SNP screening and validation in *Haliotis midae*

by
Sonja Blaauw

Thesis presented in partial fulfilment of the requirements for the degree Master of Science (MSc) in Genetics at the University of Stellenbosch

Supervisor: Dr. Rouvay Roodt-Wilding
Co-supervisor: Dr. Aletta E. van der Merwe
Faculty of Science
Department of Genetics

March 2012

Declaration

By submitting this thesis/dissertation 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 2012

Abstract

Haliotis midae (commonly referred to as perlemoen) is the only one of five endemic species in South Africa that is commercially valued both locally and internationally. Unfortunately, natural perlemoen populations have become a dwindling resource due to commercial exploitation, poaching and the influx of natural threats, such as the West Coast rock lobster, *Jasus lalandii*. To preserve the natural diversity and sustainability of natural populations as well as commercial stocks, genetic management and improvement of perlemoen is critical. Genetic management requires the utilisation of molecular markers, which aid in the construction of linkage maps and the identification of quantitative trait loci (QTL) associated with economically significant traits. This will allow improvement of commercial stock management in terms of broodstock selection as well as provide valuable insight into natural population dynamics.

Single Nucleotide Polymorphisms (SNPs) were selected as the marker of choice due to their successful employment as molecular markers and their wide distribution and abundance within the genomes of various marine species. This study focuses on the characterisation of novel SNPs from transcript sequences generated by Next Generation Sequencing technology. Approximately 40% of the transcripts facilitated the isolation of 105 putative markers, indicating a SNP frequency of ~1% within the *H. midae* genome.

A subset of 24 markers, in addition to 24 previously developed markers, was characterised using the Illumina GoldenGate genotyping assay with the VeraCode technology, a medium to high-throughput genotyping technology. This is the first reported medium- to high-throughput characterisation of SNPs in *H. midae*. The selected markers were used to determine the efficiency and overall success rate of the GoldenGate platform. Marker characterisation was completed in both natural and commercial populations to determine the utility of these markers for genetic diversity and population structure inference. An 85% genotyping success rate was achieved with the platform. Statistical analysis indicated that the markers developed in this study are suitable for applications including population genetic structure inference, genetic diversity estimation and possibly other downstream applications such as linkage mapping. These markers are considered to be invaluable for future work regarding the genetic management and conservation of *H. midae*.

Opsomming

Haliotis midae (ook bekend as perlemoen) is die enigste van vyf inheemse spesies in Suid-Afrika wat noemenswaardige kommersiële waarde toon plaaslik sowel as internasionaal. Ongelukkig het kommersiële uitbuiting, wildstropery en natuurlike bedreiging (bv. die Weskus kreef *Jasus lalandii),* wilde perlemoen populasies noemenswaardig verminder. Dus, om natuurlike diversiteit en die voortbestaan van beide wilde en kommersiële populasies te beskerm, is genetiese bestuur en verbetering absoluut noodsaaklik. Genetiese bestuur vereis die gebruik van molekulêre merkers as ’n hulpmiddel in die opstellingvan koppelingskaarte, en die identifisering van die relevante kwantitatiewe eienskap loki (QTL) tipies geassosieer met ekonomies belangrike eienskappe. Die laasgenoemde beoog om kommersiële voorraad bestuur te verbeter, kragtens deur broeidier seleksie sowel as om insig te verskaf m.b.t. wilde bevolking dinamika.

Enkel Nukleotied Polimorfismes (SNPs) is gekies as die toepaslike merker vanweë die omvattende toepaslikheid van hierdie merkers binne die genome van verskeie mariene spesies. Hierdie studie fokus op die karakterisering van nuwe SNPs vanuit transkript volgordes ontwikkel deur middel van Volgende Generasie Volgordebepaling (“*Next Generation Sequencing”)*. ’n Beraamde 40% van transkripte het gelei tot die ontwikkeling van 105 potensiëlemerkers, aanduidend van ’n SNP frekwensie van ~1% binne die *H. midae* genoom.

’n Sub-versameling van 24 merkers, tesame met 24 bestaande merkers, is gekarakteriseer deur die Illumina GoldenGate genotiperings toets met die VeraCode tegnologie, ’n medium tot hoë deurvloei genotiperingstegnologie. Hierdie is die eerste berig van medium tot hoë deurvloei karakterisering van SNPs in *H. midae.* Die geselekteerde merkers is gebruik om die doeltreffendheid van die GoldenGate platform te bepaal. Merker karakterisering is uitgevoer in beide wilde en kommersiële bevolkings om die effektiewe bruikbaarheid van hierdie merkers m.b.t. genetiese diversiteit, en bevolking struktuur bepaling, te ondersoek. Die platform het ’n 85% genotiperingsukses syfer getoon. Statistiese analise dui daarop dat merkers ontwikkel tydens hierdie studie toepaslik is vir bevolking genetiese struktuur bepaling, genetiese diversiteitberaming en moontlik ook genetiese koppelingskartering. Hierdie merkers word bestempel as onmisbaar vir toekomstige navorsing in genetiese bestuur en bewaring van *H. midae.*

Acknowledgements

I would like to thank Dr. Rouvay Roodt-Wilding for all her guidance, support and patience throughout my degree as well as her input and dedication towards the preparation of this thesis. To Dr. Aletta van der Merwe, for all her encouragement and inspiration whilst completing my thesis, especially her help with the population genetics, for which I am truly grateful.

I would also like to acknowledge the Central Analytical Facility at Stellenbosch University for their outstanding services and the Innovation Fund for providing me with funding for the duration of my studies. To my colleagues and friends from the Molecular Aquatic Research Group, thank you for all your support and making my time in the department a memorable one.

I would also like to extend my gratitude towards Dr. Zané Lombard at the University of the Witwatersrand for all her assistance and hospitality during the genotyping phase of the project.

Thanks to my family for all their love and support throughout my academic endeavours. You have been a great example of ambition and perseverance. To Deon, my husband and loving friend, thank you for your support and encouragement during my degree, especially this past year when I needed it the most.

Lastly, but by no means least, I would like to thank God for his grace and for being my foundation and life support. This project would not have been possible without Him.

“I can do all things through Christ, who strengthens me”

- Philippians 4:13

List of content

List of Figures

List of Tables

Abbreviations

%	Percentage
<	Less than
>	Greater than
℃	DegreesCelsius
µ	Mutation rate
µl	Microlitre
µM	Micromolar
®	Registered Trademark
3’	Three prime
5’	Five primer
xg	Gravity

A

A	Adenine
AFLP	Amplified Fragment Length Polymorphisms
ASE	Allele-Specific Extension
ASO	Allele-Specific Oligonucleotide
AS-PCR	Allele-Specific PCR

B

BLASTN	Basic local alignment search tool (nucleotide search)
bp	Base pair

C

C	Cytosine
cDNA	Complementary DNA
Chi^2	Chi-square
CITES	Convention on International Trade in Endangered Species of Wild Fauna and Flora
cm	Centimetre
Contig	Contiguous sequences
CTAB	Cetyltrimethylammonium bromide

D

ddNTP	Dideoxynucleotide triphosphate
df	Degrees of freedom
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate

E

EDTA	Ethylenediamine tetra-acetate ($C_{10}H_{16}N_2O_8$)
EST	Expressed Sequence Tag

F

f	Inbreeding coefficient
FCA	Factorial Correspondence Analysis
FP	Fluorescence Polarisation
FRET	Fluorescence Resonance Energy Transfer

G

G	Guanine
GA II	Genome Analyser II
GDA	Genetic Data Analysis
gDNA	Genomic deoxyribonucleic acid
GO	Gene Ontology

H

HAB	Harmful algal blooms
He	Expected Heterozygosity
Ho	Observed Heterozygosity
Hr	Hour
H-W law	Hardy-Weinberg law
HWE	Hardy-Weinberg Equilibrium

K

kb	Kilobase pair
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KOG	Eukaryote Clusters of Genes

L

LD	Linkage disequilibrium
LSO	Locus-Specific Oligonucleotide

M

M	Molar (Moles per Litre)
MAF	Minor Allele Frequency
MALDI-TOF MS	Matrix-Associated Laser Desorption Time of Flight Mass Spectrometry
MAS	Marker Assisted Selection
mb	Megabase pair
$MgCl_2$	Magnesium chloride
Min	Minutes
ml	Millilitre
mM	Millimolar
mRNA	Messenger ribonucleic acid
mtDNA	Mitochondrial deoxyribonucleic acid

N

n	Different number of alleles
NaOH	Sodium hydroxide
NCBI	National Centre of Biotechnology Information
N_e	Effective population size
ng	Nanogram

O

OLA	Oligation Ligation Assay

P

P	Probability value
PCR	Polymerase Chain Reaction
pH	Concentration of hydrogen ions in a solution
Pi	Frequency of alleles
PIC	Polymorphic Information Content
pmol	Picomole

Q

QTL	Quantitative Trait Loci

R

RAPD	Random Amplified Polymorphic DNA
RFLP	Restriction Fragment Length Polymorphisms
RNA	Ribonucleic acid
rpm	Revolutions per minute

S

SADC	Southern African Development Community
Sec	Seconds
SNP	Single Nucleotide Polymorphism
spp	Several species
SSR	Simple Sequence Repeats
SUD	Single Use DNA

T

T	Thymine
Taq	*Thermus aquaticus* DNA polymerase
TBE	Tris-Borate-EDTA buffer
TE	Tris-Ethylenediamine tetra-acetate

Tm	Melting Temperature
TM	Trade mark
Tris-HCl	Tris- (hydroxymethyl) aminomethane hydrochloric acid
Ts	Transition
Tv	Transversion

U

U	Units (enzyme)
UDG	Uracil DNA glycosylase
USD	United States Dollar
UTR	Untranslated Region

V

v	Version
V	Voltage
VBP	VeraCode Bead Plate

W

w/v	Weight per volume

Z

ZAR	South African Rand

Chapter 1

Literature Study

Introduction

Haliotis midae is a relatively large mollusc species that inhabits the coastal waters of South Africa and is the country's most commercially valuable aquatic species (Britz 1991; Britz *et al*. 2009). Due to destruction of their natural habitat, predation as well as overfishing, artificial cultivation in the form of abalone farming has been established over the last 20 years (Britz *et al*. 2009; Bester-van der Merwe *et al*. 2011). Although these farms mainly operate for commercial reasons, they inadvertently relieve stress from natural populations. In order for South African abalone farms to compete in the international market, which has become extremely lucrative and increasingly demanding, various forms of genetic management have been implemented. Issues associated with farmed populations including inbreeding and genetic drift can be addressed with the utilisation of molecular markers in such genetic management programmes (Roodt-Wilding and Slabbert 2006).

Since the start of molecular genetics within aquaculture in the 1970's, DNA-based marker technology has modernised the way genetic research is conducted within this sector. These technological advances have created a series of markers, including Amplified Fragment Length Polymorphisms (AFLPs), Restriction Fragment Length Polymorphisms (RFLPs), Simple Sequence Repeats (SSRs or microsatellites) and more recently Single Nucleotide Polymorphisms (SNPs). These markers are successfully used for the identification of stocks, the management of commercial broodstocks, parentage assignment and linkage mapping studies (Liu and Cordes 2004). With increased availability of genetic markers, commercially important traits such as growth and disease resistance, most often quantitative in nature, can be researched. These quantitative traits can be utilised in conjunction with high-resolution maps for Marker Assisted Selection (MAS) (Roodt-Wilding and Slabbert 2006). The implementation of MAS programmes would allow abalone farms to select individuals for particular advantageous traits such as increased growth rate by focusing on markers associated with genes and therefore genotypes of interest instead of selecting individuals based only on phenotypic traits.

