also observed (Figure 3.2). For final confirmation of SNPs in alignments that showed polymorphism, an additional 76 individual offspring and 10 broodstock animals were sequenced, in the forward direction only, and scored; a minor allele frequency of greater than a 5% was prerequisite for final confirmation of a SNP marker. However, SNPs were not removed based on the Hardy-Weinberg Equilibrium (HWE) p-value as the extremes of each family were selected and therefore the population was not expected to conform to HWE. This was performed for all 15 gene regions, using the same reagents, sequencing- and PCR conditions as mentioned above, with the exception of Ta which varied for each primer pair (Table S3.2). Positions showing significant variation (*e.g.* when at least 50% of the individuals in either of the cohorts are shown to have an alternative genotype to the genotype commonly identified in the other cohort) were noted and genotypic data for all individuals (cultured and wild) were collected for these positions.

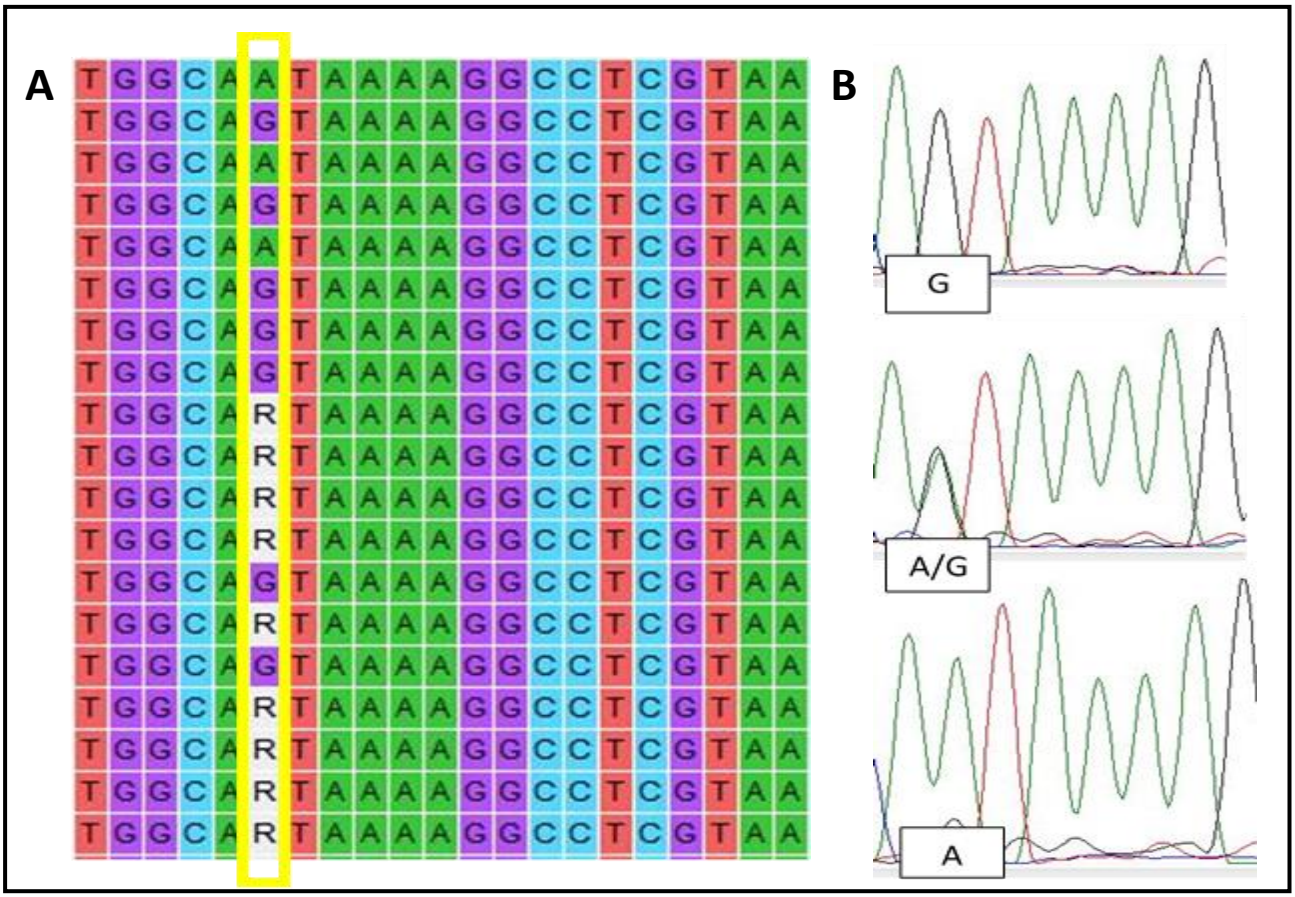



**Figure 3.2. A) A multiple alignment depicting and A>G SNP, showing the two alternative homozygotes for the A and G allele respectively and the heterozygote coded, as the "R" ambiguity (Yellow frame). B) The electropherograms of two homozygous individuals (AA and GG respectively) and a heterozygous individual, demonstrating a clear double peak (Yellow frame).**

*3.2.4) Genetic data analyses*

Allele and genotypic frequencies for the FBC cohort as a whole and individually for the large and small groups for each of the validated SNP markers were estimated and conformation to Hardy-Weinberg equilibrium was tested (exact probability test, 1000 de-memorisation, 100 batches, 1000 iterations per batch) in GenPop v4.2 (Rousset, 2008). Per marker case-control association analysis, as implemented in SNPstats (Solé et al., 2006), was performed using a regression module, correcting for family structure as a categorical covariate. Size effects were determined by estimating odds ratios, and the most likely mode of inheritance was evaluated using the Akaike's- (AIC) and Bayesian Information Criterion (BIC). Additionally, quantitative association analysis was also done by setting wet weight, standard length, and conditioning factor, respectively, as the response variable. Correlations between these phenotypic measurements were evaluated by estimating Pearson's correlation coefficients in XLStatistics v12.11.22 (Carr 2012). Because genotypic data for the broodstock was also available, it was possible to construct parental-offspring trio's and therefore a transmission disequilibrium test (TDT), more robust if genetic structure is evident

within the study population, was performed in HAPLOVIEW software v4.2 (Barrett et al. 2005) as an additional confirmation of genotype-phenotype associations. Haplotype-block structure was determined using the Solid Spine (SS) algorithm with a minimum D' value of 0.8. The SS algorithm is internal to HAPLOVIEW, this method defines a block when the first and last markers are in strong LD with all the intermediate markers. When markers in particular genes showed significant associations, a haplotypic analysis was conducted to evaluate phenotypic associations with haplotypes within those genes. Linkage disequilibrium between genes were also assessed in order to detect any possible gene-by-gene interactions and this was further assessed in terms of phenotypic association in UNPHASED v3.1.3 (Dudbridge 2003; Dudbridge 2008) (1000 permutations) using the gene-by-gene interaction model. Bonferroni adjustment of the significant p-value was done to correct for multiple tests at a 5% nominal level (Dunn 1961). To further minimise the occurrence of false positive results, additional case-control analyses were performed, including exact G-tests for allelic and genotypic differentiation (Goudet et al. 1996) (1000 de-memorisation, 100 batches, 1000 iterations per batch), which were performed in GenePop, and a permutation-based distance test using the FBC cohort to determine both allelic and genotypic association (using Prevosti's distance estimate; significance testing 1000 permutations) (Prevosti et al. 1975; Nielsen and Weir 1999), in PowerMarker v3.25 (Liu and Muse 2005). Single-locus F-tests for associations with quantitative traits (weight, length and conditioning factor) were performed in PowerMarker.

## 3.3) Results

### *3.3.1) Identification of SNP markers*

The variant data generated in chapter 2, identified a total of 4.5 million variants, of which 2.8 million were single nucleotide polymorphisms (SNPs). Using a within group variant detection tool, a total of 1,307,409 and 979,033 variants were identified as being unique to the large and small cohorts, respectively (Figure 3.3). This resulted in the elimination of 2,274,378 variants from further analyses as these variants were determined to be identical across both cohorts. Of the uniquely identified variants a total of 1,428,740 were identified as SNPs, with 263 being located within the 15 selected gene regions (Table 3.3), of which, 58 were identified as being non-synonymous (Table 3.2).

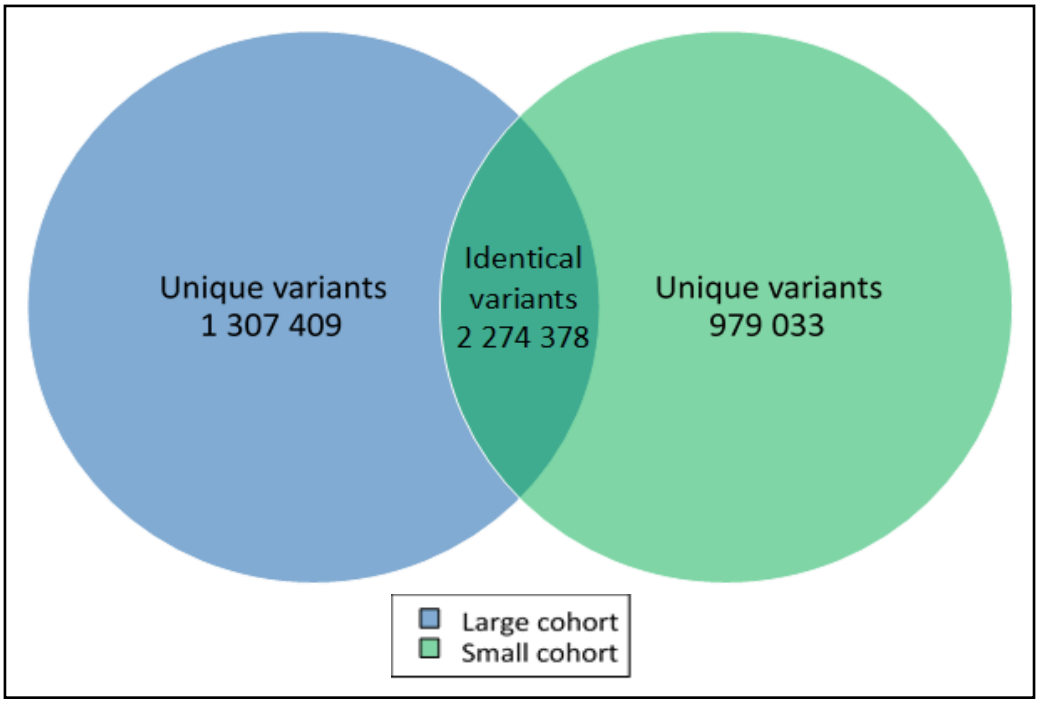



**Figure 3.3. The number of unique variants identified in each cohort, large and small, as well as the number identical variants found to occur between the two cohorts. Variants detected using the within group variant detection tool in CLC GWB. Each cohort is represented by a different colour.**

**Table 3.2. The number of variants identified as SNPs across the exome sequences of *A. japonicus*, with the number of putative and non-synonymous SNPs within the candidate gene regions indicated. The table also includes the total number of confirmed SNPs following Sanger sequencing as well as the number of confirmed SNPs shown to have a possible association with growth.**

| | |
|---|---|
| Total number of SNPs identified in the exome data | 2,840,198 |
| Putative SNPs identified as unique | 1,428,740 |
| Putative SNPs in candidate genes | 263 |
| Non-synonymous SNPs | 58 |
| SNPs confirmed in candidate genes following Sanger sequencing | 97 |
| Confirmed SNPs showing potential association | 38 |

**Table 3.3. The role that the 15 selected gene regions play in the growth and development of marine species is indicated.**

| Gene | Gene symbol | Role in growth |
|---|---|---|
| Bone morphogenetic protein 2a | *bmp2a* | Responsible for the enzymatic delivery of aminoacyl tRNAs to the ribosome. |
| STT3 oligosaccharyl transferase complex catalytic subunit B | *sttb3* | Regulate protein interactions and stability |
| Growth differentiation factor 6a | *gdf6a* | Embryonic development, cell growth, morphogenesis, tissue repair |
| Myogenic factor 5 | *myf5* | Induces cartilage and bone formation. Differentiation of myoblasts into osteoblasts |
| Myogenic factor 6 | *myf6* | Muscle development, commits undifferentiated cells to the muscle lineage |
| Fibroblast growth factor 4 | *fgf4* | Relays growth and stress signals to protein synthesis |
| Growth hormone releasing hormone receptor A | *ghrhra* | Mediates co-translational and post-translational N-glycosylation of target proteins |
| Eukaryotic translation elongation factor 1 alpha 1, like 1 | *eef1a1l1* | Controls timing of retinal neurogenesis and growth |
| Growth hormone regulated TBC protein 1a | *grtp1a* | Direct role in rib, spine, and extraocular muscle formation |
| Tubulin, alpha 8 like 2 | *tuba8l2* | Muscle growth and development |
| Myogenic differentiation 1 | *myod1* | Embryonic development |
| Eukaryotic translation elongation factor 2, like 2 | *eef2l2* | Role in cell growth and DNA damage control |
| Tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase a | *tnksa* | Microtubule cytoskeleton organization and mitotic cell cycle |
| Cadherin 4, type 1, R-cadherin | *cdh4* | Cell migration, cell adhesion, sorting and tissue morphogenesis |
| Clock circadian regulator a | *clocka* | Regulates cellular and developmental processes and provides higher fitness under diurnal conditions |

Of the 263 putative SNPs identified using *in silico* exome data, a total of 97 were confirmed *via* Sanger sequencing in the 15 gene regions (Table 3.2). Of the 97 SNPs, a total of 38 SNPs in nine genes were identified as having significant variation between the large and small cohort (*e.g.* when at least 50% of the individuals in either of the cohorts are shown to have an alternative genotype to the one commonly identified in the other cohort) (Figure 3.4), therefore indicating possible association of these positions with the growth of dusky kob. Of the 38 SNPs, the *BMP2A* gene region showed the highest level of variation, with a total of 15 SNP positions while *GDF6A* and *EEF1A1L1* had the least variation, each gene only showing a single SNP position.

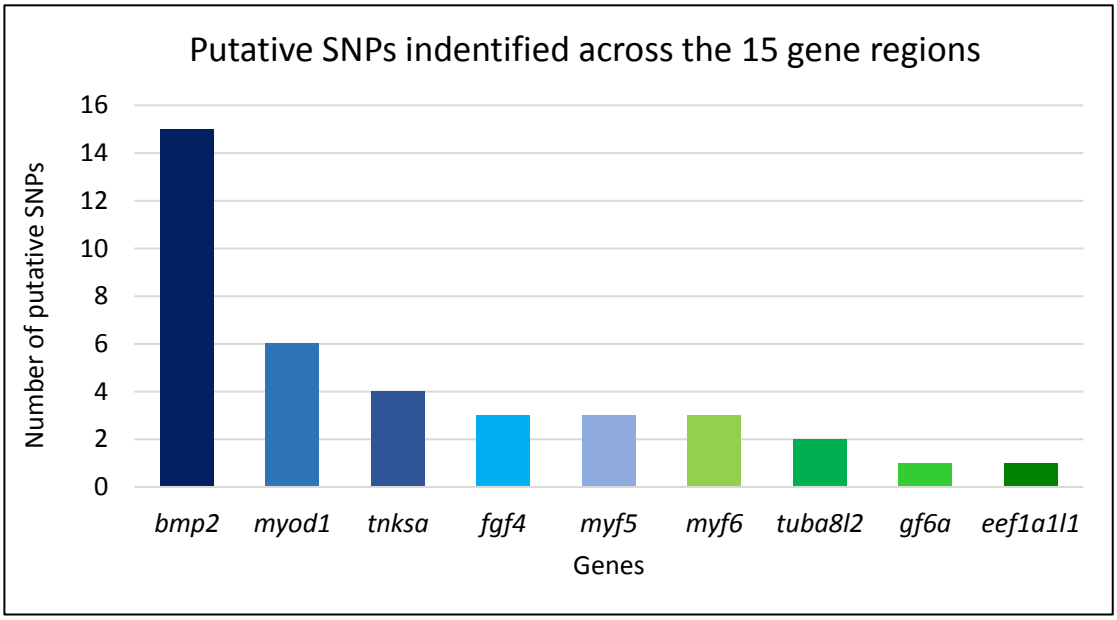



**Figure 3.4. The number of SNPs identified across the 15 gene regions as potentially having a significant association with growth as determined by sanger sequencing**

### *3.3.2) Association analysis*

Within the FBC cohort a total of eight SNPs were identified in three genes, which were found to be significantly associated with growth based on the case-control and quantitative analyses performed in SNPstats (Table 3.4). Additionally, it was determined that family was not a significant covariate, showing the successful correction of population substructure. These results were further validated by the additional analyses performed in PowerMarker and GenePop (G-tests, F-tests and distance test) (Table S3.3).

**Table 3.4. The significant SNPs identified in the FBC cohort as determined by the case-control analysis performed in SNPstats using size (Large or Small) as the response. The correlating allele frequencies and HWE p-value determined in GenePop are indicated for each of the SNPs.**

| Name | SNP | Model | Genotype | Large | Small | OR (95% CI) | p-value | AIC | BIC | Allele | Allele frequencies | | | HWE p-value | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | All | Large | Small | Large | Small |
| *tnksa-1* | c.69T>C | Dominant | C/C | 37 (92.5%) | 16 (40%) | 1.00 | <0.0001 | 87.8 | 92.5 | C | 0.70 | 0.96 | 0.44 | <0.0001 | 1.000 |
| | | | C/T-T/T | 3 (7.5%) | 24 (60%) | 18.50 (4.86-70.36) | | | | T | 0.30 | 0.04 | 0.56 | | |
| *tnksa-3* | c.74C>T | Dominant | T/T | 40 (100%) | 18 (45%) | 1.00 | <0.0001 | 75.8 | 80.6 | C | 0.82 | 0.74 | 0.90 | 0.033 | 0.094 |
| | | | C/T-C/C | 0 (0%) | 22 (55%) | 0.20 (0.07-0.57) | | | | T | 0.18 | 0.26 | 0.10 | | |
| *myod1-1* | c.34T>G | Dominant | G/G | 33 (82.5%) | 15 (37.5%) | 1.00 | <0.0001 | 97.2 | 102 | T | 0.61 | 0.56 | 0.65 | 0.037 | 0.520 |
| | | | G/T-T/T | 7 (17.5%) | 25 (62.5%) | 7.86 (2.79-22.16) | | | | G | 0.39 | 0.44 | 0.35 | | |
| *myod1-3* | c.62T>G | Over-dominant | G/G-T/T | 37 (92.5%) | 19 (47.5%) | 1.00 | <0.0001 | 93.8 | 98.6 | T | 0.78 | 0.8 | 0.75 | 0.400 | 1.000 |
| | | | G/T | 3 (7.5%) | 21 (52.5%) | 13.63 (3.60-51.55) | | | | G | 0.22 | 0.20 | 0.25 | | |
| *bmp2-1* | c.2T>G | Dominant | T/T | 31 (77.5%) | 11 (27.5%) | 1.00 | <0.0001 | 93.9 | 98.7 | T | 0.57 | 0.78 | 0.36 | <0.0001 | <0.0001 |
| | | | T/G-G/G | 9 (22.5%) | 29 (72.5%) | 9.08 (3.29-25.08) | | | | G | 0.43 | 0.22 | 0.64 | | |
| *bmp2-5* | c.25G>C | Over-dominant | G/G-C/C | 31 (77.5%) | 17 (42.5%) | 1.00 | 0.0012 | 104.4 | 109.2 | G | 0.65 | 0.74 | 0.56 | 0.350 | 0.011 |
| | | | C/G | 9 (22.5%) | 23 (57.5%) | 4.66 (1.76-12.31) | | | | C | 0.35 | 0.26 | 0.44 | | |
| *bmp2-11* | c.109T>C | Dominant | T/T | 18 (45%) | 4 (10%) | 1.00 | 0.0009 | 101.9 | 106.6 | T | 0.59 | 0.69 | 0.5 | 0.000 | 0.720 |
| | | | C/T-C/C | 22 (55%) | 36 (90%) | 7.36 (2.20-24.60) | | | | C | 0.41 | 0.31 | 0.50 | | |
| *bmp2-15* | c.123C>G | Dominant | C/C | 20 (50%) | 3 (7.5%) | 1.00 | <0.0001 | 95.7 | 100.4 | C | 0.60 | 0.72 | 0.48 | 0.000 | 0.690 |
| | | | C/G-G/G | 20 (50%) | 37 (92.5%) | 12.33 (3.26-46.63) | | | | G | 0.40 | 0.28 | 0.52 | | |

Of the eight SNPs, the majority (6) were determined as having a dominant mode of inheritance, while the two remaining SNPs were determined to be over-dominant (Table 3.4). The estimates of HWE determined that only four of the eight SNPs were found to be significant, with a p-value of 0.001 or less (Table 3.4), while the remaining four SNPs deviated significantly from HWE. All eight of the SNPs were identified as non-synonymous, with the amino acid changes between the large and small cohorts displayed in Table 3.5.

**Table 3.5. Amino acid changes for the eight non-synonymous SNPs identified as significant in the case-control and quantitative analyses.**

| SNP | Nucleotide change | Amino acids | |
|---|---|---|---|
| | | Small | Large |
| *tnksa-1* | T/C | Phe | Ser |
| *tnksa-3* | C/T | Ser | Phe |
| *myod1-1* | T/G | Pro | Arg |
| *myod1-3* | G/T | Gly | Val |
| *bmp2-1* | G/T | Arg | Ile |
| *bmp2-5* | C/G | Glu | Gln |
| *bmp2-11* | C/T | Ser | Phe |
| *bmp2-15* | G/C | Arg | Thr |

*3.3.3) Transmission disequilibrium test and Haplotypic associations*

The TDT results indicated the over transmitted allele for each of the SNP positions with the p-value for these markers ranging between 0.0000 and 0.0060 (Table 3.6). The minor allele frequency (MAF) at each marker ranged between 32 and 50% (Table 3.6), with all eight SNPs exhibiting heterozygosity. The corresponding estimated D′ values are depicted using a colour scheme in Figure 3.5, where the SS algorithm determined that both of the SNPs in *TNKSA* and three out of the four SNPs in *BMP2A* were identified as LD blocks in their respective genes. There was a lower-than-expected degree of LD observed for *bmp2-1*, in regards to the other SNPs within the same gene. Three major haplotypes for the *TNKSA* gene, accounted for all of the alleles: haplotype 1, –CT– (53.6%); haplotype 2, –TC– (32.1%), and haplotype 3, –CC– (14.3%). The *BMP2A* gene however, only had two major haplotypes which accounted for 83.8% of the alleles: haplotype 1, –GTC– (48.7%), and haplotype 2, –GCC– (35.1%) (Table 3.7). The assessment of the genes separately did not differ from the results when the genes was assessed simultaneously. The program was able to identify the same two LD blocks whether the data was combined or separated into individual genes. No linkage was observed between the genes (Figure S3.1).

**Table 3.6. Transmission disequilibrium test results and the characteristics of the SNPs in the *BMP2*, *TNKSA* and *MYOD1* genes. The over-transmitted allele, transmitted to non-transmitted (T:U) ratio, p-value, alleles (A>B where B is the minor allele) and minor allele frequency (MAF) indicated for each SNP position.**

| SNP | Over-transmitted | T:U | p-value | Alleles | MAF |
|---|---|---|---|---|---|
| *bmp2-1* | T | 62.0:18.0 | 0.0000 | G>T | 0.500 |
| *bmp2-5* | G | 43.0:23.0 | 0.0000 | C>G | 0.429 |
| *bmp2-11* | T | 47.0:25.0 | 0.0000 | C>T | 0.464 |
| *bmp2-15* | C | 47.0:17.0 | 0.0000 | G>C | 0.500 |
| *tnksa-1* | C | 61.0:3.0 | 0.0060 | T>C | 0.464 |
| *tnksa-3* | T | 48.0:0.0 | 0.0010 | C>T | 0.321 |
| *myod1-1* | G | 57.0:7.0 | 0.0000 | T>G | 0.393 |
| *myod1-3* | G | 57.0:7.1 | 0.0000 | T>G | 0.357 |

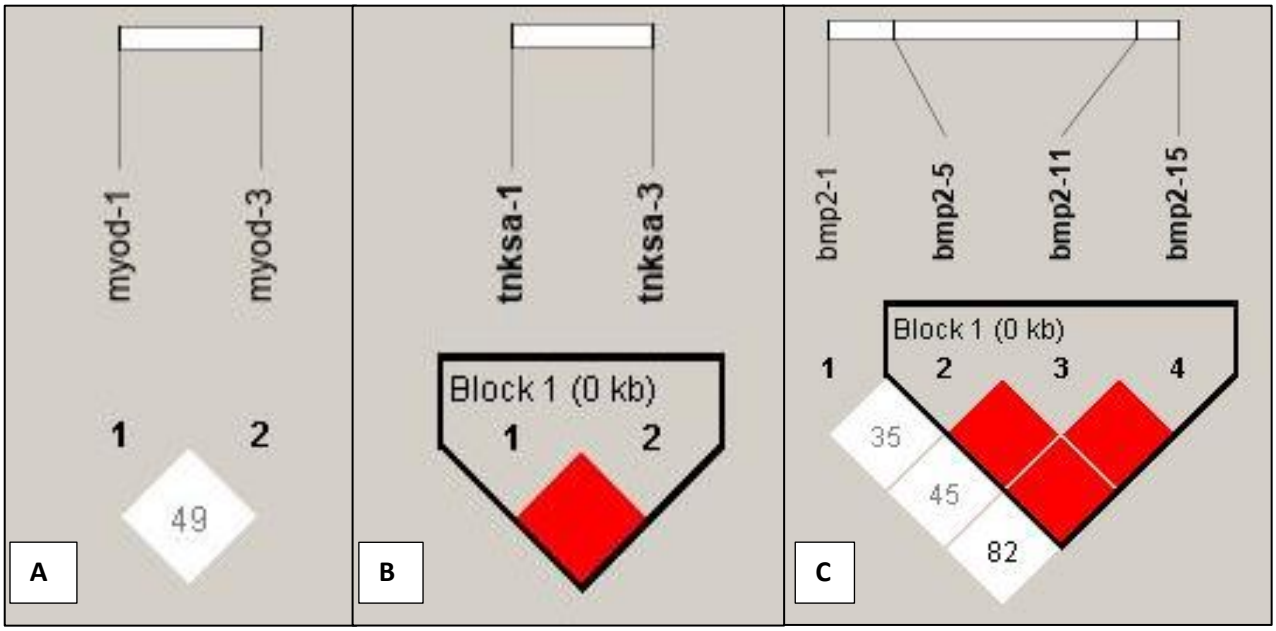



**Figure 3.5. Linkage disequilibrium (LD) block structures. LD block structure consisted of a total of ten SNPs in three different genes. Two SNPs were located in the (A) *MYOD1* gene, two SNPs in the (B) *TNKSA* gene and four SNPs in (C) *BMP2A* gene. The LD block was defined by a D' value threshold of 0.8. The colour scale ranges from red to white (colour intensity decreases with decreasing D' value, and all of D' values were = 1).**

**Table 3.7. Haplotype associations determine for the three LD blocks identified in *TNKSA* and *BMP2A*. The frequency of the haplotype, transmission to non-transmitted (T:U), Chi-Square and p-value are all indicated for each haplotype. The OR (95% CI) is given for the most frequent haplotype.**

| Gene | Haplotype | Freq | T:U | OR (95% CI) | Chi square | p-value |
|---|---|---|---|---|---|---|
| *bmp2a* | | | | | | |
| | GTC | 0.487 | 35.5 : 8.5. | 13.33 (1.04 – 170.51) | 16.600 | 0.0000 |
| | CCG | 0.351 | 4.6 : 23.6. | | 12.691 | 0.0004 |
| *tnksa* | | | | | | |
| | CT | 0.536 | 61.0 : 1.0 | 22.98 (1.20 – 438.87) | 58.065 | 0.0000 |
| | TC | 0.321 | 0.0 : 46.0 | | 46.000 | 0.0000 |
| | TT | 0.143 | 1.0 : 15.0 | | 12.250 | 0.0005 |

Gene–gene interaction analysis was performed for the six SNPs found in *TNKSA* and *BMP2A* to assess the combined effects of these genetic variants on growth. The interactions among all tested SNPs demonstrated a significant interaction between *tnksa-3* and *bmp2-5*, amounting to an OR for interaction of 1.743, $X^2$ =2.387, p-value=0.026 (Table 3.9). In addition to this finding, a p-value of less than 0.05 was observed for one other combination: *tnksa-3* and *bmp2-11* (OR =1.413, $X^2$ = 2.981, p-value = 0.049). However, these p-values did not remain statistically significant after Bonferroni correction (p-value ≥0.001) (Table 3.8).

**Table 3.8. Gene-gene interaction analysis between *BMP2A* and *TNKSA*, the corresponding OR (odds ratio), $X^2$ and p-value are given for each genotype combination**

| SNP | | SNP | | OR (95% CI) | $X^2$ | p-value |
|---|---|---|---|---|---|---|
| Gene | Identifier | Gene | Identifier | | | |
| *TNKSA* | *tnksa1* | *BMP2A* | *bmp2-1* | 1.216 | 0.370 | 0.543 |
| *TNKSA* | *tnksa1* | *BMP2A* | *bmp2-5* | 0.938 | 0.172 | 0.678 |
| *TNKSA* | *tnksa1* | *BMP2A* | *bmp2-11* | 0.812 | 1.575 | 0.210 |
| *TNKSA* | *tnksa1* | *BMP2A* | *bmp2-15* | 1.259 | 2.254 | 0.133 |
| *TNKSA* | *tnksa3* | *BMP2A* | *bmp2-1* | 1.186 | 0.781 | 0.377 |
| *TNKSA* | *tnksa3* | *BMP2A* | *bmp2-5* | 1.743 | 2.387 | 0.026 |
| *TNKSA* | *tnksa3* | *BMP2A* | *bmp2-11* | 1.413 | 2.981 | 0.049 |
| *TNKSA* | *tnksa3* | *BMP2A* | *bmp2-15* | 1.042 | 0.036 | 0.884 |

Phenotypic correlations between Fulton's conditioning factor K and the quantitative traits W and Ls were negative and not significant. The phenotypic correlation between Ls and W were significant, showing a high positive correlation between the traits (Table 3.9; Figure S3.2).

**Table 3.9. Correlation matrix (Pearson) showing the positive and negative correlations between the quantitative traits: Weight (g), Length (mm) and Conditioning factor (K).**

| Variables | Weight (g) | Length (mm) | Conditioning factor (K) |
|---|---|---|---|
| Weight (g) | **1** | 0.926 | -0.040 |
| Length (mm) | 0.926 | **1** | -0.362 |
| Conditioning factor (K) | -0.040 | -0.362 | **1** |

## 3.4) Discussion

The exome sequence reads provided a significant amount of variant data, which was narrowed down using a candidate gene approach. This approach assisted in isolating genes, which have found to be associated with growth in other marine species, thus increasing the likelihood of SNPs identified within these regions being associated with the trait of interest in kob (Kwon and Goate, 2000). The use of a within group variant detection tool assisted in identifying variant rich regions within the data, which were most likely to have an association with growth (Figure 3.2). This was achieved through the elimination of identical variants (~2.2M) and the identification of variants found to be unique to either the large or small cohort. This elimination resulted in the loss of more than half of all the variants. The most likely cause for the large number of identical variants observed within the exome data, is due to the data being generated using a single family, thus introducing ascertainment bias. The use of a single family to generate the exome data was a necessary step to try and minimise the effects of additional factors on the sequence data, such as epistasis and differing parental genotypes. Thus, to achieve the most reliable consensus sequence possible, each individual from the family was theoretically supposed to be sequenced at the same positions, therefore each individual would validate another individual's sequence data. Thus, the variation seen between the individuals of the two cohorts (large and small) could potentially be associated with the individual's size rather than the result of differing parental genotypes, in the total population, for example. Differing parental genotypes is often a consequence of commercial mass spawning events, which is essential for species such as *e.g.* rainbow trout (Pante et al., 2001), *Oreochromis niloticus* (nile tilapia) (Fessehaye et al., 2009), abalone (Kobayashi and Kijima, 2010), shrimp (Moss et al., 2007) and dusky kob (Jenkins, 2018). These mass spawning species are unable to

use single pair mating designs, thus there is no control over the contribution of broodstock within a spawning event. In a typical spawn, the broodstock contributions is likely to be skewed as a result of certain individuals not contributing towards the offspring, meaning that a single mass spawning can result in multiple different broodstock combinations (Brown et al., 2005; Herlin et al., 2008; Hillen et al., 2017; Sekino et al., 2004). Therefore, the use of a single family, which had been determined *via* parentage analysis using microsatellites (Jenkins, 2018), was essential to eliminate the possibility of differing parental genotypes. This ensured that any variation observed to be occurring between the cohorts was in fact true and not a result of differing parental genotypes.