Genetic markers are classified in either one of two categories, referred to as type I or II (Liu and Cordes 2004). Type I markers are associated with genes of known function and can also be used in comparative mapping studies where markers are mapped between

model and non-model species that are devoid of genetic maps, whereas type II markers are usually coupled with anonymous genomic content (O'Brien 1991). Single Nucleotide Polymorphisms are genetic variations that occur due to point mutations at specific loci on a nucleotide level (Liu and Cordes 2004). They can consist of up to four alternative nucleotides (tetra-allelic) (Kim and Misra 2007), but typically only contain two bases, classifying it as a bi-allelic marker (Vignal *et al*. 2002). These base substitutions involve either a transversion (purine-pyrimidine) or a transition (purine-purine and pyrimidine-pyrimidine) of a nucleotide (Kim and Misra 2007). Single Nucleotide Polymorphismsare co-dominantly inherited and can occur in both non-coding (type II) and coding regions (type I); the latter type of marker is known as gene-linked markers, which can be used for the identification of candidate genes, expression profiling and gene function, to name a few. Although type II SNPs are associated with anonymous/non-coding areas, they can affect for example transcription factor binding and gene splicing (Churbanov *et al*. 2010). The use of point mutations in research studies have become more popular than the use of microsatellites due to their lower mutation rate of 10^{-8} - 10^{-11} per locus per generation (Kondrashov 2002; Lupski 2007; Wielgoss *et al*. 2011) and higherfrequency (every 200-500 base pairs) within the genome (Brumfield *et al*. 2003; Morin *et al*. 2004; Vera *et al*. 2011). To date 264 microsatellite (Bester *et al*. 2004; Slabbert *et al*. 2008; Hepple 2010; Rhode 2010; Slabbert *et al*. 2010; Jansen 2011) and 40 SNP markers (Bester *et al*. 2008; Rhode *et al*. 2008; Rhode 2010) have been reported for *H. midae.*

Single Nucleotide Polymorphism discovery entails the isolation of polymorphisms either from cDNA libraries or whole genomic data. The use of cDNA libraries for SNP isolation is adequate for organisms that lack a complete reference genome (non-model species) (Usesche *et al*. 2001). With increased efforts in next generation sequencing (NGS), SNP discovery in non-model species have become easier and more accessible (van Bers *et al*. 2010). Single Nucleotide Polymorphism markers can be used in various applications, including genetic diversity (Ciobanu *et al*. 2009) and relatedness studies (Weir *et al*. 2006), stock identification (Wirgin *et al*. 2007), parentage assignment (Anderson and Garzam 2006) and population determination (Narum *et al*. 2008).

1 *Haliotis midae* (perlemoen)

1.1 Species Information

1.1.1 Nomenclature and Distribution

Abalone are univalve organisms, classified as marine gastropods belonging to the family Haliotidae (Table 1.1). This family includes approximately 56 species globally while four areas of endemism (temperate Australia, South Africa, New Zealand and North Pacific) have been recognised (Geiger 2000).

Table 1.1.*Haliotis* nomenclature and classification. (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?lvl=0&id=36098 [accessed December 2010]).

Taxonomy / Classification	
Phylum	Mollusca
Class	Gastropoda
Family	Haliotidae
Genus	*Haliotis*

Five of these species are found along the coast of southern Africa. The largest and most predominant species found in South Africa is *Haliotis midae* (commonly known as perlemoen(Figure 1.1)) (Muller 1986; Roodt-Wilding and Slabbert 2006). It resides in sublittoral zones along rocky surfaces from Port St Johns on the East coast to St Helena Bay on the West coast (Hecht 1994; Marine and Coastal Management 2010). The smaller, less abundant species, *H. queketti* and *H. speciosa* (= *H. alfredensis*) are typically found along the East coast of South Africa. *Haliotis midae* along with *H. spadicea* (siffie) has the widest distribution along the South African coast, while *H. parva* is mainly restricted to the South and West coast.

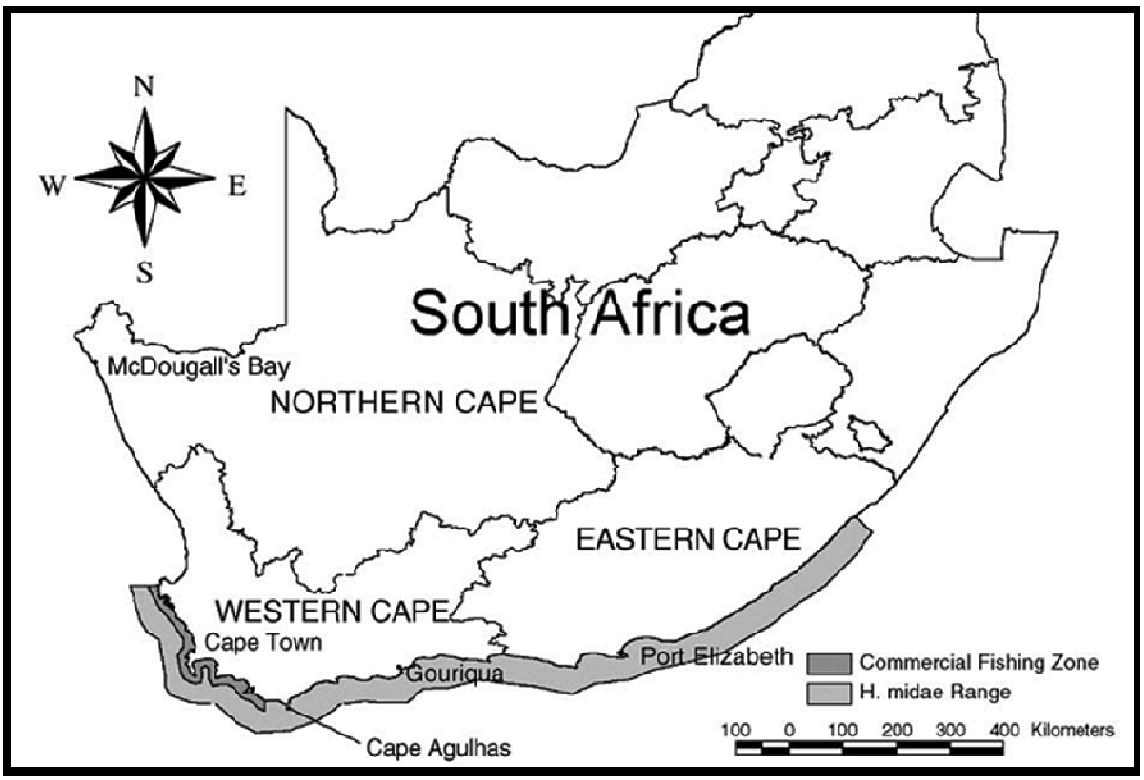

Figure 1.1. The distribution of *H. midae* along the coast of South Africa, depicting the commercial fishing zones and the natural habitats. The West and East coast of South Africa is separated by a hydrothermal break at Cape Agulhas seperating the cold Benquela current from the warm Agulhas current (De Waal *et al.* 2003).

1.1.2 Morphology and Anatomy

Perlemoen are characterised by a flat ear-shaped shell, as opposed to the usual spiral contour exhibited by many other *Haliotis* species. The shell is bordered by a ridge of small perforations, known as tremata, functioning as aerating pores to the gills (Figure 1.2) (Fallu 1991; Hecht 1994; Marine and Coastal Management 2010). The shell can reach up to 230 mm in size. The largest part of the body that is covered by the shell is the adductor muscle, commonly referred to as the foot (Hahn 1989). The foot allows the animal to adhere itself to rocky surfaces in its natural habitat and serves as a means of feeding. This muscular organ is surrounded by the epipodes, giving the foot its particular shape and the origin for the formation of the tentacles. The anterior part of the body consists of the head, which comprises of sensory organs, appendages and a mouth (Fallu 1991).

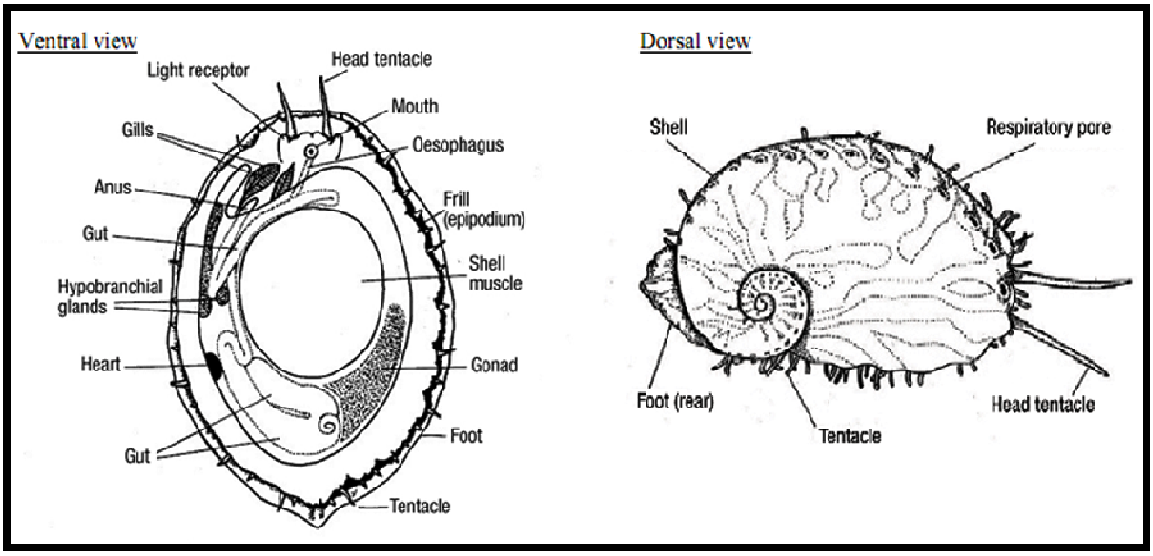

Figure 1.2. Morphological and anatomical features of abalone (Fallu and Lang 1994).

1.1.3 Feeding Habits

Abalone are nocturnal herbivores, with feeding habits dependent on their developmental stage, season and size (Bolton 2006). As a juvenile, abalone feed on benthic micro-organisms and diatoms, whereas adult abalone mainly feed on kelp (Kawamura and Takami 1995). Adult perlemoen feed on *Ecklonia maxima*, a kelp species indigenous to the western coast of South Africa (Hecht 1994). Mature abalone are sedentary feeders, remaining stationary while feeding. At most, they will attach themselves to drifting kelp, trapping it with the adductor muscle for feeding (Troell *et al*. 2006).

1.1.4 Reproduction and Life Cycle

Abalone aredioecious organisms that live in large clusters and reproduce by means of spawning. Sexual maturity in *H. midae* is reached at about 7 ½ years in the wild and is visually indicated by the colour and structure of the reproductive organs (gonads) (Sales and Britz 2001). The female gonad is a dark green or brown structure, opposed to the yellow structure in males, and is situated at the right-hand side of the body. Spawning occurs with the simultaneous release of eggs and sperm into open water; the eggs are greenish in colour while the sperm appears as whitish (Fallu 1991). Spawning only occurs twice a year during March and November (Newman 1967) when the water temperature is

optimal, which is approximately 20°C (Genade *et al*. 1988; McShane *et al*. 1988). Upon fertilisation, the eggs migrate toward the benthic zone, where it hatches between 12-24 hours after fertilisation. The newly hatched larvae, known as trochophores, are lecithotrophic organisms moving towards the photic zone of the ocean, thereby presenting phototactic behaviour (Tarr 1987). The trochophores then enter the veliger larval stage (Tarr 1987), where essential organs such as the swimming cilia develop. After approximately 35-40 days, the organisms have passed the settling phase and have developed into juvenile abalone (McShane 1989).

1.2 Constraints on perlemoen

1.2.1 Illegal Harvesting

Illegal harvesting of perlemoen started in the early 1990's and rapidly evolved into a multi-million dollar crime syndicate by the end of the decade (Hauck and Sweijd 1999). This was attributed to numerous factors such as the dwindling of the South Africa Rand (ZAR) against the American Dollar (USD), the transition of the democracy and the elimination of apartheid laws in the early 1990's as well as the establishment of Chinese organised crime networks (Steinberg 2005). This further escalated with the closure of legal and commercial fisheries in 2003 and 2008, resulting in an increase of fishermen from small fishing villages taking part in illicit harvesting (DEAT 2003, 2007). By 2007, poaching had spiralled out of control with reports of more than 2000 tons of whole mass harvested. This was due to the increased use of sea-based vessels rather than shore-based diving. Studies indicated that the amount of abalone confiscated between 2004 and 2006 increased as individual arrests declined; this is attributed to arrests occurring from transport vehicles and catch points. Although efforts were made to combat poaching, such as the listing of abalone on Appendix III of CITES (Convention on International Trade in Endangered Species of Wild Fauna and Flora) (Raemaekers *et al*. 2011), these efforts were short term and unorganised against the large crime syndicates (Raemaekers and Britz 2009). The lack of endorsing CITES permits by the South African government since 2007 have hindered the effectiveness of listing with CITES and has led to massive amounts of abalone imported into Hong Kong from neighbouring South African Development Community (SADC) countries, such as Mozambique and Zimbabwe (Groenewald 2009). All the above-mentioned has led to a continued increase in abalone poaching and with the lack of fishing rights, ineffective border and sea-based patrolling, socio-economic factors as well as the

high demand of abalone, illegal trade will persist until the resource is completely depleted (Raemaekers and Britz 2009).