Using sanger sequencing a total of 38 SNPs were verified, across nine of the 15 gene regions: Tankyrase (*TNKSA)*, Bone morphogenetic protein 2a (*BMP2A*), Fibroblast growth factor 4 (*FGF4*), Myogenic differentiation 1 (*MYOD1*), Myogenic factor 5 (*MYF5*), Myogenic factor 6 (*MYF6*), Tubulin alpha 8 like 2 (*TUBA8L2*), Eukaryotic translation elongation factor 1 alpha 1, like 1 (*EEF1A1L1*) and Growth differentiation factor 6a (*GDF6A*). By contrast, two gene regions, *GHRHRA* and *GRTPA1*, often identified in other studies as having variation associated with growth (Deane and Woo, 2009; Fuentes et al., 2013; Opazo et al., 2017), were devoid of SNPs in this present study. This unexpected result could be attributed to the fact that only a small section of each gene was able to be sequenced due to the uneven coverage of the exome data, resulting in variant dense regions remaining undetected. Therefore, showing these regions to have little to no variation between the cohorts, which is most likely not the case, but rather a consequence of the limited sample size, which consisted of only five different families. Despite the frequent mention of these genes in association studies, particularly in mammals, it is not uncommon for the sequence of genes such as the growth hormone releasing hormone (*GHRH*) gene to be fixed within a species. Meaning, that although this gene does play a critical role in the growth and development of the individual, any variation observed within this region, is unlikely to alter the phenotype of interest. It is also important to note that all though mammals and fish do share a number of similarities in regards to biological pathways they are remarkably different, meaning that although they do share the same hormones, the way in which their bodies utilise these hormones may differ. For example, in mammals both *GHRH* and pituitary adenylate cyclase-activating polypeptide (*PACAP*) belong to the same family of regulatory peptides, which are responsible for the regulated release of the growth hormone. However, in fish *PACAP* is more pronounced, having a prolonged stimulatory effect on the release of the growth hormone, while *GHRH* has been observed as having little to no effect (Tao and Boulding,

2003). This observation has been made in numerous fish species such as rainbow trout (*Oncorhynchus mykiss*, Luo et al., 1990; Blaise et al., 1995), salmon (*Oncorhynchus nerka*, Parker et al., 1997), carp (*Cyprinus carpio*, Vaughan et al., 1992) and in a tilapia hybrid (*Oreochromis niloticus* / *O. aureus*, Melamed et al., 1995). So, despite the gene regions, *GHRHRA and GRTPA*, not exhibiting any variation, a large number of putative SNPs were observed in the less frequently mentioned *BMP2A* region. Currently, no other studies have linked SNPs found within this gene to growth rate, however, *BMP2A* has been identified as a initiator of bone formation, which is an essential component in the formation of vertebrates (Maegawa et al., 2007).

Association analyses under different inheritance models revealed eight SNPs (*tnksa-1, tnksa-3, myod1-1, myod1-3, bmp2-1, bmp2-5, bmp2-11* and *bmp2-15*) to be associated with size (Table 3.4; Table S3.3; Table S3.4), following Bonferroni correction (p-value <0.001). Bonferroni correction was essential following multiple tests, tests which were necessary to try and minimise the number of false positives. Although false positives are a common occurrence when performing multiple tests, there is currently no other genomic resources for dusky kob which can be used for verification and validation, making a stringent approach necessary. Thus, all SNPs identified to be associated with growth, were done so using the conservative Bonferroni correction. This method did reduce a number of identified genetic associations; however, may have resulted in a number of false negatives. Although the elimination of true SNPs may seem counterintuitive in an association study, these SNPs can all be reanalysed in future studies once the genomic resources of dusky kob are more readily available. Another major cause of false positive results in genetic association studies, is population substructure, particularly in case-control studies. This is because the cases and controls are often sampled from genetically different underlying populations, thus any associations found could be a result of the underlying structure of the population and not actually associated with the trait of interest (Satten et al., 2001; Tian et al., 2008). This issue is found to be particularly prevalent in aquaculture as parental contributions are skewed as a result of mass spawning, making pedigree inference essential as aforementioned to try and correct for any bias caused as a result of population substructure due to high variance in family sizes. However, this expected bias was not observed within the FBC cohort, as family was determined not to be a significant covariate. This illustrates that the sampling method utilised for size selection was able to eliminate or more likely reduce, the effect of population substructure on the association results and can therefore be incorporated into future aquaculture studies (Figure 3.1).

Out of all eight SNPs, seven deviated from HWE in either the large or small cohort, with the exception of *bmp2-5* which was found to in HWE for both the large and small cohort (Table 3.4). The HWE principle states that in the absence of migration, mutation, natural selection and assortative mating, genotype frequencies at any locus are a simple function of the allele frequencies (Wigginton et al., 2005; Wittke-Thompson et al., 2005). Therefore, deviation from these principles could result in disagreement for the HWE model. This disagreement could be indicative of genotyping errors (Wigginton et al., 2005). However, the results observed in this study, indicated that the genotypic differences were occurring between the case and control groups, with the genotypes verified using sanger sequencing. Thus, making it highly unlikely that the observed deviations from HWE were the consequence of genotyping errors. It is thus generally accepted that when a variant increases the chance of a particular phenotype occurring, it may cause a deviation from HWE, meaning the variant is associated with the phenotype. Therefore, deviations from HWE are generally expected to occur within the phenotype under investigation while HWE is met in the alternative phenotype. Although, the reverse is possible, as seen for three of the eight SNPs, *bmp2-5*, *myod1-3* and *tnksa-3*. The assumption that can be made in these cases is that deviation from HWE in the small cohort is possibly still associated with size although it may not be for the desired phenotype but rather the small phenotype (Wigginton et al., 2005; Wittke-Thompson et al., 2005). This hypothesis does correlate with the OR for *tnksa-3*, which does appear to be associated with the small phenotype rather than the large although this would need to be validated using a larger sample size. Interestingly though all eight SNPs were determined to be non-synonymous (ns), altering the encoded amino acid at the variable site, causing structural and functional changes in the coding protein, which may be affecting the phenotype of interest (Table 3.5). However, not all alterations are favourable, as there are structural or functional changes which could be deleterious or damaging. Therefore, it is important to identify these changes and determine whether any of the nsSNPs are deleterious. In this study, CLC GWB was able to identify the nsSNPs, which in this case were determined to be non-deleterious, meaning that these changes to the protein are more likely to have an association with the phenotype of interest.

In this study, the gene identified as having the greatest number of significantly associated SNPs was *BMP2A*, with a total of four novel SNPs (*bmp2-1*, *bmp2-5*, *bmp2-11* and *bmp2-15*). By assessing the SNPs individually within this gene, the over transmitted alleles were determined for each SNP position: The T allele at *bmp2-1*, the G allele at *bmp2*-5, the T allele at *bmp2-11* and the C allele at *bmp2-15* (Table 3.6). Using this information, the mode

of inheritance was determined, where three of the four SNPs (*bmp2-1*, *bmp2-11* and *bmp2-15*) were identified as having a dominant mode of inheritance, where the large phenotype was observed to have a high prevalence for both the heterozygous and homozygous genotype containing the over transmitted allele. For these three SNPs the ORs ranged between 9.08 and 12.33 indicating that there was an increased likelihood of at least nine times or greater for the associated genotype to be present in the large phenotype. However, unlike the other SNPs in *BMP2A*, *bmp2-5* was found to be over-dominant, meaning that the large phenotype was showing a high prevalence to the heterozygous genotype, with an OR of 4.66 indicating a greater than 4 times increased likelihood of the heterozygous genotype to be present within the large phenotype. It was not unexpected that a large number of significant SNPs were identified within the *BMP2A* gene, as bone morphogenetic protein genes (*BMPs*), include many genes that play a fundamental role in embryonic skeletal development, vertebrate development and postnatal skeleton homeostasis (Wu et al., 2016). Therefore, because of the diversity seen in the function of these genes there has been an increase in the number of studies performed, which have focused on the roles that *BMPs* play in a wide range of biological processes.

Previously, this group of genes was thought to only be involved in bone and cartilage formation, however, it has since been determined as playing role in embryogenesis, muscle growth, adipogenesis and reproductive system development (Brazil et al., 2015; Wang et al., 2014). The fundamental role of *BMP2* was previously highlighted in a study performed in *Sparus aurata* by Rafael et al. (2006), where the *BMP2* gene was suggested as playing a key role in fish development (Sekelsky et al., 1995). More recently, a study was performed in common carp by Chen et al. (2017), where the function and number of *BMP* genes were identified and although *BMP2A* was not specifically mentioned to be significantly associated with any role within this study. This gene was mentioned as playing a fundamental role in biological processes while being highly conserved across 20 different species, thus leading to this genes inclusion within this study. When assessing the linkage disequilibrium (LD) (>0.8) of the SNPs within the *BMP2A* gene (Figure 3.5), a single linkage block was identified, where three out of the four SNPs (*bmp2-5, bmp2-11 and bmp2-15*) showed strong LD (D' >0.8 and $r^2$ >0.8). Within the linkage block two significant haplotypes were identified, GTC (*bmp2-5*, *bmp2-11* and *bmp2-15*) and CCG *(bmp2-5*, *bmp2-11* and *bmp2-15*), where the GTC haplotype was associated with a higher frequency within the large cohort, compared to small, indicating a 13.33-fold increase of displaying the desired phenotype (OR = 13.33) (Table 3.7). While the second haplotype, CCG, was associated with a higher frequency

within the small cohort, compared to large. This strong linkage disequilibrium seen between the SNPs of the *BMP2A* could potentially be utilised in marker assisted selection (MAS), as alleles of a few SNPs in a haplotype can suggest that the alleles of the other SNPs, provide redundant information. Therefore, a small number of common SNPs can be selected from each haplotype which would be sufficient to define the relevant haplotypes (Takeuchi et al., 2005).

In addition to the four SNPs identified in *BMP2A*, two novel SNPs were also identified in each of the *TNKSA* and *MYOD1* genes, with both SNPs in *TNKSA* having a dominant mode of inheritance. However, unlike *BMP2A*, both the SNPs in *TNKSA* (*tnksa-1* and *tnksa-3*) were not observed to be associated with the desired large phenotype. Unlike *tnksa-1* which showed its over transmitted allele to have a high prevalence for the large phenotype (OR = 18.50), *tnksa-3* did not have a OR greater than 1 (OR = 0.20). When assessing the LD in the gene, the program was able to identify a single linkage block in *TNKSA* where both SNPs were found to be linked (Figure 3.5). This was an interesting result for the *TNKSA* gene as the SNP *tnksa-3* had been identified as significant throughout the association tests including the case trios, despite it having an extremely low OR. The low OR in combination with the linkage results could indicate the probability of this SNP being inherited as a result of *tnksa-1* which is found to have an 18.5 greater chance of being found to occur within the large phenotype, rather than being directly involved in observation of the desired phenotype. Finally, the two SNPs in the *MYOD1* gene (*myod1-1* and *myod1-3*) were identified to be dominant and over dominant in terms of inheritance with the large phenotype with an OR of 7.86 *myod1-1* and 13.63 for *myod1-3*.Of the four novel SNP markers identified in *TNKSA* and *MYOD1*, two SNPs showed weak LD (D' <0.8 and r2 <0.8) but yielded statistical significance for the individual SNPs among the case trios. In particular, the G allele at *myod1-1* and the G allele at *myod1-3*, were shown to have a significant presence in the phenotypically large individuals (Table 3.6). However, under the standard confidence interval settings, no significant haplotype associations between the large and small cohorts were identified in the *MYOD1* gene. Therefore, the only other gene found to have haplotype associations besides *BMP2A*, was *TNKSA* which showed three significant haplotypes, however, only the CT (*tnksa-1* and *tnksa-3*) haplotype was found to have a higher association within the large cohort, compared to the small, indicating a 22.98-fold increase in displaying the desired phenotype (OR = 22.98) (Table 3.7). While the other two haplotypes TC (*tnksa-1* and *tnksa-3*) and TT (*tnksa-1* and *tnksa-3*) were not found to be associated with the large phenotype but rather the small. *TNKSA* and *MYOD1*, have been highlighted in

other studies for the fundamental role that these genes play in growth. Particularly, *TNKSA*, where a study performed by Chiang et al. (2008) on mice, showed that the knockout of this gene did result in a reduction in size of the individual, however, this knockout did not affect the development of the animal. So, although this gene has been observed to impact the growth of both humans and mice it is still under investigation within fish making this an interesting result. Unlike the other two genes, *MYOD1* has been well investigated in fish such as flounder (Zhang et al., 2006), *Sparus aurata* (gilt-head seabream) (Tan and Du 2002), *Hippoglossus hippoglossu* (Atlantic halibut) (Galloway et al. 2006) and the *Takifugu rubripes* (tiger pufferfish) (Macqueen and Johnston 2006), where this gene has been observed to influence the muscle development in growing fish. However, the normal stimulus for muscle growth in growing fish is still not well understood. Therefore, by understanding the regulation of muscle growth in fish, it can be of particular importance to aquaculture. This is because fish meat consists primarily of skeletal muscle, meaning that stimulation of muscle growth could result in fish having an increased growth rate or larger size (Zhang et al., 2006), making the *Myod1* gene a pivotal part of this study.

Interestingly all three genes (*BMP2A*, *TNKSA* and *MYOD1*) that were identified as having SNPs significantly associated with growth were part of large protein families, which are found to be highly conserved between different species (Hsiao and Smith, 2008; Rafael et al., 2006; Smith, 1992; Wozney, 1998). This conservation could explain why the exome data was able obtain a large number of sequences for these regions, enabling the identification of variant dense regions. However, despite identifying associated SNPs in all three genes, only two of the three genes (*TNKSA* and *BMP2A*) were determined as having haplotypes associated with the desired phenotype, CCG in the *BMP2A* gene and CT in the *TNKSA* gene. In addition to their identification, both haplotypes were determined to have a larger OR than the individual SNPs showing that there is variation across the gene rather than a single SNP. These genes do require further investigation, particularly, because of the diversity observed within *BMP* genes, these genes appear to have multiple biological effects on fish which, have not yet been fully studied leaving a large number of research questions unanswered. Also, the *TNKSA* gene is still in the early stages of investigation in fish species and therefore these results pertaining to LD and haplotype association are not precise largely due to the lack of chromosomal information. However, this preliminary test was able to assist in determining whether or not LD was occurring between the genes. In this case there appeared to be no interaction between the genes which was verified by the results obtained from gene-gene interaction performed in UNPHASED (Table 3.8), although this

result may change with the increase in genomic resources for the species as these results provide a basis for future studies.

Marker-assisted selection (MAS) can be based on three type of molecular markers: markers in linkage disequilibrium with a QTL (LD-MAS), markers in linkage equilibrium with quantitative trait loci (QTLs) (LE-MAS), or the causative mutation itself as in gene-assisted selection (GAS). All three types of MAS are currently being used in the livestock industries (Dekkers, 2004). However, GAS leads to the highest genetic gain, however, the identification of the gene is not easy and is resource demanding. Thus, with advances being made in high-throughput genotyping technologies allowing for the identification of thousands of SNPs (Bers et al., 2010), there is a possibility of implementing LD-MAS directly using significant SNPs from genome-wide association studies. However, unlike SSRs (multi-allelic) which have a polymorphism information content (PIC), SNPs are usually bi-allelic, meaning each marker provides less PIC, thus marker density must be increased. Various studies have promoted the use of haplotype-based analysis rather than single marker analysis as a means to overcome this limitation (Lu et al., 2012). A haplotype-based approach, however would not only assist in overcoming this limitation but could capture epistatic interactions between SNPs at a locus (Bardel et al., 2005; Clark, 2004), explain the exact biological role played by neighbouring amino-acids on a protein structure (Clark, 2004), reduce the number of tests and hence the type I error rate (Zhao et al., 2007), and provide more power than single marker when an allelic series exists at a locus (Hamblin and Jannink, 2011; Gawenda et al., 2015). Simulation studies (Calus et al., 2009; Grapes et al., 2004; Hayes et al., 2007), have already shown that in comparison to the individual SNP approach, a haplotype-based approach does improve the prediction accuracy (Cuyabano et al., 2014; Jonas et al., 2016; Ferdosi et al., 2016), which could be extremely beneficial when applied to the breeding programmes of dusky kob. Despite all the advantages of a haplotype-based approach a few studies have determined there to be no advantage in using one approach over another (Clark et al., 1998; Zhao et al., 2007b). However, this does require further investigation, which could potentially be performed utilising the significant haplotypes identified in this study as a basis for future studies in dusky kob.

The appearance traits in fish are known to affect consumer acceptance at the point of sale, with body shape and skin pigmentation being the most significant factors affecting a farm's profitability (Colihueque and Araneda, 2014). As a result, numerous studies have been conducted to determine the implications of the conditioning factor on traits affecting body shape as the fish length-weight relationship can provide important ecological insights

(Froese, 2006). Insights which can be useful for estimating community biomass or weight when only length and species data are available. This information is essential for fishery management, stock assessment, and conservation (Froese, 1998; Jellyman et al., 2013; Oscoz et al., 2005). When looking at the conditioning factor there is often an expectation that the genes influencing traits such as weight and length, from which the K is derived, may be similar and that the selection for growth, will lead to the indirect selection of K. However, this does not appear to be true, as the low underlying genetic correlations observed between the quantitative traits and K (Fishback et al., 2002; Martyniuk et al., 2003; Nilsson, 1994; Vandeputte et al., 2004). This suggests a non-linear relationship, indicating that a more complex interaction has not been accounted for, which is most likely due to epistatic or pleiotropic effects, which operates in a non-additive manner. Therefore, looking at the correlations between the K of dusky kob and the phenotypic growth traits (Figure S3.1), the conditioning factor was not found to be significantly correlated in a linear manner with either weight and length (Table 3.9), with p-values of 0.987 and 0.247 for weight and length, respectively (Figure S3.2). This observation can be attributed to the fact that the growth in dusky kob, as in most fish species, is non-linear, such that body mass at a given body length can differentially increase or decrease throughout an individual's life time. Therefore, although the growth curve historically has been estimated using a linear regression approach (Huxley, 1924; Huxley, 1932; Le Cren, 1951; Cone, 1989), a generalised non-linear approach and graphical methods are required for fish stock analysis particularly when looking at the conditioning factor (Akamine, 2009; De Giosa and Czerniejewski, 2016; Xin'an and Aijun, 2016).