1.2.2 Ecological Constraints

1.2.2.1 Sea urchin and South African West Coast rock lobster

The Cape sea urchin, *Parechinus angulosus*, belongs to the phylum Echinodermata and is indigenous to the rocky shores of the South African Western Cape. The physical nature of this organism allows it to provide shelter for juvenile abalone where little natural protection is available (Day 1998). A study conducted by Tegner and Levin (1982), indicated that depending on the availability of food supplies for sea urchins, theycan also provide juvenile abalone with access to additional food sources. Studies have also indicated that a positive relationship exists between the two species, with factors influencing the survival of both species to the same extent (Tarr *et al*. 1996). The decline in sea urchins was first noticed in the 1990's with the influx of West Coast rock lobster, *Jasus lalandii* between Hermanus and Hangklip (Cockcroft *et al*. 2008). Lobsters feed on a variety of benthic organisms, including sea urchins, turbinid snails and abalone. The increased predation of *J. lalandii* on *P. angulosus* had deteriorating effects on these populations in the vicinity, thusjeopardising the survival of juvenile abalone. This is supported by the disappearance of juvenile abalone from areas where sea urchin populations have been diminished (Tarr *et al*. 1996; Hauck and Sweijd 1999).

1.2.2.2 Algal Blooms

Another, less severe natural limitation on abalone populations is harmful algal blooms (HAB's). In 1999, the presence of paralytic shellfish poisoning (PSP) toxins was observed in populations of cultured abalone, and subsequently in wild populations as well. Paralytic shellfish poisoning toxins accumulate in filter feeders and grazers, such as oysters and abalone, when toxic dinoflagellates are consumed. A common dinoflagellate in the coastal waters of South Africa is *Alexandrium catenella* (Pitcher *et al*. 2001). Human consumption of animals containing elevated concentrations of these toxins could cause fatalities, although the effects of these toxins in abalone only inhibit growth rate (Botes *et al*. 2003).

1.3 Abalone Aquaculture

1.3.1 Historical Perspective

Only 14 species of abalone are commercially cultivated worldwide; China and the Asian-Pacific region being the largest producers and thus the main contributors to the aquaculture market. Worldwide aquaculture production contributed 51.7 million tons in 2006, of which 27% was accounted by mollusc production (FAO 2009; FISHTECH 2010). In South Africa, abalone production accounted for 81% of the total Aquaculture sector in 2008 (Britz *et al*. 2009). This high demand for perlemoen persists as its flesh resembles those of the Japanese species, making it a sought after delicacy in East Asian countries (Britz 1996). *Haliotis midae* is the only species of commercial value in South Africa due to its widespread occurrence, easily accessible habitat and size (Roodt-Wilding and Slabbert 2006). In 1949, unrestricted commercial harvesting was implemented, but due to over-exploitation of the resource, seasonal quotas and recreational fishing permits were implemented from the early 1970's (Tarr 2003; Steinberg 2005; Raemaekers *et al*. 2011). This over-exploitation of perlemoen resulted in major sustainability issues to the natural resource, which was further depleted by the rapid escalation of poaching in South Africa. In 2007, *H. midae* was registered to Appendix III of CITES. This meant that all consignments of the species for international trade required the CITES permits (Raemaekers *et al*. 2011). It enforced the co-operation of neighbouring and other countries in monitoring international abalone trade and regulating controls of large importing countries such as China, Japan and Hong Kong. Unfortunately, in May 2010 the South African government decided to remove perlemoen from the list (Bürgener 2010). This could further jeopardize wild populations, as abalone is still in high demand in the international market.

1.3.2 Abalone Farming

Abalone has been harvested commercially and illegally to maintain the increasingly high demand of the international market. Therefore, in the early 1990's cultivation of the species commenced not only to alleviate strain on natural populations, but more importantly for commercial purposes (Sales and Britz 2001). Currently, 18 abalone farms have been established along the coast of South Africa; 13 of which are on the West Coast. These farms operate by inducing spawning in adult abalone, and employing land-based

hatcheries to rear the juveniles until they reach export size (shell length of 114 mm)(Tarr 2003; Britz *et al.* 2009; Raemaekers *et al*. 2009).

1.3.3 Socio-economic Importance

Commercial abalone farms not only play an integral role in alleviating the strain on wild populations, but also contribute substantially to South Africa's socio-economic structure. Unemployment is the main contributor to the governments' socio-political problems, especially in poor coastal communities where resources are limited (Kingdom and Knight 2003; Troell *et al*. 2006). Abalone farms provide direct relief to these communities, by providing job opportunities for unskilled as well as semi-skilled workers. Due to the labour intensive nature of work on these farms, the majority of the developing farms typically employ mostly male workers. Conversely, more established farms also employ female workers for various tasks including grading and hatchery duties. Thus, it seems that gender inequalities are comparatively addressed. This industry also provides sustainability for second economy industries including abalone processing and seaweed industries (Troell *et al*. 2006).

1.3.4 Genetic Management

Commercial cultivation of *Haliotis midae* has proved successful since its establishment 20 years ago. With international demand for abalone, pressure is placed on farms to cultivate animals in a short period of time to meet market standards. Abalone are reared for three to four years, after which they have reached a suitable export size (Roodt-Wilding and Slabbert 2006; Troell *et al*. 2006). Due to their lengthy life cycle, various factors need consideration to enhance stocks and prevent disease. Therefore, various South African farms have employed genetic improvement programs to target key issues on a fundamental level such as growth rates, disease susceptibility, desirable market traits and fecundity levels with regard to age, amongst others (Roodt-Wilding and Slabbert 2006).

Genetic management includes parentage assignment and assessing levels of genetic diversity of cultured stocks. Knowledge of relatedness is essential to prohibit the occurrence of mating between related individuals (inbreeding) within populations and maintaining large effective population sizes are necessary to reduce genetic drift (the loss

of alleles within the population); both of which negatively affects genetic diversity (Allendorf and Phelps 1980; Vrijenhoek 1994; Hill 2000). This loss in variation adversely affects traits such as growth rate and disease resistance, and could ultimately lead to a genetic bottleneck. Genetic drift and bottlenecks have been observed previously in perlemoen (*H. midae*) (Slabbert *et al*. 2009), Blacklip abalone (*H. rubra*) (Evans *et al*. 2004a) and Ezo abalone (*H. discus hannai*) (Li *et al*. 2004). One way to prevent the latter is with the use of parentage assignment in which the relatedness of broodstock individuals can be determined and individuals that have ceased to spawn can be identified.

With parentage analysis, levels of genetic diversity can be maintained by mating genetically diverse animals and labour intensive labelling of offspring would become unnecessary. This could also lead to other applications including selective breeding where individuals are selected based on their genetic traits (Jerry *et al*. 2004; Roodt-Wilding and Slabbert 2006). Selective breeding requires the use of pedigree information to facilitate the development of mating designs to enhance genetic gain (Vandeputte *et al*. 2011). Successful parentage analysis have been demonstrated in Ezo abalone (*H. discus hannai*) (Selvamani *et al*. 2001; Hara and Sekino 2007), Pacific abalone (*H. discus discus*)(Li *et al*. 2003), Donkey's ear abalone (*H. asinina*) (Lucas *et al*. 2006) and perlemoen (*H. midae*)(Ruivo 2007; Slabbert *et al*. 2009; Van Den Berg and Roodt-Wilding 2010). The estimation of genetic diversity and parentage assignment within cultured populations is achieved with the employment of molecular markers.

2 Aquaculture and Genetics

2.1 Molecular Markers in Aquaculture

Molecular markers are categorised as either type I or type II markers, depending on their association with genomic content. Type I markers are associated with genes of known function or coding regions, as opposed to type II makers, which are associated with anonymous genomic sequences (O'Brien 1991; Liu and Cordes 2004). Type II markers include Random Amplified Polymorphic DNA (RAPDs), SNPs, RFLPs, AFLPs and SSRs, also referred to as microsatellites. Microsatellites are thought to be selectively neutral, suggesting that they are not influenced by selection due to their prevalence in non-coding regions (Brown and Epifanio 2003). Although type II markers are principally useful in

linkage mapping and population studies, they have limitations in that for example their discrimination power is diminished when analysing recently diverged taxa (Okumuş and Çiftci 2003; Liu and Cordes 2004). These markers also have limited potential in comparative mapping and candidate gene mapping due to their anonymous nature. Therefore, type I markers are preferred markers in various applications, particularly population studies, as well as linkage and quantitative trait loci (QTL) mapping (Liu and Cordes 2004). Common type I markers include allozymes and Expressed Sequence Tag (EST) based markers, as well as RFLPs, microsatellites and SNPs. These can be employed in studies regarding candidate gene identification, genome evolution as well as comparative studies to transfer genomic information from one species to another (Liu and Cordes 2004). Characteristics of the various markers used in aquaculture are summarised in Table 1.2.

Table 1.2.Molecular markers and the application in aquaculture research.

Makers	Type	Inheritance	Nr. of alleles	Nr. of loci	PIC[7] value	Discrimination power	Applications
Allozymes	I	Co-dominant	2-6	Single	Moderate	Low	Population studies Linkage mapping
mtDNA[1]	-	Maternal	Multiple haplotypes	Single	Moderate*	Intermediate	Population studies
RAPD[2]	II	Dominant	2	Multiple	Moderate	Intermediate	Species identification Linkage mapping Fingerprinting
RFLP[3]	I or II	Co-dominant	2	Single	Low	Low	Linkage mapping
AFLP[4]	II	Dominant	2	Multiple	Low	High	Population studies Linkage mapping
SSR[5]	I or II	Co-dominant	Multiple	Single	High	High	Population studies Linkage mapping Comparative studies
SNP[6]	I or II	Co-dominant	2-4	Single	Moderate**	High	Linkage mapping

1- Mitochondrial DNA, 2- Random Amplified Polymorphic DNA, 3- Restriction Fragment Length Polymorphism, 4- Amplified Fragment Length Polymorphism, 5- Simple Sequence Repeat, 6- Single Nucleotide Polymorphism, 7- Polymorphic Information Content.

* compared to other markers except allozymes; ** when haplotypes can be derived, otherwise low (Dodgson *et al*. 1997; Liu and Cordes 2004)

Markers are often described according to their marker efficiency, which is determined by their polymorphic information content (PIC), an indication of the level of polymorphism detected by the markers. The PIC value is calculated using the distribution frequency of the alleles detected (Botstein *et al*. 1980);$1 - \sum_{n=1}^{n} Pi^2$, where *Pi* is the frequency of alleles for a specific marker locus and *n* is the different number of alleles (Varshney *et al*. 2007). Informativeness of a marker is thus related to the number of alleles per locus. Consequently, bi-allelic markers have lower PIC scores than markers that have multiple alleles. Markers are also classified according to their mode of inheritance, of which most are in a Mendelian fashion with the exception of mitochondrial DNA (mtDNA) markers. The majority of genetic markers, e.g. microsatellites, SNPs, RFLPs and allozymes, are inherited co-dominantly, whereas AFLPs and RAPDs are considered dominant markers. Co-dominant markers allow the detection of both heterozygous and homozygous states whereas dominant markers only allow the identification of the homozygous genotype (Liu and Cordes 2004).

Several markers are popular in aquaculture genetics, depending on the required type of data to be generated and analysed. Markers are generally employed for parentage assignment, genetic diversity studies of natural and commercial populations and genetic profiling of natural and commercial populations (Subasinghe *et al*. 2003; Cenadelli *et al*. 2007). Other applications include linkage studies, QTL mapping and MAS (Liu and Cordes 2004). Quantitative traits are controlled by numerous genes and include traits such as growth rate, disease resistance and food conversion rate; all of which are economically important (Falconer and Mackay 1996). For identification of these QTLs, linkage maps are needed, which require highly informative markers that have good coverage of the genome and are easily genotyped (Martinez 2006). Unfortunately, to date only a few linkage maps are available for mollusc species, including Eastern Oyster (*Crassostrea virginica*) (Yu and Guo 2006), Zhikong scallop (*Chlamys farreri*) (Zhan *et al*. 2009), Blacklip abalone (*Haliotis rubra*) (Baranski *et al*. 2006), Pacific Abalone (*H. discus hannai*) (Sekino and Hara 2007; Qi *et al*. 2007), European flat oyster (*Ostrea edulis*) (Lallias *et al*. 2007) and the Japanese abalone (*H. diversicolor*) (Zhan *et al*. 2011). This illustrates the lack of marker resources in many species, which hampers advanced genetic technologies. It is therefore imperative for research studies to investigate marker development for, amongst others, the construction of high-resolution maps for successful implementation of genetic improvement programmes.

Molecular marker technology in aquaculture was initiated with the use of allozymes, after which various marker types were explored for large-scale marker development studies, including AFLPs and RFLPs. The markers proved to be successful in a number of fields in aquaculture, including stock identification (Lehoczky *et al*. 2005; Trape *et al*. 2009), inbreeding studies (McGoldrick and Hedgecock 1997), genetic mapping (QTL and linkage) (Li and Guo 2004; Perez 2004; Boulding *et al*. 2008),population studies (Tassanakajon *et al*. 1998) and diversity studies (Benzie *et al*. 2002). Although the above-mentioned markers have proven useful in aquaculture research, restrictions such as DNA sample quality (allozymes, mtDNA markers, RAPDs) as well as expensive and time consuming development (mtDNA markers and RFLPs) have led to the use of other marker types, including microsatellites and SNPs (Vignal *et al*. 2002; Okumuş and Çiftci 2003; Liu and Cordes 2004).

2.2 Microsatellite Markers

Microsatellites consist of short tandem repeats of up to six base pairs that are present at approximately every 10 kilobase pairs (kb) in fish species (Litt and Luty 1989, Tautz 1989; Wright 1993). These simple sequence repeats are found in coding and non-coding regions as well as in introns; distributing them evenly throughout the genome. The use of microsatellites have exceeded that of previous types of markers as their detection and isolation is fairly simple (Liu and Cordes 2004) and only require small amounts of DNA. Applications include parent-offspring identification in diverse populations as well as population studies, linkage mapping and QTL mapping in aquaculture species including perlemoen (*H. midae*) (Slabbert 2010), Red abalone (*H. rufescens*) (De La Cruz *et al*. 2010) and Japanese abalone (*H. diversicolor*) (Zhan *et al*. 2011), to name a few. A useful attribute is that microsatellite primers can potentially cross-amplify in related species; however, development is still laborious and expensive (Morris *et al*. 1996; Estoup and Angers 1998). Until recently, microsatellite markers have been the preferred marker in aquaculture genetics due to their even distribution, abundance, co-dominant nature and high mutation rate (10^{-2} - 10^{-6} per locus per generation); the latter resulting from slippage by polymerase enzymes during replication and unequal crossing-over (Levinson and Gutman 1987; Tautz 1989; Ellegren 2000; Oliveira *et al*. 2006).

2.3 Single Nucleotide Polymorphisms (SNPs)

Single Nucleotide Polymorphisms are variations that occur due to point mutations on a nucleotide level at a given locus, resulting in different alleles (Liu and Cordes 2004). These markers can have up to four different alleles, but commonly involves only two, making it a bi-allelic marker. Compared to microsatellite markers, SNPs have very low information content per locus; this is however compensated by their great abundance within the genome (Liu and Cordes 2004). The main reason for the low PIC is the small probability that more than one independent substitution will occur at a single locus. These base substitutions involve either a transversion (purine-pyrimidine) or a transition (purine-purine and pyrimidine-pyrimidine) of a nucleotide (Vignal *et al*. 2002; Liu and Cordes 2004; Kim and Misra 2007). Transversions should in theory, occur more frequently than transitions, but literature has indicated that there is a bias towards transitions (Vignal *et al*. 2002; Keller *et al*. 2007). The occurrence of SNPs within coding regions can either directly affect protein function, referred to as non-synonymous, or have no effect on the protein product, in which case it would be a synonymous mutation/substitution (Liao and Lee 2010). Non-synonymous substitutions provide direct information on gene function, which is important for understanding mutations linked to specific traits (Hayes *et al*. 2007a).

The frequency and distribution of SNPs within a genome are superior to any other molecular marker; making it the marker of choice in molecular research. These features makes SNPs ideal for the generation of higher density genetic maps in order to study economically important features associated with complex traits (Syvanen 2001; Brumfield *et al*. 2003; Hayes *et al*. 2007a). The co-dominant inheritance of SNPs furthermore allows these markers to be used in comparative genome analysis by constructing comparative linkage maps between model and non-model species (Liu and Cordes 2004). These markers are also suitable for automated analysis and high-throughput genotyping with relative ease and low cost (Moreno-Vazquez *et al*. 2003; Shen *et al*. 2005; Barbazuk *et al*. 2007). The development of SNPs in aquaculture species is increasing, as evident in reports on Atlantic cod (*Gadhus morhua*)(Moen *et al*. 2008; Hubert *et al*. 2009), Japanese scallop (*Patinopecten yessoensis*) (Liu, W. *et al*. 2011), Atlantic salmon (*Salmo salar*) (Hayes *et al*. 2007a; Andreassen *et al*. 2010) and channel catfish (*Ictalurus punctatus*) (Liu, S. *et al*. 2011). Compared to microsatellites, SNPs present lower mutation rates (10^{-8} – 10^{-11}). This, along with their abundance, makes the use of SNPs extremely

advantageous in population and pedigree studies (Landegren *et al*. 1998; Bhattramakki and Rafalski 2001; Helyar *et al*. 2011).

Unlike most markers, SNPs are categorised as both type I and II markers. Normally they are considered a type II marker, except for when they are developed from expressed sequences such as Expressed Sequence Tags (ESTs) (Liu and Cordes 2004). In the case of *H. midae*, as in many other non-model organisms, no complete genome map is available at present. This leaves ESTs as an alternative and successful source for SNP discovery (Buetow *et al*. 1999; Gurvey *et al*. 2004; Hayes *et al*. 2007b; Bester *et al*. 2008). Expressed Sequence Tags are partial sequences that arise from the random sequencing of complementary DNA (cDNA) clones (Adams *et al*. 1991). These sequences represent the transcriptional products of genes and therefore, any alterations found within these ESTs are associated with the represented gene (Sarropoulo *et al*. 2008). Despite limitations such as intron positions and gene order within the genome, using transcribed sequences such as ESTs for SNP discovery is advantageous. Such SNPs would be directly associated with genes and could therefore be used for gene-associated mapping and the identification of causative genes.

2.4 Molecular Markers in *Haliotis midae*

The employment of molecular markers in *Haliotismidae* is diverse, including markers such as AFLPs, microsatellites and more recently SNPs (Bester *et al*. 2004; Badenhorst and Roodt-Wilding 2007; Bester *et al*. 2008; Rhode *et al*. 2008; Slabbert *et al*. 2008, 2009; Rhode 2010; Slabbert 2010; Slabbert *et al*. 2010; Bester-van der Merwe *et al*. 2011; Jansen 2011). The application of a specific molecular marker largely depends on the efficiency of the marker to answer specific research questions (Klinbunga *et al*. 2003; Evans *et al*. 2004a; Badenhorst and Roodt-Wilding 2007; Qi and Kijima 2007). The majority of research conducted thus far on abalone focused on development of markers, construction of linkage maps with the purpose of identifying QTLs, parentage assignment, genetic identification of stocks and differentiation between natural and commercial populations. Due to their polymorphic attributes, microsatellites currently dominate as markers in abalone research, with 264 microsatellite (Bester *et al*. 2004; Slabbert *et al*. 2008; Hepple 2010; Rhode 2010; Slabbert 2010; Slabbert *et al*. 2010; Jansen 2011). Although microsatellites are very useful markers, limitations including development time,

size homoplasy and the detection of null alleles can be somewhat taxing. Thus, research focus is shifting towards the isolation and characterisation of SNPs with 40 SNP markers isolated for *H. midae* thus far (Bester *et al*. 2008; Rhode *et al*. 2008; Rhode 2010). Not only does the development of SNPs guarantee a wider genome representation, but these markers will be advantageous in aiding the construction of a more comprehensive linkage map for *H. midae* (Bester *et al*. 2008; Rhode *et al*. 2008; Qi *et al*. 2009).

3 SNP Discovery in Non-Model Organisms

SNP marker development includes the discovery of *de novo* polymorphisms, followed by validation and characterisation of these identified markers in population groups (Chagné *et al*. 2008). Principal methods for SNP discovery include whole genome re-sequencing strategies or sequencing of DNA fragments from candidate genes in individuals that represent a specific population's diversity, known as the targeted gene approach. Disadvantages of the aforementioned methods include the prerequisite of a premium draft genome, cost and time consumption and low coverage when re-sequencing the whole genome (Kwok and Chen 2003; Le Dantec *et al*. 2004; Van Tassel *et al*. 2008). Due to these disadvantages, the development of SNPs in non-model organisms is still somewhat limited (Belfiore *et al*. 2003). SNP development is complicated without a reference genome since there is little means of verifying the presence of true SNPs. Conventional SNP isolation procedures in non-model organisms include AFLP-based analysis, the employment of sister taxa to develop sequencing primers as well as the construction and sequencing of reduced representation genomic or cDNA libraries (Meksem *et al*. 2001; Primmer *et al*. 2002; Bensch *et al*. 2002; Nicod and Largiader 2003; Bester *et al*. 2008). The development of ESTs has been successfully demonstrated in various fish and mollusc species including abalone (Zeng and Gong 2002; Bester *et al*. 2008; Qi *et al*. 2008; Wynne *et al*. 2008; Qi *et al*. 2009, 2010; Zhang *et al*. 2010).

Currently, SNP isolation procedures are greatly improved because of advances in DNA sequence technology. Recent approaches towards identifying SNPs include the utilisation of datasets generated from next generation sequencing (NGS) technologies. Platforms, such as the Illumina Genome Analyser and 454 Life Sciences' Genome Sequencer, expedites the generation of sequence data, by reducing the cost and time expenditure as well as increasing the production of sequence data to several thousand megabase pairs

(mb) (van Bers *et al*. 2010). Read lengths generated by these platforms allow sufficient assembly of contigs for non-model species (Kerstens *et al*. 2009; Renaut *et al*. 2010). Mining of SNPs in NGS generated ESTs mainly involve creating, clustering and assembly of unprocessed ESTs followed by the identification of SNPs either by means of *in vitro* or *in silico* approaches (Le Dantec *et al*. 2004). *In vitro* methods involve the re-sequencing of EST or genomic sequences to identify nucleotide variations; whereas *in silico* methods refer to the use of bioinformatics pipelines to identify polymorphisms (Useche *et al*. 2001). Programs such as autoSNP (Barker *et al*. 2003), PolyPhred (Nickerson *et al*. 1997) and PolyBayes have simplified automatic SNP identification significantly (Smith *et al*. 2005; Hubert *et al*. 2009). Methods such as these coupled with high-throughput genotyping techniques, allow for far more efficient means of SNP discovery and utilisation than previously.

4 SNP Genotyping

4.1 Introduction

Genotyping is mainly performed to determine the genetic differences between individuals and when typed in a number of individuals are used to validate polymorphism in a population. Over the years, various methods have been established in genotyping different markers, but it is the increase in SNP marker usage that has resulted in advances in SNP genotyping methodologies. Various criteria such as, simplicity of assay development, robustness, cost and time effectiveness, uncomplicated data analysis and easy automation are important aspects when developing the ideal genotyping method (Kwok 2001; Kwok and Chen 2003). Currently, there is no genotyping technique that complies with all of the above-mentioned criteria.