## 3.5) Conclusion

Currently, the vast majority of the aquaculture facilities in South Africa are still making use of ineffective phenotypic based selection procedures which, have hindered the development of species. This has largely been due to the limited number of genomic resources available for the non-model organism dusky kob. However, this study represents one of the first in aquaculture to employ a candidate gene approach in the identification of novel and previously uncharacterised genetic variants associated with growth rate in a non-model species. Three of the 15 candidate genes assessed using both case-control and quantitative analyses, yielded significant variants associated with the growth rate of dusky kob. In conclusion, the work presented here yields important findings pertaining to the biological role that genes play in the development and growth of dusky kob, information which can be utilised for the effective management and utilisation of the species. The sample size remains

to be one of the most important limitations, requiring follow up in a larger sample for validation. Also, further assessment of the haplotypes is required as the utilisation of these haplotypes in the development of a MAS breeding programme can drastically reduce the generational interval of the species, while simultaneously improving the accuracy of broodstock selection. This is because MAS allows for the implementation of early selection procedures, which accelerates the improvement of the species through the genetic gain of commercially favourable traits. Overall, the candidate gene approach proved to be an effective method for the development of SNP markers in non-model species, providing a basis for all future studies aiding in the development of molecular markers.

## References

Akamine, T., 2009. Non-linear and Graphical Methods for Fish Stock Analysis with Statistical Modeling. Aqua-BioScience Monographs, 2(3). https://doi.org/10.5047/absm.2009.00203.0001

Barrett, J.C., Fry, B., Maller, J., Daly, M.J., 2005. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics, 21, 263–265.

Bardel, C., Danjean, V., Hugot, J.P., Darlu, P., Genin, E., 2005. On the use of haplotype phylogeny to detect disease susceptibility loci. BMC Genetics. 6(1), 24.

Bers, N., Oers, K., Kerstens, H., Dibbits, B., Crooijmans, R., Visser, M., Groenen, M., 2010. Genome-wide SNP detection in the great tit *Parus majorusing* high throughput sequencing. Molecular Ecology, 19, 89-99. https://doi.org/10.1111/j.1365-294X.2009.04486.x

Blaise, O., Weil, C., Le Bali, P., 1995. Role of IGF-I in the control of GH secretion in Rainbow trout. Growth regulation. 5: 142-150.

Brazil, D., Church, R., Surae, S., Godson, C. and Martin, F., 2015. BMP signalling: agony and antagony in the family. Trends in Cell Biology, 25(5), 249-264. https://doi.org/10.1016/j.tcb.2014.12.004

Caldarone, E. M., MacLean, S. A., Sharack, B., 2012. Evaluation of bioelectrical impedance analysis and Fulton's condition factor as nonlethal techniques for estimating short-term responses in postsmolt Atlantic salmon (*Salmo salar*) to food availability. Fishery Bulletin. 110 (2), 257-270.

Calus, M., Meuwissen, T., Windig, J., Knol, E., Schrooten, C., Vereijken, A., 2009. Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy

of QTL mapping and prediction of genomic breeding values. Genet Sel Evol. 41(1), 11.

Carr, R., 2012. XLStatistics 12.11.22. XLent Works, Australia.

Chen, L., Dong, C., Kong, S., Zhang, J., Li, X., Xu, P., 2017. Genome wide identification, phylogeny, and expression of bone morphogenetic protein genes in tetraploidized common carp (*Cyprinus carpio*). Gene, 627, 157–163. https://doi.org/10.1016/j.gene.2017.06.020

Chiang, Y., Hsiao, S., Yver, D., Cushman, S., Tessarollo, L., Smith, S., Hodes, R., 2008. Tankyrase 1 and Tankyrase 2 Are Essential but Redundant for Mouse Embryonic Development. PLoS ONE, 3(7), 2639. https://doi.org/10.1371/journal.pone.0002639

Clark, A.G., Weiss, K.M., Nickerson, D.A., Taylor, S.L., Buchanan, A., Stengard, J., 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. American Journal of Human Genetics. 1998; 63(2):595–612. https://doi.org/10.1086/301977

Clark, A.G., 2004. The role of haplotypes in candidate gene studies. Genetic Epidemiology. 27(3), 321–33. https://doi.org/10.1002/gepi.20025

Colihueque, N., Araneda, C., 2014. Appearance traits in fish farming: progress from classical genetics to genomics, providing insight into current and potential genetic improvement. Frontiers in Genetics, 5. https://doi.org/10.3389/fgene.2014.00251

Cone, R.S., 1989. The need to reconsider the use of condition indices in fishery science. Trans. Am. Fish. Soc. 118, 510-514.

Cuyabano, B.C.D., Su, G., Lund, M.S., 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics. 15(1), 1171.

Davidson, E. H., 2010. Emerging properties of animal gene regulatory networks. Nature. 468, 911–920.

Deane, E., Woo, N., 2008. Modulation of fish growth hormone levels by salinity, temperature, pollutants and aquaculture related stress: a review. Reviews in Fish Biology and Fisheries, 19(1), 97-120. https://doi.org/10.1007/s11160-008-9091-0

De Giosa, M., Czerniejewski, P., 2016. A generalized, nonlinear regression approach to the length-weight relationship of European perch (*Perca fluviatilis L.)* from the Polish coast of the southern Baltic Sea. Archives of Polish Fisheries, 24(4), 169-175. https:///doi.org/10.1515/aopf-2016-0014

Dekkers, J.C., 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. J Anim Sci. 82: 313-328. https://doi.org/10.2527/2004.8213_supplE313x

De-Santis, C., Jerry, D., 2007. Candidate growth genes in finfish — Where should we be looking?. Aquaculture, 272(1-4), 22-38. https://doi.org/10.1016/j.aquaculture.2007.08.036

Du, S., Gong, Z., Fletcher, G., Shears, M., King, M., Idler, D., Hew, C., 1992. Growth Enhancement in Transgenic Atlantic Salmon by the Use of an “All Fish” Chimeric Growth Hormone Gene Construct. Nature Biotechnology, 10(2), 176-181. https://doi.org/10.1038/nbt0292-176

Dudbridge, F., 2003. Pedigree disequilibrium tests for multilocus haplotypes. Genet Epidemiol. 25, 115-21.

Dudbridge, F., 2008. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. Hum Hered. 66, 87-98.

Dunn, O.J., 1961. Multiple comparisons among means. J Am Stat Assoc 56:52–64 Goudet, J., Raymond, M., de Meeüs, T., Rousset, F., 1996. Testing differentiation in diploid populations. Genetics. 144:1933–1940.

Elliott, M., Hemingway, K. L., Costello, M. J., Duhamel, S., Hostens, K., Labropoulou, M., Marshall, S., Winkler, H., 2002. Links between fish and other trophic levels. In Fishes in. 124– 216.

Ferdosi, M.H., Henshall, J., Tier, B., 2016. Study of the optimum haplotype length to build genomic relationship matrices. Genet Sel Evol. 48(1), 75. https://doi.org/10.1186/s12711-016-0253-6

Fessehaye, Y., Bovenhuis, H., Rezk, M., Crooijmans, R., van Arendonk, J. and Komen, H., 2009. Effects of relatedness and inbreeding on reproductive success of Nile tilapia (*Oreochromis niloticus*). Aquaculture, 294(3-4), 180-186. https://doi.org/10.1016/j.aquaculture.2009.06.001

Fishback, A., Danzmann, R., Ferguson, M., Gibson, J., 2002. Estimates of genetic parameters and genotype by environment interactions for growth traits of rainbow trout (*Oncorhynchus mykiss*) as inferred using molecular pedigrees. Aquaculture, 206(3-4), 137-150. https://doi.org/10.1016/S0044-8486(01)00707-4

Froese, R., 1998. Length-weight relationships for 18 less-studied fish species. J. Appl. Ichthyol. 14, 117-118.

Froese, R., 2006. Cube law, condition factor and weight–length relationships: history, meta-analysis and recommendations. J. Appl. Ichthyol. 22, 241-253.

Fuentes, E., Valdés, J., Molina, A., Björnsson, B., 2013. Regulation of skeletal muscle growth in fish by the growth hormone – Insulin-like growth factor system. General and Comparative Endocrinology, 192, 136-148. https://doi.org/10.1016/j.ygcen.2013.06.009

Galloway, T.F., Bardal, T., Kvam, S.N., Wiborg, S., Gaute, D., Nesse, J., Randøl, M., Kjørsvik, E., Andersen, O., 2006. Somite formation and expression of MyoD, myogenin and myosin in Atlantic halibut (*Hippoglossus hippoglossus L*.) embryos incubated at different temperatures: Transient asymmetric expression of MyoD. Journal of Experimental Biology 209, 2432-41. https://doi.org/10.1242/jeb.02269

Gawenda, I., Thorwarth, P., Günther, T., Ordon, F., Schmid, K.J., 2015. Genome-wide association studies in elite varieties of German winter barley using single-marker and haplotype-based methods. Plant Breeding. 134(1), 28–39.

Gjedrem, T., Robinson, N., Rye, M., 2012. The importance of selective breeding in aquaculture to meet future demands for animal protein: A review. Aquaculture. https://doi.org/10.1016/j.aquaculture.2012.04.008

Grapes, L., Dekkers, J.C.M., Rothschild, M.F., Fernando, R.L., 2004. Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. Genetics. 166(3), 1561–70.

Griffiths, M.H., 1996. Life history of the dusky kob *Argyrosomus japonicus* (Sciaenidae) off the east coast of South Africa. South African J. Mar. Sci. 17, 135–154. https://doi.org/10.2989/025776196784158653

Gawenda, I., Thorwarth, P., Günther, T., Ordon, F., Schmid, K., 2015. Genome-wide association studies in elite varieties of German winter barley using single-marker and haplotype-based methods. Plant Breeding, 134(1), 28-39. https://doi.org/10.1111/pbr.12237

Hamblin, M.T., Jannink, J.L., 2011. Factors affecting the power of haplotype markers in association studies. The Plant Genome. 4, 145–53. https://doi.org/10.3835/plantgenome2011.03.0008

Han, Z., Xiao, S., Li, W., Ye, K., 2018. The identification of growth, immune related genes and marker discovery through transcriptome in the yellow drum (*Nibea albiflora*). Genes and genomics 40(4), 1–27. https://doi.org/10.1007/s13258-018-0697-x

Hayes, B.J., Chamberlain, A.J., McPartlan, H., Macleod, I., Sethuraman, L., Goddard, M.E., 2007. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genetics Research. 89(4),215–20.

Herlin, M., Delghandi, M., Wesmajervi, M., Taggart, J., McAndrew, B. Penman, D., 2008. Analysis of the parental contribution to a group of fry from a single day of spawning from a commercial Atlantic cod (*Gadus morhua*) breeding tank. Aquaculture, 274(2-4), 218-224. https://doi.org/10.1016/j.aquaculture.2007.11.034

Hillen, J., Coscia, I., Vandeputte, M., Herten, K., Hellemans, B., Maroso, F., Vergnet, A., Allal, F., Maes, G., Volckaert, F., 2017. Estimates of genetic variability and inbreeding in experimentally selected populations of European sea bass. Aquaculture, 479, 742-749. https://doi.org/10.1016/j.aquaculture.2017.07.012

Hsiao, S. and Smith, S., 2008. Tankyrase function at telomeres, spindle poles, and beyond. Biochimie, 90(1), 83-92. https://doi.org/10.1016/j.biochi.2007.07.012

Hu, G., Gu, W., Bai, Q., Wang, B., 2013. Estimation of genetic parameters for growth traits in a breeding program for rainbow trout (*Oncorhynchus mykiss*) in China. Genetics and Molecular Research, 12(2), 1457-1467. https://doi.org/10.4238/2013.april.26.7

Huxley, J.S., 1924. Constant differential growth-ratios and their significance. Nature, 114: 895-896.

Huxley, J.S., 1932. Problems of relative growth. Metheun, London, UK, 319.

Jenkins, S.F., 2018. Genetic and phenotypic characterisation of commercial dusky kob (*Argyrosomus japonicus*) cohorts, MSc thesis. Stellenbosch University, Stellenbosch.

Jellyman, P.G., Booker, D.J., Crow, S.K., Jellyman, D.J., 2013. Does one size fit all? An evaluation of length-weight relationships for New Zealand's freshwater fish species. New Zeal. J. Mar. Fresh. 47, 450-468.

Jonas, D., Ducrocq, V., Fouilloux, M.N., Croiseau, P., 2016. Alternative haplotype construction methods for genomic evaluation. Journal of Dairy Science. 99(6), 4537–46. https:///doi.org/10.3168/jds.2015-10433

Kamenskaya, D.N., Pankova, M. V., Atopkin, D.M., Brykov, V.A., 2015. Fish growth-hormone genes: Evidence of functionality of paralogous genes in Levanidov's charr *Salvelinus levanidovi*. Mol. Biol. 49, 687–693. https://doi.org/10.1134/S002689331505009X

Kang, J., Lee, S., Park, S., Ryu, H., 2002. DNA polymorphism in the growth hormone gene and its association with weight in olive flounder *Paralichthys olivaceus*. Fisheries Science, 68(3), 494-498. https://doi.org/10.1046/j.1444-2906.2002.00453.x

Kause, A., Paananen, T., Ritola, O., Koskinen, H., 2007. Direct and indirect selection of visceral lipid weight, fillet weight, and fillet percentage in a rainbow trout breeding program. Journal of Animal Science, 85(12), 3218-3227. https://doi.org/10.2527/jas.2007-0332

Kause, A., Ritola, O., Paananen, T., Eskelinen, U., Mantysaari, E., 2003. Big and beautiful? Quantitative genetic parameters for appearance of large rainbow trout. J. Fish Biol. 62, 610–622. https://doi.org/10.1046/j.1095-8649.2003.00051.x

Kobayashi, T., Kijima, A., 2010. Effects of Inbreeding Depression in Pacific Abalone *Haliotis Discus Hannai*. Journal of Shellfish Research, 29(3), 643-649. https://doi.org/10.2983/035.029.0313

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets, Molecular Biology and Evolution, 33 (7), 1870–1874, https://doi.org/10.1093/molbev/msw054

Kwon, J.M., Goate, A.M., 2000. The Candidate Gene Approach. https://doi.org/10.1007/springerreference_34556

Le Cren, E.D., 1951. The length–weight relationship and seasonal cycle in gonad weight and condition in the perch (*Perca fluviatilis*). J. Anim. Ecol. 20, 201-219.

Li, Z., Chen, F., Huang, C., Zheng, W., Yu, C., Cheng, H, Zhou, R., 2017. Genome-wide mapping and characterization of microsatellites in the swamp eel genome. Scientific Reports, 7(1). https://doi.org/10.1038/s41598-017-03330-7

Li, N., Zhou, T., Geng, X., Jin, Y., Wang, X., Liu, S., Xu, X., Gao, D., Li, Q., Liu, Z., 2018. Identification of novel genes significantly affecting growth in catfish through GWAS analysis. Mol. Genet. Genomics 293, 587–599. https://doi.org/10.1007/s00438-017-1406-1

Liu, K., Muse, S.V., 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics. 21: 2128–2129.

Liu, F., Sun, F., Xia, J.H., Li, J., Fu, G.H., Lin, G., Tu, R.J., Wan, Z.Y., Quek, D., Yue, G.H., 2014. A genome scan revealed significant associations of growth traits with a major QTL and GHR2 in tilapia. Sci. Rep. 4, 1–9. https://doi.org/10.1038/srep07256

Lu, Y., Xu, J., Yuan, Z., Hao, Z., Xie, C., Li, X., Shah, T., Lan, H., Zhang, S., Rong, T. and Xu, Y. 2011. Comparative LD mapping using single SNPs and haplotypes identifies QTL for plant height and biomass as secondary traits of drought tolerance in maize. Molecular Breeding, 30(1), 407-418.