4.2 Genotyping Methodologies

SNP genotyping involves the discrimination and detection of allele-specific products (Kim and Misra 2007). Allele discrimination can be performed in two reaction formats, homogeneous or solid phase reactions, and are detected by various methods. Homogeneous assays refer to solution state reactions; no separation or purification is required, robustness is increased in solution and labour is minimal. A major constraint of

homogeneous solutions is its limitation in terms of multiplexing. Solid phase assays refer to the use of solid supports (microbeads, chips and microtitre plates) for allele discrimination. This kind of format allows for easy multiplexing, thus increasing throughput; unfortunately assay design and optimisation is costly and time inefficient (Kwok 2001).

4.2.1 Allele Detection Strategies

Allele detection is known as a non-sequence specific approach based on changes observed during electrophoresis such as mobility, cleavage and capture (Kwok 2001). The major detection methods are mass spectrometry, fluorescence and chemiluminescence. Mass spectrometry**,** such as Matrix-associated laser desorption time of flight mass spectrometry (MALDI-ToF MS), is the detection of changes in DNA based on molecular weight. Apart from the requirement of a particularly pure sample, considered a limitation, MALDI-ToF requires no labelled probes or primers, yields fast analysis of every sample and has the potential for multiplexing (Kwok 2001). Fluorescent detection is coupled with direct sequencing, where primers are labelled with different fluorescent dyes and each dye represents a different size product or a base nucleotide (Sanger *et al*. 1977). This is applicable to both homogenous solutions and solid supports (Kwok 2001). Other adaptations of fluorescence are fluorescence polarisation (FP) (Chen *et al*. 1999) and fluorescence resonance energy transfer (FRET) (Kim and Misra 2007). Due to its simplicity and versatility, all versions of fluorescent signal based detection can be applied to all the different allele discrimination techniques (Kwok 2001) and thus are utilised in most genotyping methodologies. The final detection strategy employed for SNP genotyping is chemiluminescence. It has several advantages including rapid detection, high signal-to-noise ratio and is well suited for automation (Sobrino *et al*. 2005). A well-known method that utilises chemiluminescence for the purposes of SNP genotyping is Pyrosequencing (Roche), a method thatcan determine between 400 bp - 800 bp of template DNA in real time (Robison 2011).

4.2.2 Allele Discrimination Strategies

Unlike allelic detection, allele discrimination relies on biochemical changes that occur due to DNA alterations and are therefore sequence-specific. The four basic strategies of allele

discrimination are primer extension, hybridisation, ligation and enzymatic cleavage (Kwok 2001).

4.2.2.1 Primer Extension

Primer extension is based on distinguishing between allelic variants caused either by a polymorphism (SNP) or by incorporating single dideoxynucleotide triphosphates (ddNTPs) to identify the polymorphic site, known as allele-specific PCR (AS-PCR) and minisequencing, respectively (Kwok 2001). Allele-specific PCR utilises a set of primers that are identical to each other except for their 3'-ends. Each contain the nucleotide with which the polymorphism (SNP) is distinguished and a fluorescently end-labelled universal reverse primer (Figure 1.3a) (Kim and Misra 2007). For the minisequencing approach, alternatively known as allele-specific nucleotide incorporation, a primer is designed to anneal adjacent to the 5'-end of the SNP and DNA polymerase incorporates a single nucleotide that is complementary to the SNP (Sokolov 1990) (Figure 1.3b). One such platform is the SNaPshot® assay (Applied Biosystems), where a primer flanks the SNP at the 5' end and extends by inserting single nucleotides.

Primer extension reactions are robust, easy to design and applicable to high-throughput screening as well as multiplexing, which makes it an easy and cost effective system (Syvänen 2001; Black and Vontas 2007). These assays are detectable by means of mass spectrometry (MALDI-ToF), chemiluminescence (pyrosequencing), fluorescence and melting curve analysis in either homogeneous or solid phase reactions (Hansson and Kawabe 2005; Kim and Misra 2007).

4.2.2.2 Hybridisation

This use of allele-specific hybridisation probes for genotyping was first demonstrated by Wallace *et al.* (1979). This method utilises two allele-specific probes that are designed to anneal to the target sequence, with a nucleotide in the centre of the probe complementary to the polymorphism (Figure 1.3c). The single base pair mismatch destabilises the probe and prevents it from hybridising to the sequence, making it the simplest genotyping method currently used. No enzymes are required and the thermal stability of the approach lies in the sophisticated probe design, SNP flanking sequence and optimisation of the

assay (Mir and Southern 1999; Sobrino *et al*. 2005). Although there are various hybridisation techniques used for genotyping SNPs, the most popular one seems to be the 5' exonuclease TaqMan® assay (Applied Biosystems).

4.2.2.3 Allele-specific Oligonucleotide Ligation

Oligonucleotide ligation assay (OLA) utilises the ability of a thermostable enzyme, *Thermus thermophilus* andDNA ligase, to covalently restore phosphodiester bonds in DNA. The application of the ligases allows for recurring ligation reactions, resulting in the exponential amplification of the template DNA (Barany 1991). The assay consists of four different components, two allele-specific probes that differ only in their 3'-ends complementarity to the SNP, DNA ligase and one universal probe that hybridises immediately downstream of the SNP (Figure 1.3d) (Sobrino *et al*. 2005; Kim and Misra 2007). Oligonucleotide ligation assay mainly employs FRET as a means of detection.

4.2.2.4 Invasive Cleavage

Enzymes that recognise certain structures and sequences can be employed for the discrimination of alleles when a SNP occurs within the recognition site of the enzyme. This forms the principal for invasive cleavage (Kim and Misra 2007). The method requires the use of probes that hybridise to the template DNA with the SNP at the overlapping site; when hybridisation occurs the correct structure is formed and invasive cleavage occurs. Invasive cleavage can be utilised for SNP genotyping with the use of RFLPs (Figure 1.3e) or the more sophisticated approach, the Invader® assay (Third Wave Technologies Inc.). This is a homogeneous isothermal assay that utilises a 5' endonuclease for specificity recognition and cleavage, called Flap endonuclease (Figure 1.3f) (Syvänen 2001). The genotype is inferred on agarose gel electrophoresis by size selection. Although this is relatively simple and requires no probes, it is limited to low-throughput genotyping (Kim and Misra 2007).

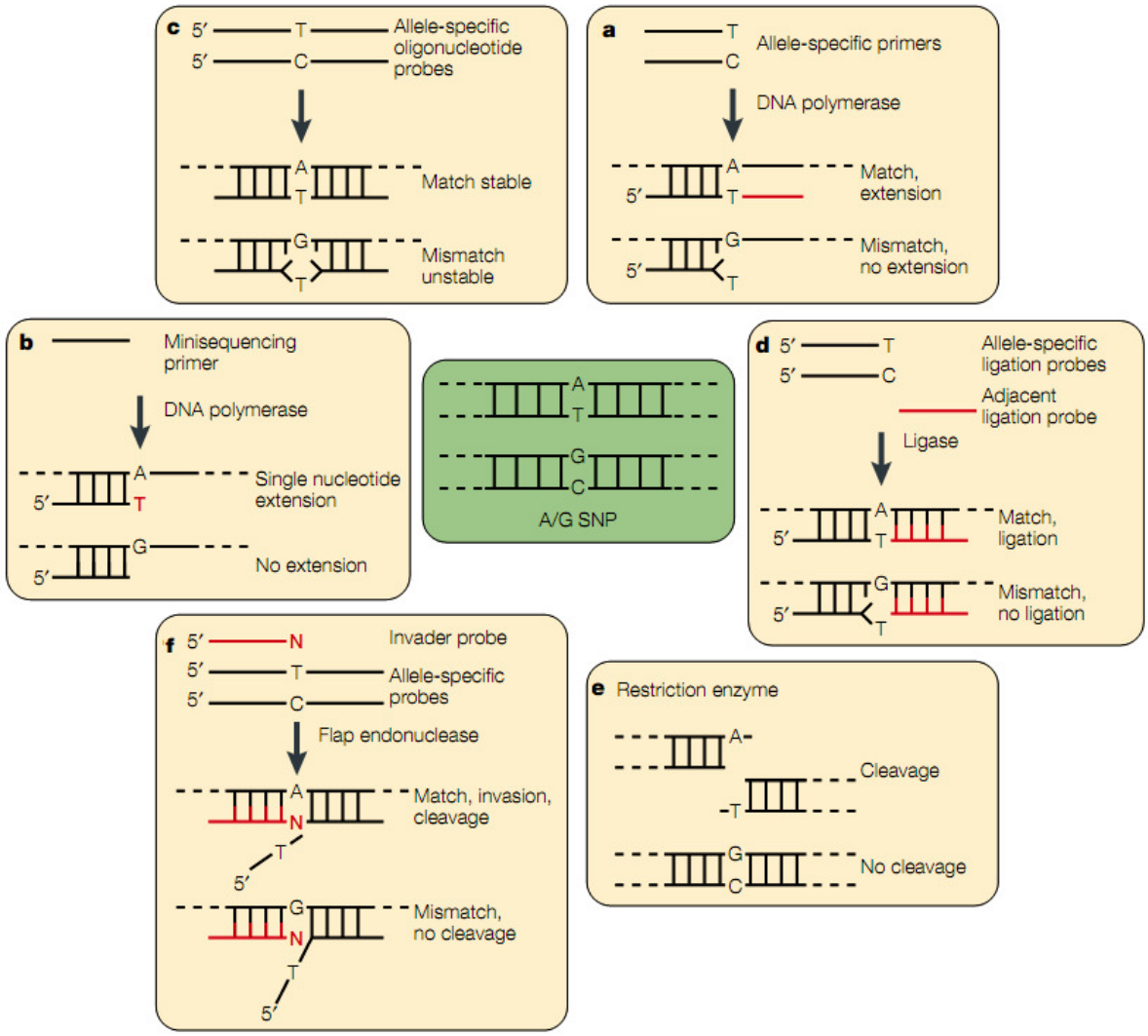

Figure 1.3.A modification of the allele discrimination methods taken from Syvänen (2001).

4.3 Illumina GoldenGate® Genotyping Assay with the VeraCode™ Technology on the BeadXpress®Platform

This genotyping system combines the established biochemistry of the GoldenGate genotyping assay with the novel VeraCode technology in the BeadXpress platform. The GoldenGate assay, a robust assay that has been successfully employed in large research projects such as the human HapMap project (The International HapMap Consortium 2003), includes locus identification by means of hybridisation, enzymatic allele discrimination and the exponential amplification of the target sequence (Lewis *et al*. 2007). Various degrees of multiplexing can be applied to the GoldenGate, which minimises cost and time (Illumina 2008). The novel VeraCode technology employs 28 x 240 micron-sized

cylindrical rods, referred to as silica glass microbeads that are inscribed with digital holographic barcodes (Illumina 2010). These act as a solid substrate in solution (Figure 1.4). The microbeads allow researchers to customise assays by pooling the beads in solution, which allow for the rapid reaction kinetics of solution-based assays. It also circumvents the decoding process that is associated with microarrays, making it suitable for smaller research projects (Lin *et al*. 2009). The final component of the genotyping assay, known as the BeadXpress Reader, is a platform that contains a dual-colour laser apparatus that scans over the microbeads and identifies the unique code within each bead.

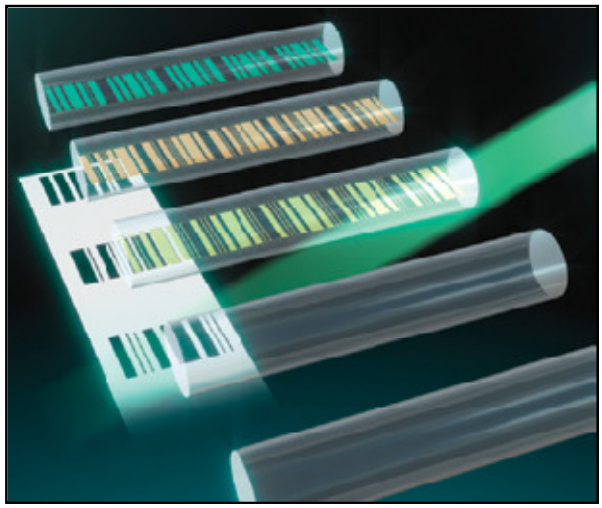

Figure 1.4. Silica glass microbeads embedded with holographic content (www.illumina.com).

The platform is highly robust, requiring only 250 ng of template DNA, which is activated to enable the binding of paramagnetic particles to the DNA (Figure 1.5, step 1). Oligonucleotides, hybridisation buffer and paramagnetic particles are then added to the activated DNA. This is followed by the binding of both the paramagnetic particles and oligonucleotides to the sample, referred to as the hybridisation step (Figure 1.5, step 2). Hybridisation occurs prior to amplification, eliminating any amplification bias. Three assay oligonucleotides are designed for each SNP locus. The first two oligonucleotides are Allele-Specific Oligonucleotides (ASO) followed by the third oligonucleotide, the Locus-Specific Oligonucleotide (LSO), which is complementary to the sequence, thus hybridising downstream from the SNP. All three oligonucleotides are complementary to the universal primers, but the LSO also has a unique address sequence that is complementary to a specific VeraCode bead. Several wash steps are implemented after hybridisation to

remove any non-hybridised and excess oligonucleotides. The third part in the procedure consists of the extension and ligation procedure, where the relevant ASOs are extended to the LSO (Figure 1.5, step 3). This results in a specific sequence (SNP) that is linked to the address sequence on the LSO and thus a genotype is inferred. The ligation product then serves as a PCR template in the universal PCR (Figure 1.5, step 4). The universal PCR consists of three primers (P1, P2 and P3), where the P1 and P2 are Cy3- and Cy5-labeled. Following downstream processing (Figure 1.5, step 5) the dye-labelled single stranded PCR amplicon is hybridised to its complementary VeraCodebead through the unique address sequence (Figure 1.5, step 6). Hybridisation allows the solution-based assay products to separate onto a solid surface (bead) for genotyping (Figure 1.5, step 7). The BeadXpress Reader analyses the fluorescent signal and holographic barcode on each bead (Figure 1.5, step 8) (Lewis *et al*. 2007). The intensity of fluorescence is used to infer a genotype. A heterozygous state is indicated by an equal intensity of both fluorescent dyes (Cy3/Cy5), whereas any other signal ratios (1:0 or 0:1) corresponds to alternate homozygous states (Akhunov *et al*. 2009). The flexibility and simplicity of the procedure allows several pause or stop steps and can be completed within two to three days.

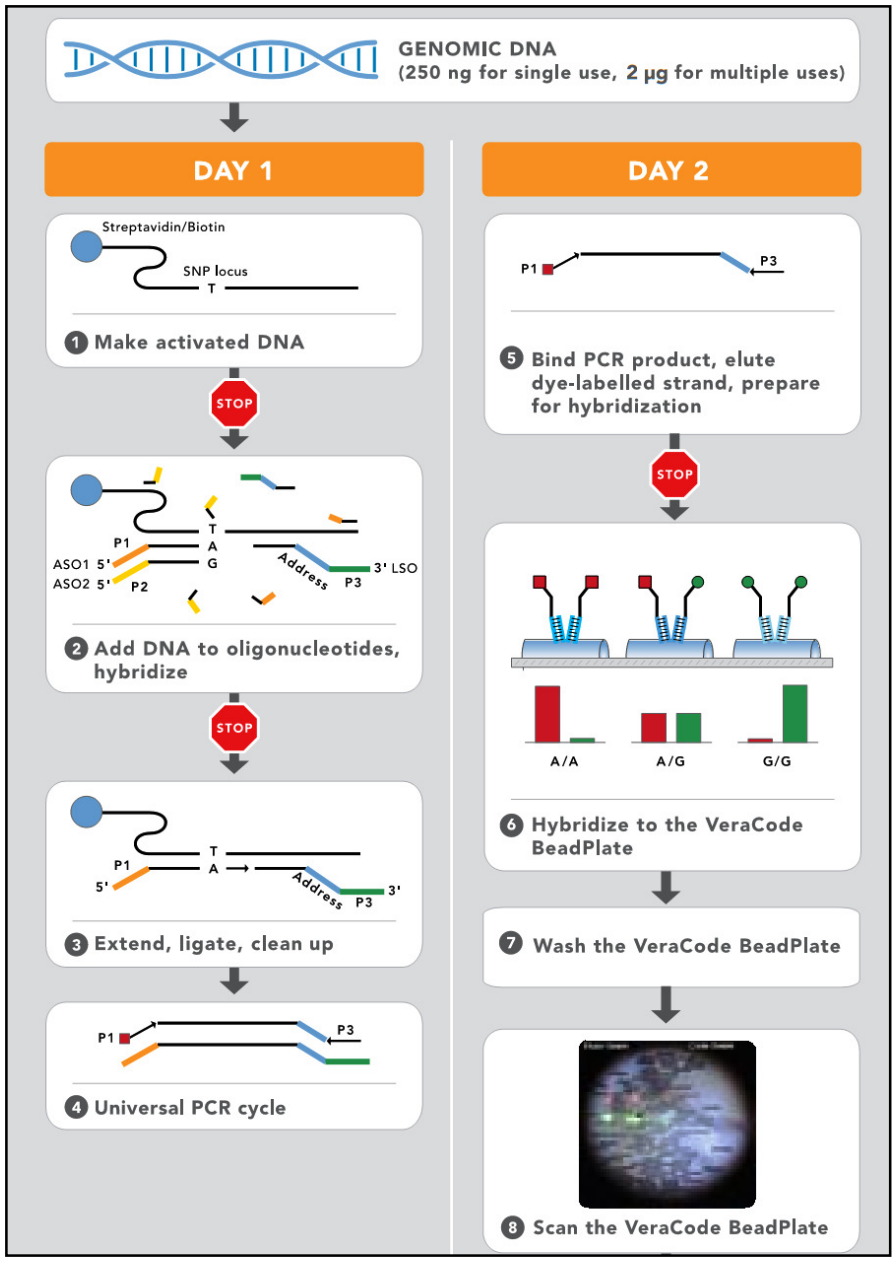

Figure 1.5. GoldenGate VeraCode assay (Illumina 2008).

Data generation involves the clustering of the fluorescent intensity values and generating a polar plot for each SNP, referred to as a GenoPlot (Figure 1.6) (Akhunov *et al*. 2009). The y-axis represents the normalised intensity, which is the sum of the signal intensities in the two channels that is normalised to explain any nominal variations, background differences and possible crosstalk between the dyes. The x-axis represents the theta value, indicating the allelic angle. Theta values near 0 and 1 are representative of the respective homozygotic states, whereas heterozygotes fall within this range. The values (in colour)

underneath each cluster represent the amount of samples that fall within a specific genotype (Lewis *et al*. 2007). The assay provides a high level of confidence as well as minimal room for error with the internal controls that monitor contamination, extension, amplification and hybridisation efficiency and annealing specificity (Illumina 2008).

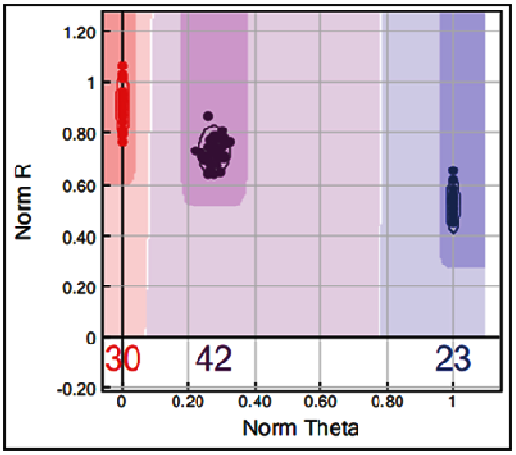

Figure 1.6.GenoPlot (Illumina 2008).

Aims and Objectives

The aim of this study is to identify and genotype SNP markers in the South African aquaculture species, *Haliotis midae*, and to determine the utility of these markers for population genetic studies.

1. Marker development will be facilitated through the application of transcriptome data generated from next generation sequencing of *H. midae*. The EST data generated will be assembled into contiguous sequences, thereby serving as the basis for primer design and subsequent SNP discovery. Primer efficiency will be tested on a panel of unrelated individuals, followed by sequencing by means of Sanger sequencing and subsequent *in vitro* discovery of SNPs.

2. Putative polymorphic markers will be genotyped in four linkage mapping families and three natural populations with the use of Illumina GoldenGate genotyping assay with VeraCode technology on the BeadXpress platform.

3. Data analysis will be performed to infer genetic diversity and population differentiation within and between the above-mentioned populations. This will serve as an indication of the utility of these SNPs for genetic diversity and population structure inference.

A linkage map is currently in development for *H. midae*; thus the development of the SNP markers in this study will prove important in the development of such a high resolution linkage map and will aid in the identification of QTLs linked to commercially important traits. With the availability of a linkage map, further research could ultimately lead to a more comprehensive understanding of the abalone genome and important traits for commercial sustainability, the latter of which will prove to be invaluable for the viability of the abalone industry in the South African context.

Chapter 2

Marker Development

1 Introduction

Genomic studies of non-model organisms have introduced an elevated demand for molecular markers. These studies will benefit more from molecular markers that are associated to coding regions, and thus markers that are linked to gene function, than anonymous markers (non-coding regions); and will therefore provide more insight into the genomes of non-model species. Unfortunately, knowledge of candidate genes is limited in such species, forcing researchers to rely on other means of identifying significant mutations. Consequently, the utilisation of Expressed Sequence Tags (ESTs) has increased due to the direct association of these tags to gene transcripts.

Earlier work on ESTs mostly involved the construction of cDNA libraries, which is a tedious and time-consuming endeavour. Newer strategies have moved to the use of next generation sequencing (NGS) technologies to facilitate the generation of these tags. Next generation sequencing technologies have paved the way for sequencing, genotyping and high-throughput marker discovery at an affordable rate (Stapley *et al*. 2010). It relies on the digestion of template DNA and sequencing of millions of short reads which is subsequently assembled into longer contiguous overlapping segments, referred to as contigs (Flicek and Birney 2009). The contigs are then used for marker discovery and stored in databases, such as Unigene. Next generation sequencing can be applied to transcriptomic or genomic studies, the former of which includes the synthesis of cDNA by reverse transcription of mRNA prior to sequencing (Ekblom and Galindo 2010). A popular application of NGS for non-model species includes transcriptome characterisation, the first of which was completed on the wasp (*Polistes meticus*) (Toth *et al.* 2007). Other applications include gene expression profiling, the construction of microarrays and alternative splicing research, to name a few (Ekblom and Galindo 2010; Harr and Turner 2010).

The use of EST sequence data for the development of markers, such as SNPs, provides an easier and inexpensive means of generating genomic resources for various species, including non-model organisms (Bouck and Vision 2007). A greater amount of functionally related markers can be isolated from ESTs due to the high redundancy of sequences that represents expressed genes (Picoult-Newberg *et al.* 1999). Another advantage of using ESTs is the interspecific transferability of markers, which is facilitated by readily available

EST databases of closely related species. This proves to be of great importance in non-model species, where knowledge of the genome is limited (Wang *et al*. 2010). Although ESTs are becoming a more popular resource for the isolation of SNPs, some factors still play a crucial role in the utility of these SNPs. These factors include the diversity of the selected species and its representation within the database, the range of tissue used for generating the EST database and the assembly depth of the EST sequences (Picoult-Newberg *et al*. 1999; Rafalski 2002; Ganal *et al*. 2009). Thus, this chapter focuses on the utilisation of ESTs from next generation sequencing of the *H. midae* transcriptome for SNP discovery and to evaluate the efficiency of generating markers from NGS data.

2 Experimental Design

2.1 EST Construction, Clustering and Sequence Assembly

Samples and techniques used for RNA extraction, cDNA library construction and EST contig assembly have been performed previously and are described in detail in Van der Merwe (2010). In brief, RNA extractions were performed on the soft tissue of 19 animals selected from a single family. mRNA molecules containing poly-A stretches were isolated, fragmented and copied into cDNA, to convert total RNA into a collection of template molecules suitable for sequencing on the Illumina Genome Analyzer II (GA II). Expressed Sequence Tags clustering and assembly of the sequencing data of the 19 samples was completed using Velvet v0.7.52 (k-mer size 23, expected coverage 100, cut off 5, insert length 250 and minimum length 80 base pairs) (Zerbino and Birney 2008; Franchini *et al*. 2011; Van der Merwe *et al*.2011). Sequences were annotated utilising dCAS v1.4; a cDNA annotation software package. Annotation was completed against the Eukaryotic Clusters of Genes (KOG) (Tatusov *et al*. 2003), Gene Ontology (GO) (Ashburner *et al.* 2000) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Ogata *et al*. 1999) databases.