Luo, D., McKeown, B., Rivier, J., Vale, W., 1990. In vitro responses of rainbow trout (*Oncorhynchus mykiss*) somatotrophs to carp growth hormone-releasing factor (GRF) and somatostatin. General and Comparative Endocrinology, 80(2), 288-298. https://doi.org/10.1016/0016-6480(90)90173-j

Macqueen, D.J., Johnston, I.A., 2006. A novel salmonid myoD gene is distinctly regulated during development and probably arose by duplication after the genome tetraploidization. FEBS Letters. 580, 4996–5002

Maegawa, N., Kawamura, K., Hirose, M., Yajima, H., Takakura, Y., Ohgushi, H., 2007. Enhancement of osteoblastic differentiation of mesenchymal stromal cells cultured by selective combination of bone morphogenetic protein-2 (*BMP-2*) and fibroblast growth factor-2 (*FGF-2*). Journal of Tissue Engineering and Regenerative Medicine, 1(4), 306-313. https://doi.org/10.1002/term.41

Martyniuk, C., Perry, G., Mogahadam, H., Ferguson, M., Danzmann, R., 2003. The genetic architecture of correlations among growth-related traits and male age at maturation in rainbow trout. Journal of Fish Biology, 63(3), 746-764. https://doi.org/10.1046/j.1095-8649.2003.00188.x

Melamed, P., Eliahu, N., Ofir, M., Levavi-Sivan, B., Smal, J., Rentier-Delrue, F., Yaron, Z., 1995. The effects of gonadal development and sex steroids on growth hormone secretion in the male tilapia hybrid (*Oreochromis niloticus O. aureus*). *Fish* Physiology and Biochemistry, 14(4), 267-277.https://doi.org/10.1007/bf00004065

Moss, D., Arce, S., Otoshi, C., Doyle, R. and Moss, S. (2019). Effects of inbreeding on survival and growth of Pacific white shrimp Penaeus (*Litopenaeus*) vannamei. 272: 30-37. https://doi.org/10.1016/j.aquaculture.2007.08.014

Mozsár, A., Boros, G., Sály, P., Antal, L., Nagy, S., 2014. Relationship between Fulton's condition factor and proximate body composition in three freshwater fish species. Journal of Applied Ichthyology, 31(2), 315-320. https://doi.org/10.1111/jai.12658

Muchlisin, Z., Musman, M., Siti Azizah, M., 2010. Length-weight relationships and condition factors of two threatened fishes, *Rasbora tawarensis* and *Poropuntius tawarensis*, endemic to Lake Laut Tawar, Aceh Province, Indonesia. Journal of Applied Ichthyology, 26(6), 949-953. https://doi.org/10.1111/j.1439-0426.2010.01524.x

Nielsen, D.M., Weir, B.S., 1999. A classical setting for associations between markers and loci affecting quantitative traits. Genet Res 74:271–277.

Nilsson, J., 1994. Genetics of Growth of Juvenile Arctic Char. Trans. Am. Fish. Soc. 123, 430–434. https://doi.org/10.1577/1548-8659(1994)123<0430:GOGOJA>2.3.CO;2

Opazo, R., Valladares, L., Romero, J., 2017. Comparison of gene expression patterns of key growth genes between different rate growths in zebrafish (*Danio rerio*) siblings. Lat. Am. J. Aquat. Res. 45, 766–775. https://doi.org/10.3856/vol45-issue4-fulltext-12

Oscoz, J., Campos, F., Escala, M.C., 2005. Weight–length relationships of some fish species of the Iberian Peninsula. J. Appl. Ichthyol. 21, 73-74

Pante, M., Gjerde, B., McMillan, I., 2001. Effect of inbreeding on body weight at harvest in rainbow trout, *Oncorhynchus mykiss*. Aquaculture, 192(2-4), 201-211. https://doi.org/10.1016/S0044-8486(00)00467-1

Parker, D., Power, M., Swanson, P., Rivier, J., Sherwood, N., 1997. Exon Skipping in the Gene Encoding Pituitary Adenylate Cyclase-Activating Polypeptide in Salmon Alters the Expression of Two Hormones that Stimulate Growth Hormone Release1. Endocrinology, 138(1), 414-423. https://doi.org/10.1210/endo.138.1.4830

Prevosti, A., Ocaña, J., Alonso, G., 1975. Distances between populations for *Drosophila subobscura* based on chromosome arrangement frequencies. Theor Appl Genet 45:231–241.

Rafael, M., Laizé, V., Cancela, M., 2006. Identification of *Sparus aurata* bone morphogenetic protein 2: Molecular cloning, gene expression and in silico analysis of protein conserved features in vertebrates. Bone, 39(6), 1373-1381. https://doi.org/10.1016/j.bone.2006.06.021

Renaville, R., Hammadi, M., Portetelle, D., 2002. Role of the somatotropic axis in the mammalian metabolism. Domestic Animal Endocrinology, 23(1-2), 351-360. https://doi.org/10.1016/S0739-7240(02)00170-4

Richmond, J., Norris, T., Zinn, S., 2010. Re-alimentation in 109arbour seal pups: Effects on the somatotropic axis and growth rate. General and Comparative Endocrinology, 165(2), 286-292. https://doi.org/10.1016/j.ygcen.2009.07.007

Rousset, F., 2008. GENEPOP'007: a complete re-implementation of the GENEPOP software for windows and Linux. Mol Ecol Resour. 8:103–106.

Saghai-Maroof, M.A., Solima, K.M., Jorgenson, R.A., Allard, R.W., 1984. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location and population dynamics. Proc Natl Acad Sci USA 81: 8014-8018

Saillant, E., Ma, L., Wang, X., Gatlin, D.M., Gold, J.R., 2007. Heritability of juvenile growth traits in red drum (*Sciaenops ocellatus L.).* Aquacult. Res. 38, 781–788

Satten, G., Flanders, W., Yang, Q., 2001. Accounting for Unmeasured Population Substructure in Case-Control Studies of Genetic Association Using a Novel Latent-Class Model. The American Journal of Human Genetics, 68(2), 466-477. https://doi.org/10.1086/318195

Sekino, M., 2004. Relatedness inferred from microsatellite genotypes as a tool for broodstock management of Japanese flounder *Paralichthys olivaceus*. Aquaculture, 233(1-4), 163-172. https://doi.org/10.1016/j.aquaculture.2003.11.008

Slabbert, R., 2010. Identification of growth-related quantitative trait loci within the abalone, Haliotis midae, using comparative microsatellite bulked segregant analysis. PhD Dissertation, University of Stellenbosch, Stellenbosch, South Africa.

Sekelsky, J.J., Newfeld, S.J., Raftery, L.A., Chartoff, E.H., Gelbart, W.M., 1995. Genetic characterization and cloning of mothers against dpp, a gene required for decapentaplegic function in *Drosophila melanogaster*. Genetics. 139(3),1347-58.

Smith, S., 1998. Tankyrase, a Poly(ADP-Ribose) Polymerase at Human Telomeres. Science, 282(5393), 1484-1487. https://doi.org/10.1126/science.282.5393.1484

Sole, X., Guino, E., Valls, J., Iniesta, R., Moreno, V., 2006. SNPStats: a web tool for the analysis of association studies. Bioinformatics. 22(15), 1928-1929.

Takeuchi, F., Yanai, K., Morii, T., Ishinaga, Y., Taniguchi-Yanai, K., Nagano, S., Kato, N., 2005. Linkage Disequilibrium Grouping of Single Nucleotide Polymorphisms (SNPs)

Reflecting Haplotype Phylogeny for Efficient Selection of Tag SNPs. Genetics, 170(1), 291-304. https://doi.org/10.1534/genetics.104.038232

Tan, X., Du, S.J., 2002. Differential expression of two MyoD genes in fast and slow muscles of gilthead seabream (*Sparus aurata*). Dev Genes Evol. 212(5), 207-17.

Tao, W., Boulding, E., 2003. Associations between single nucleotide polymorphisms in candidate genes and growth rate in Arctic charr (*Salvelinus alpinus L*.). Heredity, 91(1), 60-69. https://doi.org/10.1038/sj.hdy.6800281

Thomson, J.D., Higgins, D.G., Gibson, T.J., 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequences weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673-4680.

Tian, C., Gregersen, P., Seldin, M., 2008. Accounting for ancestry: population substructure and genome-wide association studies. Human Molecular Genetics, 17(2), 143-150. https://doi.org/10.1093/hmg/ddn268

Tsai, H., Hamilton, A., Guy, D., Tinch, A., Bishop, S., Houston, R., 2015. The genetic architecture of growth and fillet traits in farmed Atlantic salmon (*Salmo salar*). BMC Genetics, 16(1). https://doi.org/10.1186/s12863-015-0215-y

Vandeputte, M., Kocour, M., Mauger, S., Dupont-Nivet, M., De Guerry, D., Rodina, M., Gela, D., Vallod, D., Chevassus, B., Linhart, O., 2004. Heritability estimates for growth-related traits using microsatellite parentage assignment in juvenile common carp (*Cyprinus carpio L*.). Aquaculture, 235(1-4),.223-236. https://doi.org/10.1016/j.aquaculture.2003.12.019

Vaughan, J., Rivier, J., Spiess, J., Peng, C., Chang, J., Peter, R., Vale, W., 1992. Isolation and Characterization of Hypothalamic Growth-Hormone Releasing Factor from Common Carp, *Cyprinus carpio*. Neuroendocrinology, 56(4), 539-549. https://doi.org/10.1159/000126272

Wang, R., Green, J., Wang, Z., Deng, Y., Qiao, M., Peabody, M., Zhang, Q., Ye, J., Yan, Z., Denduluri, S., Idowu, O., Li, M., Shen, C., Hu, A., Haydon, R., Kang, R., Mok, J., Lee, M., Luu, H. and Shi, L., 2014. Bone Morphogenetic Protein (BMP) signalling in development and human diseases. Genes & Diseases, 1(1), 87-105. https://doi.org/10.1016/j.gendis.2014.07.005

Wargelius, A., Fjelldal, P., Benedet, S., Hansen, T., Björnsson, B., Nordgarden, U., 2005. A peak in gh-receptor expression is associated with growth activation in Atlantic

salmon vertebrae, while upregulation of igf-I receptor expression is related to increased bone density. General and Comparative Endocrinology, 142(1-2), 163-168. https://doi.org/10.1016/j.ygcen.2004.12.005

Wigginton, JE, Cutler, DJ, Abecasis, GR. 2005. A Note on Exact Tests of Hardy-Weinberg Equilibrium. Am. J. Hum. Genet. 76: 887–893.

Wittke-Thompson, JK, Pluzhnikov, A, Cox, NJ. 2005. Rational Inferences about Departures from Hardy-Weinberg Equilibrium. Am. J. Hum. Genet. 76: 967–986.

Wozney, J., 1992. The bone morphogenetic protein family and osteogenesis. Molecular Reproduction and Development, 32(2),160-167. https://doi.org/10.1002/mrd.1080320212

Wringe, B., Devlin, R., Ferguson, M., Moghadam, H., Sakhrani, D., Danzmann, R., 2010. Growth-related quantitative trait loci in domestic and wild rainbow trout (*Oncorhynchus mykiss*). BMC Genetics, 11(1), 63. https://doi.org/10.1186/1471-2156-11-63

Wu, M., Chen, G., Li, Y., 2016. TGF-β and BMP signaling in osteoblast, skeletal development, and bone formation, homeostasis and disease. Bone Research, 4(1).

Xin'an, W., Aijun, M., 2016. Comparison of four nonlinear growth models for effective exploration of growth characteristics of turbot *Scophthalmus maximus* fish strain. African Journal of Biotechnology, 15(40), 2251-2258. https://doi.org/10.5897/ajb2016.15490

Zhang, Y., Tan, X., Zhang, P., Xu, Y., 2006. Characterization of Muscle-Regulatory Gene, MyoD, from Flounder (*Paralichthys olivaceus*) and Analysis of Its Expression Patterns During Embryogenesis. Marine Biotechnology, 8(2), 139-148. https://doi.org/10.1007/s10126-005-5042-0

Zhao, H.H., Fernando, R.L., Dekkers, J.C.M., 2007. Power and Precision of Alternate Methods for Linkage Disequilibrium Mapping of Quantitative Trait Loci. Genetics. 175(4), 1975–86. https://doi.org/10.1534/genetics.106.066480

Zhao, K., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., Tang, C., 2007b. An Arabidopsis Example of Association Mapping in Structured Samples. PLoS Genetics. 3(1), 4. https://doi.org/10.1371/journal.pgen.0030004

# CHAPTER 4

# Study Conclusions

## 4.1) Overview

Dusky kob, a marine finfish, is an emerging aquaculture species in South Africa. The shift towards aquaculture was initiated in response to the collapse of the natural populations, which has been a direct result of overexploitation of the species. In South Africa, cultured kob are currently derived from unimproved wild broodstock; but there has been a considerable amount of effort made to maintain the faster growing F1-generation animals, for the implementation of a selective breeding programme. However, this species' late onset of sexual maturity does pose a problem for early selection, as currently, selection methods are based upon phenotypic traits rather than genetic data. Thus, for the broodstock to be accurately selected, the fish are required to be of a certain age for measurement and comparison, however this approach is often time consuming, unreliable and ineffective. Hence the shift towards the use of genetic markers in the selection process of economically important marine species, as it allows for early, reliable selection without negatively impacting the health of the animal. Also, the development of markers associated with growth, will assist in understanding the biological role that genes play in the growth rate of dusky kob, which is critical for the development of effective management and improvement strategies.

Chapter 2 therefore investigated the transferability of a solution-based exome capture kit, designed for the model organism zebrafish, in the capture and sequencing of a non-model organism's, the dusky kob's, exome. The exome data were analysed and used for variant identification, specifically, single nucleotide polymorphisms (SNPs), which could assist in future marker development. Sequencing was performed using 16 individuals from a single F1 generation family. Although a recent study was performed in 2011 by Cosart et al., using a model organism's capture kit in non-model species, the study was limited in its assessment, as it only assessed the ability of a commercial cattle kit to sequence a select few regions in closely related wild bovine species. Therefore, chapter 2 is the first to assess the full capability of a commercially available exome capture kit to sequence the entire exome of a non-model species such as dusky kob, which is considerably diverged from the model organism, zebrafish.

In chapter 3, a candidate gene approach was utilised for the selection of 15 gene regions observed to be associated with growth in other aquaculture species. Following primer optimisation and SNP confirmation, an additional four families, collected from two facilities, were used for the characterisation of SNP markers associated with the growth rate in dusky kob. Association was determined using both case-control and quantitative analyses which utilised the quantitative values: body weight (W), standard length (Ls) and Fulton's condition factor (K) as well as the categorical measure, size (large or small). Lastly, to exploit the potential of a selective breeding programme, the correlations between these phenotypic measurements were evaluated by estimating Pearson's correlation.

## 4.2) Transferability of the exon-capture

The use of gene-targeted, genome-wide markers are essential for the effective utilisation and management of species in areas such as animal production, and conservation. This is because the information obtained when developing such markers can assist in understanding the genetic processes underlying complex traits, adaptation and speciation. Unfortunately, one of the greatest challenges for developing genome wide resources such as genome-wide SNPs is the fact that many commercially important eukaryotic species have large genomes, making the development of resources costly. To overcome the limitation of cost, gene targeted method's such as transcriptome sequencing have been utilised despite the methods known limitations, which includes ascertainment bias due to differential gene expression (Ozsolak and Milos, 2010). Ascertainment bias can negatively impact results as it decreases the power of tests of association between SNPs and complex traits, introducing false-positive inferences and has been shown to distort population genetic inferences (Lachance and Tishkoff, 2013). Therefore, a more flexible, gene-targeted method is required for the identification of genome-wide SNPs in multiple individuals of a non-model organism. This study thus demonstrated the usefulness of exome capture, in the discovery of genome-wide markers, with exome sequencing performed in *Argyrosomus japonicus* using a model organism's exome capture kit, zebrafish (*Danio rerio*), on the ion-torrent™ platform. With a significant amount of divergence between the study species, *A. japonicus* and *D. rerio*, the exome capture kit was able to be tested in order to determine the feasibility of using a model organism capture kit in future studies, for the development of genomic resources in resource scarce species. The efficiency of the kit required careful evaluation as studies have shown that the performance of whole exome sequencing does decline when faced with even a limited amount of divergence between species (Jones and Good, 2016; Vallender, 2011). However, by applying this method to sixteen F1-generation individuals, the capture kit was

able to successfully sequence millions of reads per sample, of which thousands were homologous exons of zebrafish as well as a large number of potentially species-specific exons. Similar success has been reported in studies where whole exome sequencing was performed using model organism's capture kits, however the species used were closely related (*e.g.* Cosart et al., 2011; Jia et al., 2013; Ryan et al., 2013). Although additional studies have been previously performed in more divergent species, the success of these studies was shown to occur as a consequence of losing comparative data (*e.g.* Faircloth et al., 2012; Lemmon et al., 2012; McCormack et al., 2012). This reported loss of comparative data was observed in this study as regions other than the expected target regions had been sequenced as a result of divergence, affecting the reproducibility of the study. However, the kit was not limited by the divergence between the species and was still able to produce a large number of usable reads with a large majority of reads being exons which may potentially be species specific, although this does require further investigation. Overall, this study was able to produce data that was more than sufficient, showing that exon capture can be customised for genome-wide SNP discovery in non-model organisms without prior information regarding the species' genome. Characterisation of the exon regions was able to determine that the exons were distributed among the various functional classes of GO and KOG databases indicating how the exome data, even though not covering the entire genome, encompasses a broad gene functional diversity. The exome data was shown to be a valuable resource for the identification of variants, which assisted in the discovery of thousands of SNPs and a few thousand tandem repeats, which will greatly assist in marker development for the species.

### 4.3) SNP markers associated with growth

In general, the limited resources for non-model organisms such as dusky kob, has hindered research efforts aiming to characterise and understand genetic variation. To circumvent this issue, early molecular work on dusky kob was based on microsatellite markers that did not require prior knowledge of the species' genome (Archangi et al., 2009; Mirimin et al., 2013). Currently, microsatellite markers are still the only molecular marker being used in dusky kob for the evaluation of pedigree relationships (Liu et al., 2012; Vandeputte et al., 2014; Vandeputte and Haffray, 2014) and although these markers have been extremely useful in pedigree inferences, they are limited in regards to association. Therefore, in order to understand the genetic basis of phenotypic variation in complex traits (Tsai et al, 2015), SNP markers needed to be developed. Despite the need for gene-targeted genome wide SNPs the vast majority of the aquaculture facilities in South Africa are still making use of a

conditioning factor, for the selection of their broodstock (Kause et al., 2003). However, when dealing with complex traits such as growth rate, the implementation of genetic markers in the selection process can greatly assist in shortening generational intervals thus assisting the accelerated development of species. The results from this study showed the candidate gene approach to be an effective method for the development of SNP markers in non-model species as this approach was able to identify eight SNPs in three key genes to be associated with growth. These associations were made using both case-control and quantitative analyses, with all the markers found to be significant, reanalysed and verified using a TDT. Validation using TDTs is commonly seen in studies where the parental genotypes are available. This is because the method in which a TDT makes an association between the phenotype and SNP provides confidence in the result (McGinnis et al., 2002), showing that an observed SNP/phenotype association is not simply a result of a sampling error (Lander and Schork 1994). The linkage disequilibrium results for the three genes in which the SNPs were identified showed that the genes were not linked to one another, which was further verified using a gene-gene interaction analysis. Although no linkage appeared to be occurring between the genes, two of the three genes were identified as having a single linkage block with high LD correlations between markers within the respective genes. All these markers were in close proximity to one another, indicating their potential suitability for MAS (Abasht et al., 2009). Simple MAS often relies of the marker's mode of inheritance, in relation to the desired phenotype, which was determined for each of the markers in this study when performing the case-control analyses. Although this method may be simple, the benefits of such MAS has been observed in a number of commercial populations such as tilapia (McAndrew et al., 1988), and rainbow trout (Blanc et al., 2006). In addition to association analyses this study was able to estimate the correlation between quantitative traits: wet weight, standard length and conditioning factor. These results were comparable to a study performed by Jenkins (2018) which showed the conditioning factor to have no linear correlation to the weight or length prediction of the fish. This candidate gene study was the first step in exploring the biological pathways between genetic determinants and growth in dusky kob. The findings from these studies will significantly contribute to further biological function analysis of the identified candidate genes and potential utilisation of these markers in MAS.

## 4.4) Considerations for the implementation of MAS in the breeding programmes of dusky kob

The data generated in both experimental chapters highlights the need for gene-targeted genome wide SNPs for the improvement and effective utilisation of the species. Previously, the development of such markers would not have been feasible due to the large size of this species genome. However, chapter 2 was able to sufficiently demonstrate the use of exome capture in the identification of novel markers, which can potentially be used in the selective breeding programme of dusky kob. Over the last few years multiple authors have suggested the need for genetic markers in the selection of complex traits such as growth, disease resistance and fillet quality (Goddard and Hayes, 2009; Newton-Cheh and Hirschhorn, 2005; Midtlyng et al., 2001). Yet despite multiple studies concluding that the only reliable method for the selection of complex traits is the use of molecular markers, the majority of South Africa's aquaculture farms still rely on only phenotypic traits. Although there is a chance of success using these traditional methods, the result is not always guaranteed as decisions are made on subjective observational data rather than factual genetic information. These traditional methods may appear to be more cost effective, as genotyping and the development of markers is an expensive procedure. However, for a species such as dusky kob, which does have such a long generation interval, the long-term genetic gains and benefits that can be achieved will greatly exceed any initial apprehension. As early selection methods are not possible using phenotypic measures, slow growing individuals are spending more time within the production cycle, using resources which could be better utilised on improved, faster growing individuals that will generate a considerable amount of profit for the farm. Thus, selection procedures are extremely important, because by incorrectly selecting poor broodstock can have detrimental consequences on the farm and the industry as a whole. Utilising genetic markers in the selective breeding programme of dusky kob will enable for the accurate selection of a variety of different traits at early developmental stages. In particular, genetic parameters for traits relating to growth could be estimated early on in the production cycle to reveal wild broodstock with superior growth which can be utilised for the production of faster growing F1 individuals. The implementation of such selection methods in selective breeding programmes will allow managers to effectively manage the population by constantly producing animals from only the best performing broodstock, while continually improving the broodstock individuals through the selection of multiple economically important traits, will assist in the long-term improvement of the species.

## 4.5) Shortcomings and perspectives on future undertakings

Overall, the cross-species targeted capture method used in this study was shown to be successful despite the non-uniform coverage across the genome and the unintended sequencing of non-protein coding regions. Yet despite the significant impact this will have on the reproducibility of this study, making the consolidation of data between projects extremely difficult, these singleton and non-coding sequences do require further investigation into their nature, as these sequences could contain valuable information, which may be species specific. Thus, although the divergence between the species did affect the number of usable reads in this study, as a consequence of non-uniform coverage, which resulted in the elimination of multiple reads, the information obtained regarding these sequences' could be applied to aid not only in the development of additional markers but also in future genomics research for the species. The development of more markers in these unique regions may prove to be extremely useful in applications such as linkage mapping and marker-assisted selection (MAS).

Numerous methods have been used for the development of markers associated with traits of interest, such as growth rate, with the candidate gene approach shown to be highly successful in this study. However, this strategy does limit the study to genes of known or suggested involvement in the trait, thereby excluding the discovery of novel genes that could influence the trait of interest. Yet despite the limitations regarding novel discovery, it is a good starting point for the development of SNP markers in non-model species (Alghamdi and Padmanabhan, 2014; Holloway et al., 2017). However, the discovery of novel genes in this study was not exclusively a result of the candidate gene approach but rather a combination of this approach with the large number of eliminated sequences. Future studies should however not only investigate the large number of presumed species-specific reads but should be directed towards the development of a SNP based linkage map as the construction of such a map will be useful in the further dissection of quantitative trail loci (QTL) (Lander and Botstein, 1989). Thus, bringing new genomic insights to poorly characterised species, while assisting in understanding the complex interaction between genetic factors and environmental effects. The discovery of large numbers of new genetic markers and the construction of a dense genetic map of the dusky kob genome using both available and newly discovered markers would be of great value for future breeding programmes.

One clear limitation was the small sample sizes (*i.e.* few families and few individuals for each family) that were used. Many more samples will be needed to obtain a more accurate

correlation between the weight, length and conditioning factor of dusky kob. The conditioning factor showed no significant correlation to either quantitative measure, however these characteristics could impact the market acceptance in a developing and more competitive dusky kob industry, therefore this factor requires further investigation with studies looking into body depth (BD) and body shape index (H), as more reliable indicators of body conformation and condition. This could be promising for future studies as studies have shown a positive genetic correlation between the weight of the fish and BD or H (Domingos et al., 2013; Gjedrem and Thodesen, 2005). Estimation of genotype by environment (G x E) interactions for growth-related traits in dusky kob should also be conducted in order to fully exploit the potential of a selective breeding programme (Dupont-Nivet et al., 2008; Vandeputte et al., 2014; Vlok et al., 2016).

## 4.6) Concluding statement

This study is unique in regards to marker development as it evaluates the transferability of a model organisms exome capture kit in a non-model organism, for the development of genomic resources. Additionally, this study represents one of first attempts to develop SNP markers associated with an economically important trait of interest, growth rate, for dusky kob. The obtained results were able to identify eight associated SNPs, showing the successful combination of exome capture and a candidate gene approach in the development of SNP markers. While growth rate will always remain one of the most commercially important traits selected for in dusky kob culture, a selection programme for the species may also need to considered when expanding to include the simultaneous selection of multiple traits utilising MAS. Thus, molecular genetics is expected to play a major role in the development of aquaculture breeding programmes, particularly in dusky kob which should substantially benefit from any genetic improvement as a result of the species extremely long generation interval.

## References

Abasht, B., Sandford, E., Arango, J., Settar, P., Fulton, J., O'Sullivan, N., Hassen, A., Habier, D., Fernando, R., Dekkers, J., Lamont, S., 2009. Extent and consistency of linkage disequilibrium and identification of DNA markers for production and egg quality traits in commercial layer chicken populations. BMC Genomics, 10, p.S2. https://dx.doi.org/10.1186%2F1471-2164-10-S2-S2

Alghamdi, J., Padmanabhan, S., 2014. Fundamentals of Complex Trait Genetics and Association Studies. Handbook of Pharmacogenomics and Stratified Medicine, 235-257.

Archangi, B., Chand, V., Mather, P.B., 2009. Isolation and characterization of 15 polymorphic microsatellite DNA loci from *Argyrosomus japonicus* (mulloway), a new aquaculture species in Australia. Mol. Ecol. Resour. 9, 412–414. https://doi.org/10.1111/j.1755-0998.2008.02464.x

Blanc, J., Poisson, H., Quillet, E., 2006. A Blue Variant in the Rainbow Trout, Oncorhynchus mykiss Walbaum. Journal of Heredity, 97(1), 89-93. https://doi.org/10.1093/jhered/esj010

Cosart, T., Beja-Pereira, A., Chen, S., Ng, S.B., Shendure, J., Luikart, G., 2011. Exome-wide DNA capture and high throughput sequencing in domestic and wild species. BMC Genomics 12, 347.

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

Dupont-Nivet, M., Vandeputte, M., Vergnet, A., Merdy, O., Haffray, P., Chavanne, H., Chatain, B., 2008. Heritabilities and GxE interactions for growth in the European sea bass (*Dicentrarchus labrax L*.) using a marker-based pedigree. Aquaculture 275, 81–87. https://doi.org/10.1016/j.aquaculture.2007.12.032

Faircloth, B., McCormack, J., Crawford, N., Harvey, M., Brumfield, R., Glenn, T., 2012. Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. Systematic Biology, 61(5), 717-726. https://doi.org/10.1093/sysbio/sys004

Gjedrem, T., Thodesen, J., 2005. Selection and Breeding Programs in Aquaculture, 89–111. http://doi.org/10.1007/1-4020-3342-7_7

Goddard, M., Hayes, B., 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics, 10(6), 381-391. https://doi.org/10.1038/nrg2575

Holloway, J., Prescott, S., 2017. The Origins of Allergic Disease. Middleton's Allergy Essentials, 29-50.

Jenkins, S.F., 2018. Genetic and phenotypic characterisation of commercial dusky kob (*Argyrosomus japonicus*) cohorts, MSc thesis. Stellenbosch University, Stellenbosch.

Jia, X., Zhang, F., Bai, J., Gao, L., Zhang, X., Sun, H., Sun, D., Guan, R., Sun, W., Xu, L., Yue, Z., Yu, Y., Fu, S., 2013. Combinational analysis of linkage and exome sequencing identifies the causative mutation in a Chinese family with congenital cataract. BMC Med. Genet. 14, 107. https://doi.org/10.1186/1471-2350-14-107

Jones, M., Good, J., 2015. Targeted capture in evolutionary and ecological genomics. Molecular Ecology, 25(1), 185-202. https://doi.org/10.1111/mec.13304

Kause, A., Ritola, O., Paananen, T., Eskelinen, U., Mantysaari, E., 2003. Big and beautiful? Quantitative genetic parameters for appearance of large rainbow trout. J. Fish Biol. 62, 610–622. https://doi.org/10.1046/j.1095-8649.2003.00051.x

Lachance, J., Tishkoff, S., 2013. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays, 35(9), 780-786. https://doi.org/10.1002/bies.201300014

Lander E.S., Botstein D., 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 121, 185–99.

Lemmon, A., Emme, S., Lemmon, E., 2012. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Systematic Biology, 61(5),727-744. https://doi.org/10.1093/sysbio/sys049

Liu, P., Xia, J.H., Lin, G., Sun, F., Liu, F., Lim, H.S., Yue G.H., 2012. Molecular Parentage Analysis Is Essential in Breeding Asian Seabass. PLoS ONE 7, e51142. https://doi.org/10.1371/journal.pone.0051142

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

McCormick, M., Delaney, J., Tsuchiya, M., Tsuchiyama, S., Shemorry, A., Sim, S., Chou, A., Ahmed, U., Carr, D., Murakami, C., Schleit, J., Sutphin, G., Wasko, B., Bennett, C., Wang, A., Olsen, B., Beyer, R., Bammler, T., Prunkard, D., Johnson, S., Pennypacker, J., An, E., Anies, A., Castanza, A., Choi, E., et al., 2015. A

Comprehensive Analysis of Replicative Lifespan in 4,698 Single-Gene Deletion Strains Uncovers Conserved Mechanisms of Aging. Cell Metabolism, 22(5), 895-906. https://doi.org/10.1016/j.cmet.2015.09.008

McGinnis, R., Shifman, S., Darvasi, A., 2002. Power and Efficiency of the TDT and Case-Control Design for Association Scans. Behavior Gen, 32(2), 135-144.

Midtlyng, P.J., Storset, A., Michel, C., Slierendrecht, W.J., Okamoto, N., 2002. Breeding for disease resistance in fish. Bull. Eur. Ass. Fish Pathol., 22(2), 166.

Mirimin, L., Ruiz Guajardo, J.C., Vervalle, J., Bester-Van der Merwe, A., Kerwath, S., Macey, B., Bloomer, P., Roodt-Wilding, R., 2013. Isolation and validation of microsatellite markers from a depleted South African sciaenid species, the dusky kob (*Argyrosomus japonicus*), by means of the FIASCO/454 approach. Conserv. Genet. Resour. 5, 841–844. https://doi.org/10.1007/s12686-013-9922-8

Newton-Cheh, C., Hirschhorn, J., 2005. Genetic association studies of complex traits: design and analysis issues. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 573(1-2), 54-69. https://doi.org/10.1016/j.mrfmmm.2005.01.006

Ozsolak, F., Milos, P., 2010. RNA sequencing: advances, challenges and opportunities. Nature Reviews Genetics, 12(2), 87-98. https://doi.org/10.1038/nrg2934

Ryan, S., Willer, J., Marjoram, L., Bagwell, J., Mankiewicz, J., Leshchiner, I., Goessling, W., Bagnat, M., Katsanis, N., 2013. Rapid identification of kidney cyst mutations by whole exome sequencing in zebrafish. Development 140, 4445–4451.

Tsai, H., Hamilton, A., Guy, D., Tinch, A., Bishop, S., Houston, R., 2015. The genetic architecture of growth and fillet traits in farmed Atlantic salmon (*Salmo salar*). BMC Genetics, 16(1). https://doi.org/10.1186/s12863-015-0215-y

Vallender, E.J., 2011. Expanding whole exome resequencing into non-human primates. Genome Biol. 12, R87

Vandeputte, M., Haffray, P., 2014. Parentage assignment with genomic markers: A major advance for understanding and exploiting genetic variation of quantitative traits in farmed aquatic animals. Front. Genet. Frontiers Media SA. http://doi.org/10.3389/fgene.2014.00432

Vlok, A.C., Difford, G.F., Rhode, C., Brink, D., 2016. An Assessment of Hatchery Cohort Growth Rates of South African Abalone, *Haliotis midae*, Across Four Commercial Environments. J. World Aquac. Soc. 47, 658–666. https://doi.org/10.1111/jwas.1229

# Appendix A

## Supplementary Information Chapter 3

---

**Table S3.1.** The 15 gene regions identified through literature to be associated with growth in other aquaculture species. The genes name, gene symbol, accession number and location in the zebrafish genome is provided in the table.

**Table S3.2.** Primers designed for the 15 gene regions. Sequence shown for the reverse and forward primer in the 5'-3' orientation. The optimised annealing temperature (Ta) is indicated for each primer pair.

**Table S3.3.** Summary of the quantitative analyses performed using the FBC cohort with altered responses: (A) weight (B) conditioning factor and (C) length. The genotypes for the large and small phenotypes are depicted with the correlating statistics. The odds ratio (OR) with a confidence interval (CI) of 95%, p-value, the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) are shown for each SNP. The HWE p-value and correlating allele frequencies are indicated for each of the significant SNPs.

**Table S3.4.** Results from the association tests performed in PowerMarker. FBC cohort: Distance-based test, F-tests for weight, length, and conditioning factor, and an exact G-test.

**Figure S3.1.** Linkage disequilibrium (LD) block structures. LD block structure consisted of a total of ten SNPs in three different genes. Two SNPs were located in the *MYOD1* gene, two SNPs in the *TNKSA* gene and four SNPs in *BMP2* gene. The LD block was defined by a D' value threshold of 0.8. The colour scale ranges from red to white (colour intensity decreases with decreasing D' value, and all of D' values were = 1).

**Figure S3.2.** Scatterplots illustrating correlation analysis for Fulton's conditioning factor K versus body weight (**A**) and length (**B**). Trend line equations and R2-values are also indicated.

**Figure S3.3.** Scatterplots illustrating correlation analysis weight versus length. Trend line equations and R2-values are also indicated.