2.2 EST Contiguous Sequence Selection and Annotation

Annotated contigs from the above-mentioned databases (KOG, GO and KEGG) were screened manually to identify contigs that illustrated significant hits against genes with known functions. A significant hit was defined as having an expected value (E-value) of less than 1.0×10^{-17}. Annotated contigs with significant hits to genes of interest in the

KOG, GO and KEGG databases were submitted to the National Centre of Biotechnology Information (NCBI) to determine if the equivalent genes were associated with specific gene function in aquaculture species. These BLASTN searches were conducted using the default parameters of the nr-nucleotide database on the National Centre of Biotechnology Information (NCBI) (Altschul *et al*. 1990) (http://blast.ncbi.nlm.nih.gov/Blast.cgi). A set of 58 annotated contigs that had an E-value < 1.3×10^{-19} and an identity value > 65% for BLASTN results in aquaculture species were selected for this study (Table 2.1).

2.3 Oligonucleotide Primers

2.3.1 Primer Design

The 58 contigs facilitated the design of ninety-seven primer pairs (Table S1) using BatchPrimer3 software (http://probes.pw.usda.gov/cgi-bin/batchprimer3/batchprimer3.cgi) with the following criteria: primer size ≥ 20 bp, maximum Tm difference 1.0°C, Tm 55.0°C – 60.0°C, GC content 40.0% – 60.0% (You *et al.* 2008). To adhere to maximum sequence amplifiable lengths, 20 of the 58 contigs were manually fragmented into smaller sections of no more than 700 bp. The sections from the same contig were fragmented in such a way to include an overlapping sequence over the first and last 77 bp of each section. This allowed the design of a subset of primers for each contig that would overlie at these overlapping sequences. This was done to ensure that the entire contig could be amplified by the subset of primers.

2.3.2 Primer Optimisation

Primers were optimised and PCR conditions for each primer are indicated in Table S1. PCR optimisation reactions contained 20 ng genomic DNA, 200 µM dNTPs (ABgene), 2.0 mM $MgCl_2$ (Promega), 2.0 pmol of each primer (WhiteSci) and 0.25 U Go*Taq*® Flexi DNA polymerase (Promega) in a total volume of 10 µl for each primer pair. Amplification cycles included an initial denaturing step of 95°C for 5 min; 30 cycles of 94°C for 30 sec, Tm (specific for each primer pair) for 45 sec, 72°C for 45 sec; and a final elongation step of 10 min at 72°C in the Gene-Amp System 2700 thermal cycler (Applied Biosystems).

Primers that showed non-specific amplification or failed to amplify were optimised using the Px2 Thermal Cycler (Thermo Electron Corporation) with three different $MgCl_2$ concentrations, 1.5 mM, 2.0 mM and 2.5 mM as follows: 95°C for 5 min; 30 cycles of 95°C for 30 sec, 50°C for 30 sec and 72°C for 45 sec; followed by a final step of 72°C for 7 min. A temperature range of 50.1°C - 65.2°C was established by the Px2 Thermal Cycler (Thermo Electron Corporation). Primers that amplified a single PCR amplicon were considered as optimised. Primer pairs that could not be optimised were deemed unusable for this study and were thus excluded.

2.4 Primer Verification

2.4.1 Polymerase Chain Reactions and Conditions

Optimised primers were amplified in a panel of eight unrelated (randomly selected) *Haliotis midae* individuals. For amplification of genomic DNA, the following PCR conditions were used for each primer pair: 20 ng template DNA, 2.0 mM $MgCl_2$ (Promega), 200 µM dNTPs (ABgene), 2.0 pmol of each primer (WhiteSci) and 0.25 U Go*Taq*® Flexi DNA polymerase (Promega) in a 10 µl reaction volume. Amplification cycles included an initial denaturing step of 95°C for 5 min; 35 cycles of 94°C for 30 sec; Tm (specific for each primer pair) for 45 sec; 72°C for 45 sec; and a final elongation step of 10 min at 72°C in the Gene-Amp System 2700 thermal cycler (Applied Biosystems).

2.4.2 Agarose Gel Electrophoresis

Assessment of PCR amplification was conducted through 2% (w/v) horizontal agarose gel electrophoresis and visualised using 0.1% (v/v) Ethidium Bromide. Three microlitres of PCR product was combined with 2 µl of cresol red loading buffer (Appendix A). A 100 bp ladder (1 µl ladder mixed with 2 µl blue/orange 6x loading dye) (Promega) was added as a molecular size marker to visually determine if amplicons of the correct size were generated. PCR amplicons were electrophoresed at 100V for 1.5 hr and visualised with ultraviolet light transillumination in the GeneSnap MultiGenius Bio Imaging System (Syngene).

2.5 Semi-automated Sequencing and Analysis

2.5.1 DNA Purification and Sequencing

Post-PCR purification of the PCR amplicons of all eight individuals for each primer pair was conducted with the SigmaSpin™ Post-reaction cleanup kit (Sigma) according to manufacturer's specifications. Sequencing reactions were carried out by means of standard Sanger sequencing using the BigDye® Terminator v3 cycle sequencing kit (Applied Biosystems). DNA quantification for each purified sample was performed prior to sequencing. Purified PCR amplicons were sequenced from the 5'-end with the respective forward primer. Amplicons that failed to sequence from the 5'-end were also sequenced from the 3'-end with the respective reverse primer. Sequencing reactions were sent for capillary electrophoresis on the ABI PRISM® 3100 DNA automated sequencer (Applied Biosystems) at the Central Analytical Facility (DNA sequencing unit) at Stellenbosch University.

2.5.2 Sequence Quality Control

DNA sequence chromatograms for each primer pair were manually analysed using Sequence Scanner v1.0 (Applied Biosystems) to determine sequence quality. Sequences with a poor signal-to-noise ratio were omitted from further analysis. Successfully sequenced sequences from all eight individuals for each primer pair were aligned to each primer pairs' original contig with ClustalW v1.4 (available in the BioEdit v7.0.9.0 software package) and annotated using BLASTN to determine that the correct amplicons were amplified and sequenced (Thompson *et al*. 1994; Hall 1999). Subsequently sequences suitable for SNP discovery were trimmed at the endpoints where base calling had dropped off (Figure 2.1).

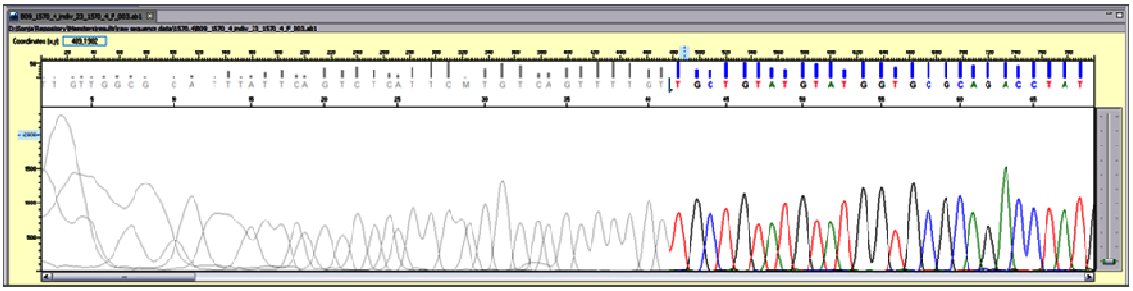

Figure 2.1. DNA chromatogram depicting a sequence with a trimmed 5'-endpoint.

2.6 PutativeSNP discovery

The concatenation of fragmented sections was prohibited due to the failed optimisation of some of the primer pairs that formed part of the subset of primers that were designed for a particular contig. Therefore, amplified sections from each primer pair were considered as a contig and analysed as such. Trimmed sequences from the eight individuals representing a contig were aligned (reverse compliment where applicable) with ClustalW v1.4 (in BioEdit v7.0.9.0) using default parameters (Thompson *et al*. 1994; Hall 1999). This was completed for all the optimised primer pairs. Alignments were visually screened for putative SNPs. To eliminate false positives, verification of putative SNPs was done manually using chromatograms of all eight individuals for each primer pair. Only base positions with two peaks, of which the height ratio was approximately ≥ 1:2, were considered as putative SNP loci for heterozygous individuals.

3 Results

3.1 EST Construction, Clustering and Contig Assembly

The Velvet assembly of transcriptome data of 19 individuals generated by Illumina Genome Analyser (GA II) yielded 127 687 contigs with a maximum length of 5 740 nucleotides and a mean length of 276 bp. Only 30 689 of these contigs that were more than 80 bp in length and comprised of at least two reads were used for annotation and further analysis (Franchini *et al*. 2011; Van der Merwe *et al*.2011).

3.2 EST Contiguous Sequence Selection and Annotation

A set of 58 contigs that adhered to above-mentioned criteria was selected and searched for significant similarity against genes of known function (Table 2.1). BLAST hits indicated that 51 (~88%) of the contigs showed significant similarity to genes of interest in aquaculture species at an E-value <1.0 x 10^{-19}. The majority of hits (48%) were classified in the Mollusca phylum followed by 38% hits in the phylum Chordata, and only 14% in other phyla (Figure 2.2). Within the phylum Mollusca, 82% were classed as Gastropoda and 18% as Bivalvia, of which 71% of the contigs showed significant similarity to other Haliotid species. Seven (12%) of the contigs did not show significant similarity to any

cultured fish or shellfish species. These, however, did show significant similarity to genes such as Glyceraldehyde-3-phosphate dehydrogenase and Electron transfer flavoprotein and were therefore still included as they conferred important functions in other organisms, including glycolysis and electron transfer (http://www.ebi.ac.uk [accessed August 2011]).

Table 2.1. Characterisation and annotation of 58 contigs generated from the *Haliotis midae* transcriptome using BLASTN.

Assembled Contig	Organism		E-value	Genbank Accession number	Significant similarity	Identity (%)	Coverage (%)
	Scientific name	Common name					
Conitg_30	*Danio rerio*	Zebrafish	0	NM_199218.1	S-adenosylhomocysteine hydrolase	72	97
Conitg_54	*Salmo salar*	Atlantic salmon	1e-171	BT059645.1	Elongation factor 2	77	92
Conitg_101	*Pecten maximus*	King scallop	3e-58	AF134172.1	Myosin heavy chain mRNA	77	99
Conitg_146	*Haliotis diversicolor*	Japanese abalone	3e-80	EU244376.1	ADP/ATP carrier protein	75	25
Conitg_149	*Haliotis tuberculata*	Green ormer	0	AM283516.1	Heat shock protein (hsp71 gene)	90	99
Conitg_153	*Pyrus pyrifolia*	Chinese sand pear	0	AB266449.1	Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	94	100
Conitg_210	*Haliotis tuberculata*	Green ormer	7e-49	FN566841.1	Fibrillar collagen (col1A1 gene)	93	98
Conitg_214	*Aplysia californica*	California sea slug	0	S51239.1	Calreticulin	79	66
Conitg_216	*Haliotis diversicolor supertexta*	Taiwanese abalone	3e-161	FJ446700.1	QM-like protein mRNA	96	100
Conitg_237	*Danio rerio*	Zebrafish	0	NM_001002055.1	COP9 constitutive photomorphogenic homolog subunit	76	70

Conitg_342	*Haliotis diversicolor*	Japanese abalone	0	FJ812177.1	Heat shock inducible protein 70 mRNA	94	99
Conitg_379	*Danio rerio*	Zebrafish	6e-124	BC063965.1	Eukaryotic translation elongation factor 2	73	81
Conitg_393	*Danio rerio*	Zebrafish	1e-175	NM_001190982.1	Tubulin, alpha 1, like 2	84	90
Conitg_449	*Haliotis rufescens*	Red abalone	4e-112	DQ087489.1	Methionine adenosyltransferase mRNA	92	40
Conitg_516	*Haliotis discus discus*	Disk abalone	2e-157	EF103424.1	Ribosomal protein S14 mRNA	97	99
Conitg_569	*Lymnaea stagnalis*	Great pond snail	0	Z23105.1	G protein beta subunit	81	18
Conitg_783	*Haliotis cracherodii*	Black abalone	2e-124	EF650053.1	Rab7 mRNA	93	32
Conitg_984	*Danio rerio*	Zebrafish	2e-116	AY398322.1	Cell division cycle 42 (CDC42) mRNA	78	71
Conitg_1030	*Salmo salar*	Atlantic salmon	0	BT072562.1	ATP-dependent RNA helicase DDX39	97	86
Conitg_1095	*Strongylocentrotus purpuratus*	California purple sea urchin	0	XM_001177920.1	Similar to Peptidase	72	76
Conitg_1264	*Salmo salar*	Atlantic salmon	4e-19	NM_001141008.1	Ribonucleoprotein complex subunit 1 putative mRNA	68	100
Conitg_1527	*Haliotis discus discus*	Disk abalone	0	EF103366.1	Calcineurin A mRNA	99	99