**Table S3.1. The 15 gene regions identified through literature to be associated with growth in other aquaculture species. The genes name, gene symbol, accession number and location in the zebrafish genome is provided in the table.**

| Gene | Gene Symbol | GenBank | Position |
|---|---|---|---|
| Bone morphogenetic protein 2a | *bmp2a* | NC_007128.7 | Chromosome 17 (4227274...4231388) |
| STT3 oligosaccharyl transferase complex catalytic subunit B | *sttb3* | NC_007127.7 | Chromosome 16 (39385483...39477790) |
| Growth differentiation factor 6a | *gdf6a* | NC_007127.7 | Chromosome 16 (39125608...39131666) |
| Myogenic factor 5 | *myf5* | NC_007115.7 | Chromosome 4 (21741228...21745107) |
| Myogenic factor 6 | *myf6* | NC_007115.7 | Chromosome 4 (21717793...21720943) |
| Fibroblast growth factor 4 | *fgf4* | NC_007118.7 | Chromosome 7 (54617076...54624873) |
| Growth hormone releasing hormone receptor A | *ghrhra* | NC_007113.7 | Chromosome 2 (50906106…50966124) |
| Eukaryotic translation elongation factor 1 alpha 1, like 1 | *eef1a1l1* | NC_007130.7 | Chromosome 19 (43119014…43122337) |
| Growth hormone regulated TBC protein 1a | *grtp1a* | NC_007112.7 | Chromosome 1 (184908…191971) |
| Tubulin, alpha 8 like 2 | *tuba8l2* | NC_007112.7 | Chromosome 1 (5402484...5419116) |
| Myogenic differentiation 1 | *myod1* | NC_007136.7 | Chromosome 25 (31421253...31423459) |
| Eukaryotic translation elongation factor 2, like 2 | *eef2l2* | NC_007116.7 | Chromosome 5 (41485467...41494872) |
| Tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase a | *tnksa* | NC_007132.7 | Chromosome 21 (19926061...20110923) |
| Cadherin 4, type 1, R-cadherin | *cdh4* | NC_007122.7 | Chromosome 11 (20371182...20826753) |
| Clock circadian regulator a | *clocka* | NC_007131.7 | Chromosome 20 (22068860...22176005) |

**Table S3.2. Primers designed for the 15 gene regions. Sequence shown for the reverse and forward primer in the 5'-3' orientation. The optimised annealing temperature (Ta) is indicated for each primer pair**

| Name of primer | Sequence (5'-3' orientation) | Number of bases | Optimised Ta |
|---|---|---|---|
| Ajap_cdh4_F | TTCACAGGCATGGTTCTGCT | 20 | 54°C |
| Ajap_cdh4_R | GGGTACTGCGACTCAGCATT | 20 | |
| Ajap_myf5_F | GGACATTGCCTCCAGTGGG | 19 | 60°C |
| Ajap_myf5 R | CGTCAGAGCAGTTGGACAGT | 20 | |
| Ajap_myf6_F | TCTGCAAGAGGAAGTCAGCG | 20 | 60°C |
| Ajap_myf6_R | GTCTTCTCCTGCTCGTCCAG | 20 | |
| Ajap_stt3b_F | TTGTCGTAGAGGTTTCCGGC | 20 | 56°C |
| Ajap_stt3b_R | TCTGTTTCCTTCCAGTGGCC | 20 | |
| Ajap_fgf4_F | CAGTCCGTCAGAACCGTAGC | 20 | 60°C |
| Ajap_fgf4_R | TACCAGCCAACACAACAGCA | 20 | |
| Ajap_bmp2_F | TCCCTCCACCACCATATCCT | 20 | 60°C |
| Ajap_bmp2_R | CCCTGGTTACGAGGCCTTTT | 20 | |
| Ajap_clocka_F | GCACTCGTCTTCTCCACAGT | 20 | 56°C |
| Ajap_clocka_R | GCTGTATGATGCTGCTGTTGA | 21 | |
| Ajap_tnksa_F | CGGAGGTGTCTTCAGCAGAT | 20 | 60°C |
| Ajap_tnksa_R | CGCTCGTTGTGATGGTTGTG | 20 | |
| Ajap_eef1a1l1_F | GGCCTTCATCCATTTCCCCA | 20 | 54°C |
| Ajap_eef1a1l1_R | AGGAGGGTAGTTGGAGAAGCT | 21 | |
| Ajap_ghrhra_F | ACGATGTGGTCCATTGCAGT | 20 | 54°C |
| Ajap_ghrhra_R | CGTCCAACCGAAACAGATGC | 20 | |
| Ajap_grtp1a_F | AGTCCTCTCAGCCAATCGC | 19 | 54°C |
| Ajap_grtp1a_R | GCACAGAACCTTCCCAGACA | 20 | |
| Ajap_gdf6_F | CTACCTGCACCCACACTGC | 19 | 58°C |
| Ajap_gdf6_R | GAGACACGGCAAGAAGTCCA | 20 | |
| Ajap_myod1_F | TCACCATGCCATCAGAGCAG | 20 | 60°C |
| Ajap_myod1_R | CAAGGCCTGCAAGAGGAAGA | 20 | |
| Ajap_tuba8l2_F | AACCTGAACCGCCTCATCAG | 20 | 60°C |
| Ajap_tuba8l2_R | CAGTGGGAGGCTGGTAGTTG | 20 | |
| Ajap_eef2l2_F | TGATGATGGGCCGGTATGTG | 20 | 56°C |
| Ajap_eef2l2_R | GCAAGCATGATCCTCCTCCA | 20 | |

**Table S3.3. Summary of the quantitative analyses performed using the FBC cohort with altered responses: (A) weight (B) conditioning factor and (C) length. The genotypes for the large and small phenotypes are depicted with the correlating statistics. The odds ratio (OR) with a confidence interval (CI) of 95%, p-value, the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) are shown for each SNP. The HWE P-value and correlating allele frequencies are indicated for each of the significant SNP**

**A**

**Weight**

| SNP | Model | Genotype | Large | Small | OR (95% CI) | p-value | AIC | BIC | Allele | Allele frequencies | | | HWE P-value | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | All | Large | Small | Large | Small |
| *tnksa-1* | Dominant | C/C | 53 | 461.08 (32.61) | 0 | 0.001 | 1090.7 | 1097.8 | C | 0.7 | 0.96 | 0.44 | <0.0001 | 1.000 |
| | | C/T-T/T | 27 | 292.96 (31.43) | -168.11 (-268.00 - -68.23) | | | | T | 0.3 | 0.04 | 0.56 | | |
| *tnksa-3* | Dominant | T/T | 58 | 464.84(31.16) | 0 | 0.0001 | 1084.6 | 1091.7 | T | 0.81 | 1 | 0.62 | 0.170 | 1.000 |
| | | C/T-C/C | 22 | 244.82 (17.8) | -220.03 (-321.84 - -118.22) | | | | C | 0.19 | 0 | 0.38 | | |
| *fgf4-2* | --- | C/C | 56 | 346.29 (29.72) | 0 | 0.0005 | 1087.9 | 1095.1 | C | 0.85 | 0.72 | 0.98 | 1.000 | 0.020 |
| | | G/C | 24 | 539.79 (37.57) | 193.51 (92.21 – 294.80) | | | | G | 0.15 | 0.28 | 0.02 | | |
| *fgf4-3* | --- | G/G | 60 | 354.55 (28.58) | 0 | 0.0005 | 1088.7 | 1095.9 | G | 0.88 | 0.76 | 0.99 | 1.000 | 0.082 |
| | | G/C | 20 | 553.7 (41.13) | 199.15 (91.43 – 306.87) | | | | C | 0.12 | 0.24 | 0.01 | | |
| *myod1-1* | Dominant | G/G | 48 | 478.46 (35.06) | 0 | 0.0002 | 1087.2 | 1094.4 | C | 0.61 | 0.56 | 0.65 | 0.037 | 0.520 |
| | | G/T-T/T | 32 | 293.16 (26.51) | -185.30 (-279.64 - -90.96) | | | | G | 0.39 | 0.44 | 0.35 | | |
| *myod1-3* | Over dominant | G/G-T/T | 56 | 470.77 (30.91) | 0 | <0.0001 | 1083.3 | 1090.5 | T | 0.78 | 0.8 | 0.75 | 0.400 | 1.000 |
| | | G/T | 24 | 249.33 (25.6) | -221.43 (-319.87 - -123.00) | | | | C | 0.22 | 0.2 | 0.25 | | |
| *bmp2-1* | Dominant | T/T | 42 | 501.9 (36.08) | 0 | <0.0001 | 1082.9 | 1090.1 | T | 0.57 | 0.78 | 0.36 | <0.0001 | <0.0001 |
| | | T/G-G/G | 38 | 296.5 (27.28) | -205.40 (-295.49 - -115.32) | | | | G | 0.43 | 0.22 | 0.64 | | |
| *bmp2-5* | Over dominant | G/G-C/C | 48 | 462.94 (36.2) | 0.00 | 0.0003 | 1092.7 | 1099.9 | G | 0.65 | 0.74 | 0.56 | 0.350 | 0.011 |
| | | C/G | 32 | 316.44 (27.68) | -146.50 (-244.14 - -48.86) | | | | C | 0.35 | 0.26 | 0.44 | | |
| *bmp2-8* | Over dominant | G/G-C/C | 34 | 507.15 (43.91) | 0 | 0.0004 | 1088 | 1095.2 | G | 0.61 | 0.7 | 0.52 | 0.004 | 0.720 |
| | | C/G | 46 | 328.35 (25.45) | -178.80 (-272.75 - -84.84) | | | | C | 0.39 | 0.3 | 0.48 | | |
| *bmp2-10* | Recessive | A/A-A/G | 62 | 353.89 (26.49) | 0 | 0.0004 | 1086.3 | 1093.4 | G | 0.5 | 0.6 | 0.4 | 0.007 | 0.330 |
| | | G/G | 18 | 578.11 (50.27) | 224.22 (114.22 – 334.23) | | | | A | 0.5 | 0.4 | 0.6 | | |
| *bmp2-11* | Codominant | T/T | 22 | 530.18 (49.53) | 0.00 | 0.0003 | 1091.3 | 1100.8 | T | 0.59 | 0.69 | 0.5 | 0.000 | 0.720 |
| | | C/T | 51 | 342.12 (25.15) | -188.06 (-295.56 - -80.57) | | | | C | 0.41 | 0.31 | 0.5 | | |
| | | C/C | 7 | 462.14 (139.16) | -68.04 (-250.92 - 114.84) | | | | | | | | | |
| *bmp2-15* | Dominant | C/C | 23 | 538.48 (41.1) | 0 | 0.0006 | 1089 | 1096.2 | C | 0.6 | 0.72 | 0.48 | 0.000 | 0.690 |
| | | C/G-G/G | 57 | 350.21 (29.05) | -188.27 (-291.53 - -85.00) | | | | G | 0.4 | 0.28 | 0.52 | | |

**B**

| SNP | Model | Genotype | Conditioning Factor: Large | Small | OR (95% CI) | p-value | AIC | BIC | Allele | Allele frequencies: All | Large | Small | HWE P-value: Large | Small |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *tnksa-1* | Codominant | C/C | 53 | 1.32 (0.03) | 0 | 0.0009 | -19.6 | -10 | C | 0.7 | 0.96 | 0.44 | <0.0001 | 1.000 |
| | | C/T | 6 | 1.37 (0.1) | 0.05 (-0.12 - 0.23) | | | | T | 0.3 | 0.04 | 0.56 | | |
| | | T/T | 21 | 1.12 (0.05) | -0.20 (-0.30 - -0.09) | | | | | | | | | |
| *tnksa-3* | Dominant | T/T | 58 | 1.32 (0.03) | 0 | 0.00038 | -15.8 | -8.6 | T | 0.81 | 1 | 0.62 | 0.170 | 1.000 |
| | | C/T-C/C | 22 | 1.16 (0.06) | -0.16 (-0.26 - -0.05) | | | | C | 0.19 | 0 | 0.38 | | |
| *myod1-1* | Codominant | G/G | 48 | 1.33 (0.03) | 0 | 0.00087 | -15 | -5.4 | C | 0.61 | 0.56 | 0.65 | 0.037 | 0.520 |
| | | G/T | 31 | 1.21 (0.04) | -0.12 (-0.21 - -0.02) | | | | G | 0.39 | 0.44 | 0.35 | | |
| | | T/T | 1 | 0.83 (0) | -0.49 (-0.92 - -0.07) | | | | | | | | | |
| *myod1-2* | Recessive | C/C-C/G | 63 | 1.24 (0.03) | 0 | 0.00057 | -15 | -7.9 | G | 0.82 | 0.91 | 0.74 | 0.039 | 0.017 |
| | | G/G | 17 | 1.41 (0.04) | 0.17 (0.05 - 0.28) | | | | T | 0.18 | 0.09 | 0.26 | | |
| *myod1-3* | Recessive | G/G-G/T | 78 | 1.28 (0.03) | 0 | 0.0011 | -9.8 | -2.6 | T | 0.78 | 0.8 | 0.75 | 0.400 | 1.000 |
| | | T/T | 2 | 1.02 (0.01) | -0.26 (-0.57 - 0.05) | | | | C | 0.22 | 0.2 | 0.25 | | |
| *myod1-4* | Recessive | T/T-C/T | 78 | 1.26 (0.02) | 0 | 0.001 | -13.9 | -6.8 | T | 0.74 | 0.79 | 0.7 | 0.072 | 0.160 |
| | | C/C | 2 | 1.67 (0.14) | 0.41 (0.10 - 0.71) | | | | C | 0.26 | 0.21 | 0.3 | | |
| *bmp2-1* | Codominant | T/T | 42 | 1.31 (0.03) | 0 | 0.001 | -18.3 | -8.8 | T | 0.57 | 0.78 | 0.36 | <0.0001 | <0.0001 |
| | | T/G | 7 | 1 (0.09) | -0.31 (-0.48 - -0.15) | | | | G | 0.43 | 0.22 | 0.64 | | |
| | | G/G | 31 | 1.29 (0.04) | -0.02 (-0.12 - 0.08) | | | | | | | | | |
| *bmp2-5* | Over dominant | G/G-C/C | 48 | 1.32 (0.03) | 0 | 0.001 | -13.1 | -6 | G | 0.65 | 0.74 | 0.56 | 0.350 | 0.011 |
| | | C/G | 32 | 1.2 (0.04) | -0.12 (-0.22 - -0.03) | | | | C | 0.35 | 0.26 | 0.44 | | |
| *bmp2-6* | Codominant | T/T | 28 | 1.29 (0.03) | 0 | 0.001 | -14.2 | -4.7 | T | 0.63 | 0.72 | 0.54 | 0.054 | 0.690 |
| | | C/T | 45 | 1.23 (0.03) | -0.06 (-0.16 - 0.04) | | | | C | 0.37 | 0.28 | 0.46 | | |
| | | C/C | 7 | 1.49 (0.09) | 0.21 (0.03 - 0.38) | | | | | | | | | |
| *bmp2-11* | Dominant | T/T | 25 | 1.2 (0.05) | 0 | 0.00047 | -11.2 | -4.1 | T | 0.59 | 0.69 | 0.5 | 0.000 | 0.720 |
| | | C/T-C/C | 55 | 1.31 (0.03) | 0.11 (0.00 - 0.21) | | | | C | 0.41 | 0.31 | 0.5 | | |
| *bmp2-14* | Recessive | A/A-A/G | 71 | 1.29 (0.03) | 0 | 0.0002 | -12.7 | -5.6 | A | 0.53 | 0.6 | 0.46 | <0.0001 | 0.190 |
| | | G/G | 9 | 1.11 (0.04) | -0.18 (-0.33 - -0.03) | | | | G | 0.47 | 0.4 | 0.54 | | |
| *bmp2-15* | Recessive | C/C-C/G | 73 | 1.29 (0.03) | 0 | 0.00037 | -11.6 | -4.5 | C | 0.6 | 0.72 | 0.48 | 0.000 | 0.690 |
| | | G/G | 7 | 1.11 (0.09) | -0.18 (-0.35 - -0.01) | | | | G | 0.4 | 0.28 | 0.52 | | |

**C**

**Length**

| SNP | Model | Genotype | Large | Small | OR (95% CI) | p-value | AIC | BIC | Allele | Allele frequencies | | | HWE P-value | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | All | Large | Small | Large | Small |
| *tnksa-3* | Over dominant | T/T-C/C | 66 | 317.58 (7.59) | 0 | 0.001 | 885.7 | 892.8 | T | 0.81 | 1 | 0.62 | 0.170 | 1.000 |
| | | C/T | 14 | 260.36 (13.26) | -57.22 (-91.73 - -22.71) | | | | C | 0.19 | 0 | 0.38 | | |
| *fgf4-2* | --- | C/C | 56 | 292.59 (8.56) | 0 | 0.0009 | 884.5 | 891.6 | G | 0.85 | 0.72 | 0.98 | 1.000 | 0.020 |
| | | G/C | 24 | 342.5 (9.46) | 49.91 (21.50 - 78.32) | | | | C | 0.15 | 0.28 | 0.02 | | |
| *fgf4-3* | --- | C/C | 56 | 292.59 (8.56) | 0 | 0.0009 | 884.5 | 891.6 | C | 0.88 | 0.76 | 0.99 | 1.000 | 0.082 |
| | | G/C | 24 | 342.5 (9.46) | 49.91 (21.50 - 78.32) | | | | G | 0.12 | 0.24 | 0.01 | | |
| *myf5-1* | --- | A/A | 36 | 281.53 (11.41) | 0 | 0.0006 | 883.8 | 890.9 | A | 0.72 | 0.68 | 0.78 | 0.160 | 0.003 |
| | | A/G | 44 | 328.86 (7.58) | 47.34 (21.29 - 73.38) | | | | G | 0.28 | 0.32 | 0.22 | | |
| *myf5-2* | --- | T/T | 46 | 288.91 (9.85) | 0 | 0.001 | 885.7 | 892.9 | T | 0.79 | 0.75 | 0.82 | 0.580 | 0.081 |
| | | T/G | 34 | 332.79 (8.37) | 43.88 (17.35 - 70.41) | | | | G | 0.21 | 0.25 | 0.18 | | |
| *myf5-3* | Over dominant | T/T-G/G | 46 | 288.04 (9.3) | 0 | 0.001 | 884.7 | 891.8 | T | 0.78 | 0.72 | 0.82 | 1.000 | 0.020 |
| | | T/G | 34 | 333.97 (9.28) | 45.93 (19.56 - 72.29) | | | | G | 0.22 | 0.28 | 0.18 | | |
| *myod2-1* | | C/C | 45 | 287.89 (8.97) | 0 | 0.001 | 885.1 | 892.2 | C | 0.61 | 0.56 | 0.65 | 0.037 | 0.520 |
| | | C/G | 35 | 332.86 (9.95) | 44.97 (18.63 - 71.31) | | | | G | 0.39 | 0.44 | 0.35 | | |
| *myod2-2* | Over dominant | A/A-G/G | 54 | 292.5 (8.47) | 0 | 0.001 | 885.7 | 892.8 | G | 0.82 | 0.91 | 0.74 | 0.039 | 0.017 |
| | | A/G | 26 | 338.85 (10.66) | 46.35 (18.34 - 74.35) | | | | T | 0.18 | 0.09 | 0.26 | | |
| *myod1-3* | Over dominant | G/G-T/T | 56 | 324.2 (8.07) | 0 | 0.0002 | 881.6 | 888.7 | T | 0.78 | 0.8 | 0.75 | 0.400 | 1.000 |
| | | G/T | 24 | 268.75 (10.86) | -55.45 (-83.34 - -27.55) | | | | C | 0.22 | 0.2 | 0.25 | | |
| *bmp2-1* | Dominant | T/T | 42 | 330.12 (8.92) | 0 | 0.0006 | 883.6 | 890.7 | T | 0.57 | 0.78 | 0.36 | <0.0001 | <0.0001 |
| | | T/G-G/G | 38 | 282.63 (9.8) | -47.49 (-73.41 - -21.57) | | | | G | 0.43 | 0.22 | 0.64 | | |
| *bmp2-5* | Over dominant | G/G-C/C | 48 | 462.94 (36.2) | 0.00 | 0.0043 | 1092.7 | 1099.9 | G | 0.65 | 0.74 | 0.56 | 0.35 | 0.011 |
| | | C/G | 32 | 316.44 (27.68) | -146.50 (-244.14 - -48.86) | | | | C | 0.35 | 0.26 | 0.44 | | |
| *bmp2-10* | Recessive | A/A-A/G | 62 | 294.84 (8.05) | 0 | 0.0006 | 883.7 | 890.8 | G | 0.5 | 0.6 | 0.4 | 0.007 | 0.330 |
| | | G/G | 18 | 351.39 (9.47) | 56.55 (25.53 - 87.57) | | | | A | 0.5 | 0.4 | 0.6 | | |
| *bmp2-11* | Dominant | T/T | 22 | 336.82 (11.97) | 0.00 | 0.001 | 889 | 896.1 | T | 0.59 | 0.69 | 0.5 | 0.000 | 0.72 |
| | | C/T-C/C | 58 | 296.47 (8.25) | -40.35 (-70.34 - -10.37) | | | | C | 0.41 | 0.31 | 0.5 | | |
| *bmp2-15* | Dominant | C/C | 23 | 340.65 (11.51) | 0.00 | 0.0025 | 886.4 | 893.5 | A | 0.59 | 0.54 | 0.65 | 0.013 | 0.01 |
| | | C/G-G/G | 57 | 294.21 (8.2) | -46.44 (-75.54 - -17.34) | | | | G | 0.41 | 0.46 | 0.35 | | |

**Table S3.4. Results from the association tests performed in PowerMarker. FBC cohort: Distance-based test, F-tests for weight, length, and conditioning factor, and an exact G-test.**

| Marker | F-test for Weight p-value | F-test for Length p-value | Distance (Prevosti) p-value | F-test for K p-value | Exact G-test p-value |
|---|---|---|---|---|---|
| *tnksa-1* | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 |
| *tnksa-2* | 0.000 | 0.000 | 0.000 | 0.008 | 0.000 |
| *tnksa-3* | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 |
| *myod-1* | 0.000 | 0.000 | 0.000 | 0.009 | 0.000 |
| *myod-3* | 0.002 | 0.000 | 0.000 | 0.001 | 0.000 |
| *bmp2-1* | 0.000 | 0.002 | 0.000 | 0.002 | 0.000 |
| *bmp2-5* | 0.006 | 0.003 | 0.004 | 0.001 | 0.001 |
| *bmp2-6* | 0.007 | 0.003 | 0.001 | 0.002 | 0.002 |
| *bmp2-11* | 0.004 | 0.006 | 0.001 | 0.004 | 0.000 |
| *bmp2-15* | 0.003 | 0.007 | 0.000 | 0.004 | 0.000 |

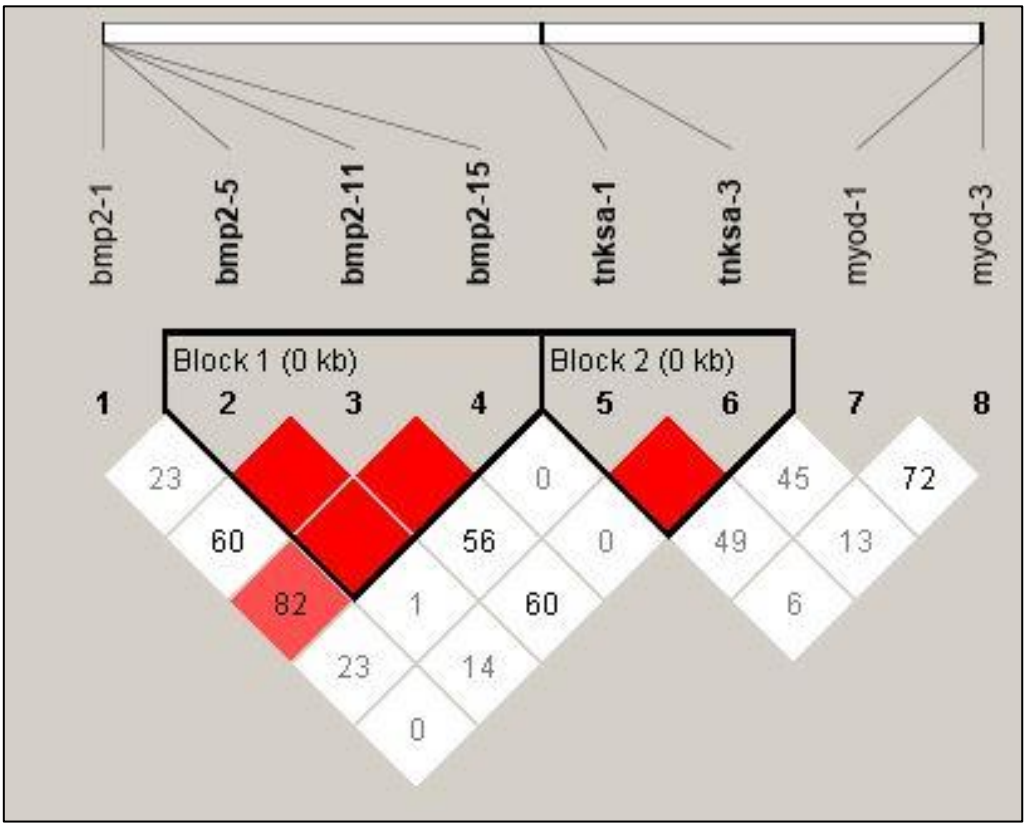
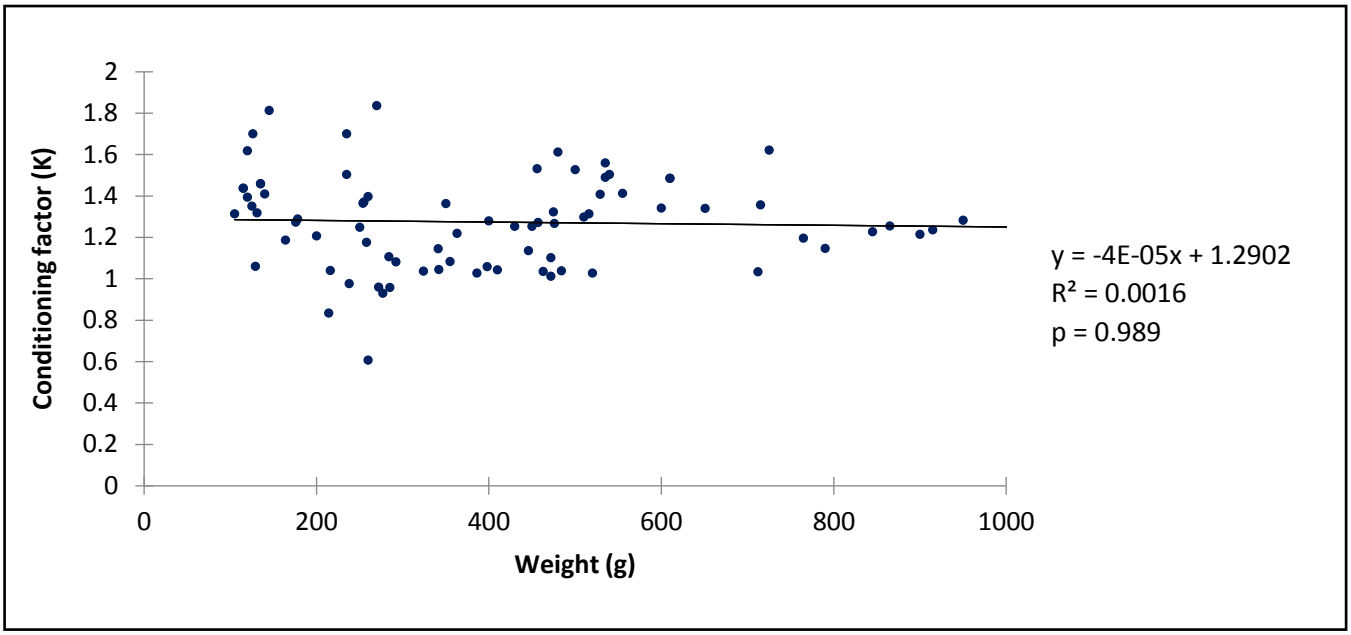



**Figure S3.1. Linkage disequilibrium (LD) block structures. LD block structure consisted of a total of ten SNPs in three different genes. Two SNPs were located in the *MYOD1* gene, two SNPs in the *TNKSA* gene and four SNPs in *BMP2* gene. The LD block was defined by a D' value threshold of 0.8. The colour scale ranges from red to white (colour intensity decreases with decreasing D' value, and all of D' values were = 1).**

A




B

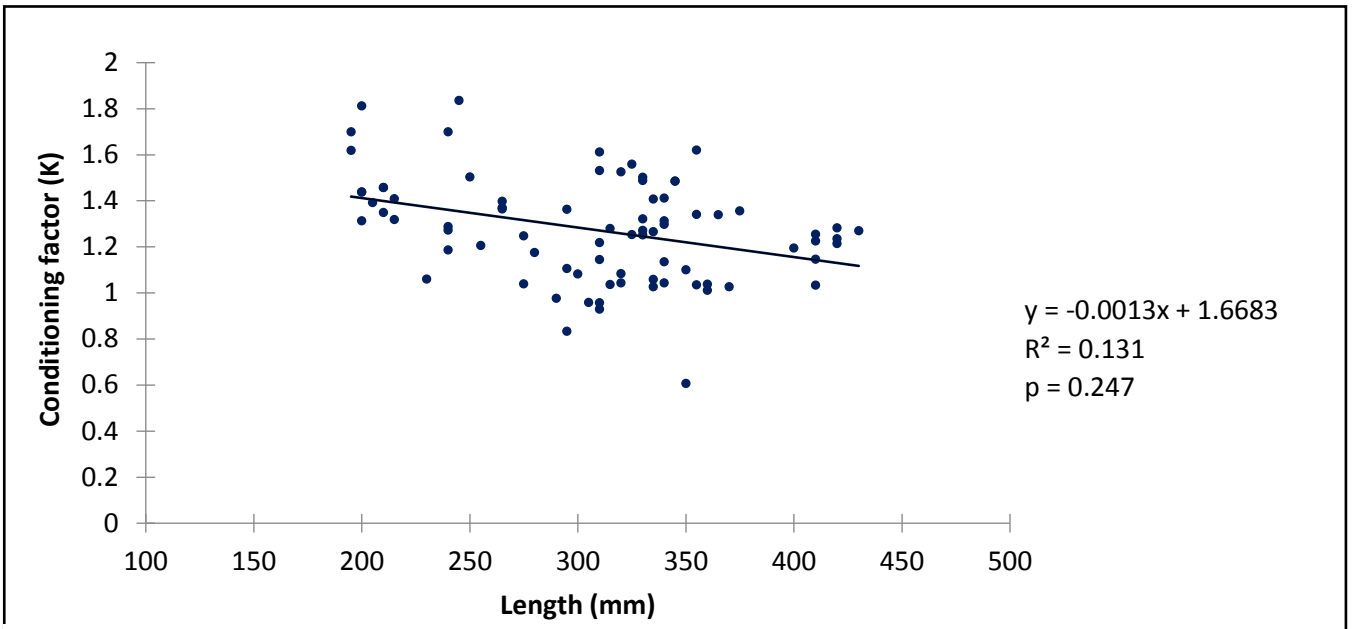



**Figure S3.2. Scatterplots illustrating correlation analysis for Fulton's conditioning factor K versus body weight (A) and length (B). Trend line equations and R2-values are also indicated**

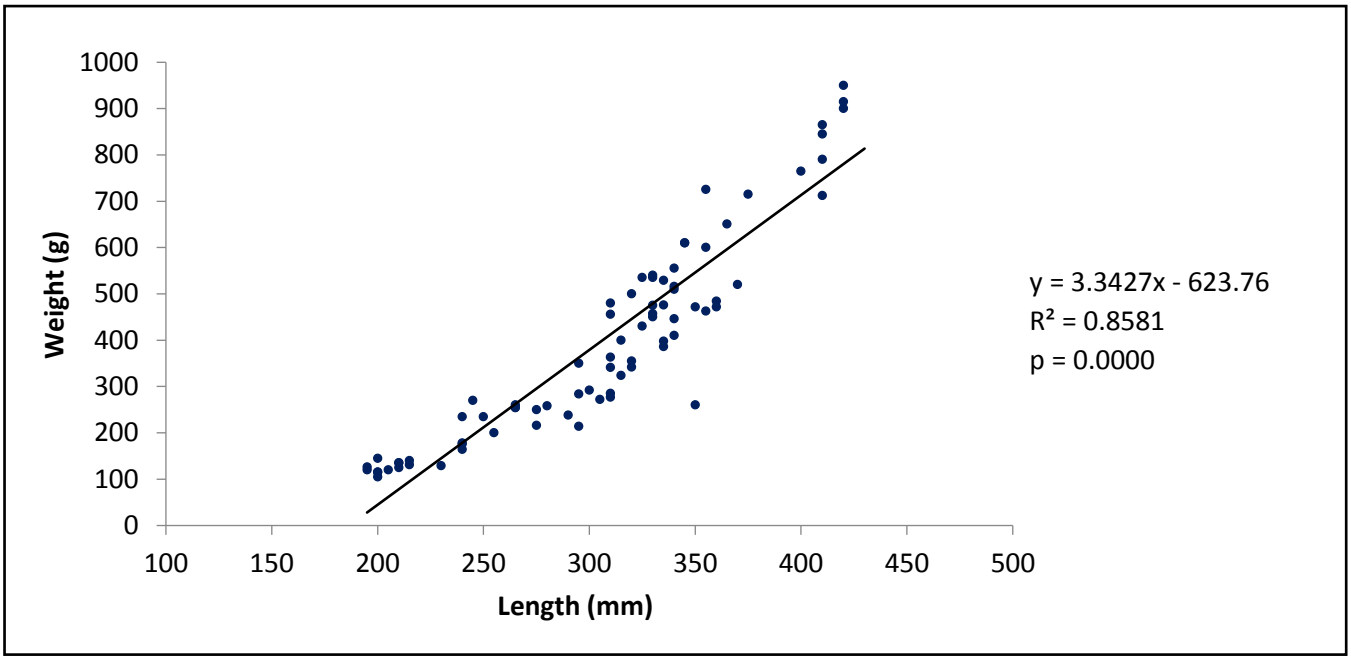



**Figure S3.3. Scatterplots illustrating correlation analysis weight versus length. Trend line equations and R2-values are also indicated.**