Conitg_1570	*Placopecten magellanicus*	Sea scallop	3e-63	AF175578.2	Omega-crystallin mRNA	68	95
Conitg_1659	*Haliotis discus discus*	Disk abalone	7e-56	EU247757.1	Ribosomal protein S9 mRNA	93	100
Conitg_1718	*Haliotis tuberculata*	Green ormer	1e-68	AY941073.1	HTUB1 tRNA-Leu gene	87	100
Conitg_1827	*Haliotis discus discus*	Disk abalone	8e-97	EF103429.1	Ribosomal protein l mRNA	91	100
Conitg_1833	*Salmo salar*	Atlantic salmon	0	BT045702.1	Tubulin alpha chain putative mRNA	86	100
Conitg_1834	*Aplysia californica*	California sea slug	0	AF481056.1	Alpha tubulin 2 mRNA	82	100
Conitg_1881	*Haliotis diversicolor*	Japanese abalone	2e-92	EF553516.1	Elongation factor 1 alpha (EF1a) mRNA	96	92
Conitg_1949	*Chlamys farreri*	Zhikong scallop	3e-79	AF526241.1	Host defence genes	96	98
Conitg_2050	*Danio rerio*	Zebrafish	5e-131	NM_213318.1	Linked to protein phosphatase 2	78	73
Conitg_2187	*Haliotis discus discus*	Disk abalone	0	FJ380208.1	Histone H2A isoform 2 mRNA	100	98
Conitg_2962	*Danio rerio*	Zebrafish	4e-86	AF128240.1	Ubiquitin-conjugating enzyme 9 (ubc9) mRNA	77	75
Conitg_3112	*Haliotis discus discus*	Disk abalone	0	DQ530211.1	Catalase mRNA	85	91

Conitg_3516	*Salmo salar*	Atlantic salmon	0	BT059085.1	Medium-chain specific acyl-CoA dehydrogenase	73	62
Conitg_3566	*Pinctada fucata*	Pearl oyster	3E-102	AB354635.1	PfAF mRNA for antisecretory factor-like protein	72	95
Conitg_4691	*Crassostrea gigas*	Pacific oyster	7e-76	AB122063.1	HSP70 mRNA for 70kDa heat shock protein	73	98
Conitg_5153	*Salmo salar*	Atlantic salmon	1e-96	NM_001140183.1	Eukaryotic translation initiation factor 2 subunit 1	69	68
Conitg_5470	*Oncorhynchus mykiss*	Rainbow trout	9e-65	NM_001124685.1	Cystathionine gamma-lyase inhibitor (LOC100136725)	67	90
Conitg_5553	*Salmo salar*	Atlantic salmon	0	BT059030.1	Glucose-6-phosphate isomerase	72	84
Conitg_5879	*Ixodes scapularis*	Deer tick	5e-119	XM_002412644.1	Electron transfer flavoprotein, beta subunit	73	78
Conitg_7544	*Haliotis asinina*	Donkey's ear abalone	2e-90	AF231942.2	Class III POU protein mRNA	97	99
Conitg_7614	*Salmo salar*	Atlantic salmon	4e-144	NM_001140349.1	Coatomer subunit beta (copb)	76	94
Conitg_8704	*Aedes aegypti*	Yellow fever mosquito	2e-65	XM_001651344.1	Glycyl-tRNA synthetase partial mRNA	69	94
Conitg_11848	*Salmo salar*	Atlantic salmon	0	BT045352.1	Aldehyde dehydrogenase	71	61
Conitg_11960	*Haliotis discus discus*	Disk abalone	0	EF103413.1	26S protease regulatory subunit 6B mRNA	94	69

Conitg_13064	*Homo sapiens*	Human	2E-75	NM_006842.2	Splicing factor 3b, subunit 2	79	98
Conitg_14549	*Branchiostoma belcheri*	Japanese lancelet	3E-100	AF397147.1	G10-like protein mRNA	81	71
Conitg_15845	*Haliotis discus discus*	Disk abalone	0	EF103387.1	Proteasome alpha type 2 mRNA,	94	99
Conitg_16644	*Salmo salar*	Atlantic salmon	0	BT058697.1	NADH dehydrogenase flavoprotein 1	74	63
Conitg_17023	*Salmo salar*	Atlantic salmon	0	BT058697.1	NADH dehydrogenase flavoprotein 1	74	63
Conitg_17550	*Danio rerio*	Zebrafish	0	NM_001005391.2	Clathrin	77	60
Conitg_17935	*Salmo salar*	Atlantic salmon	7E-52	NM_001139952.1	Carboxypeptidase, vitellogenic-like (cpvl)	80	49
Conitg_18559	*Haliotis diversicolor supertexta*	Taiwanese abalone	1E-24	FJ943416.1	Cathepsin L-like cysteine proteinase	75	26
Conitg_20089	*Homo sapiens*	Human	2E-55	AK316444.1	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A member 5	78	95
Conitg_21833	*Haliotis discus discus*	Disk abalone	0	EF103382.1	l-3-hydroxyacyl-coenzyme a dehydrogenase	91	91
Conitg_21867	*Danio rerio*	Zebrafish	3E-135	BC071366.1	Acyl-Coenzyme A dehydrogenase	72	38
Conitg_22644	*Haliotis discus discus*	Disk abalone	0	EF103362.1	Chaperonin containing tcp1 mRNA	91	94

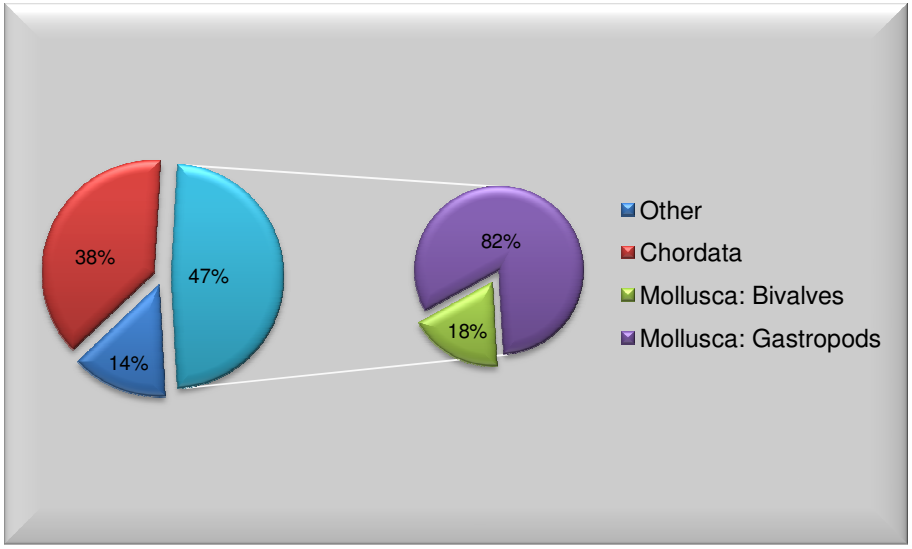

Figure 2.2. Classification of 28 contigs that showed significant similarity to the Mollusca phylum, classified into Bivalves and Gastropods.

3.3 Primer Optimisation and Verification

Fifty-eight contigs, of which 20 were fragmented, facilitated the design of 97 primer pairs in total. Only 38 (39%) of the primer pairs were successfully optimised as indicated by the presence of a single PCR product. Non-optimised primers were regarded as primers that produced multiple products, failed to amplify any product or showed non-specific amplification; 59 of the primer pairs were regarded as such (Figure 2.3). Twenty-one (22%) of the original primer pairs consistently amplified multiple products. These showed sequence similarity using BLAST to multigene families including catalase, histones and acyl-CoA dehydrogenase and were excluded from further analysis.

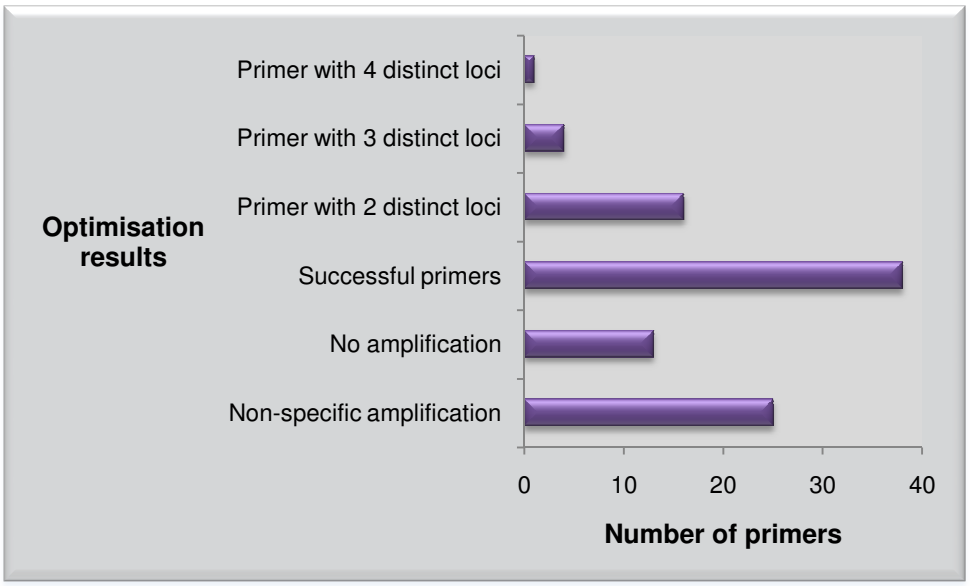

Figure 2.3. Optimisation results of the 97 primer pairs.

Fifty percent of the 38 optimised primer pairs amplified the expected size fragment in a panel of unrelated individuals (Figure 2.4) and 45% of the primer pairs amplified products that exceeded the predicted EST sequence lengths. The amplification of larger than expected amplicons could be attributed to the presence of an intron or multiple introns that were not present in the original EST contig. The majority of these possible introns were found to be between 300 bp – 400 bp in size. The remaining two (5%) successfully optimised primers amplified fragments that were shorter than predicted from the EST sequences.

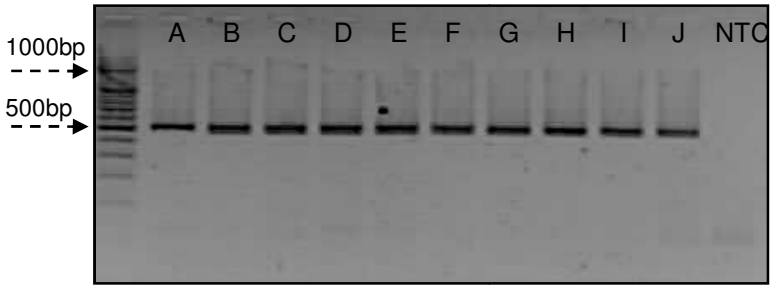

Figure 2.4.Verification of primer pair 1570.4 in panel of ten unrelated (A-J) individuals. The product size was determined with a 100 bp ladder and No Template Control (NTC) was used as an indication of contamination.

3.4 Semi-automated Sequencing and Analysis

PCR amplicons of the successfully optimised primer pairs were subjected to automated sequencing using capillary electrophoresis, of which 66% yielded trace quality that were adequate for putative SNP discovery. The remaining 34% of the primer pairs failed; possibly due to slippage of the *Taq* polymerase after a homopolymer or repeat region (3), non-recognisable sequences (6) and/or possible insertion/deletions (indels) or isoforms (4) (Figure 2.5).

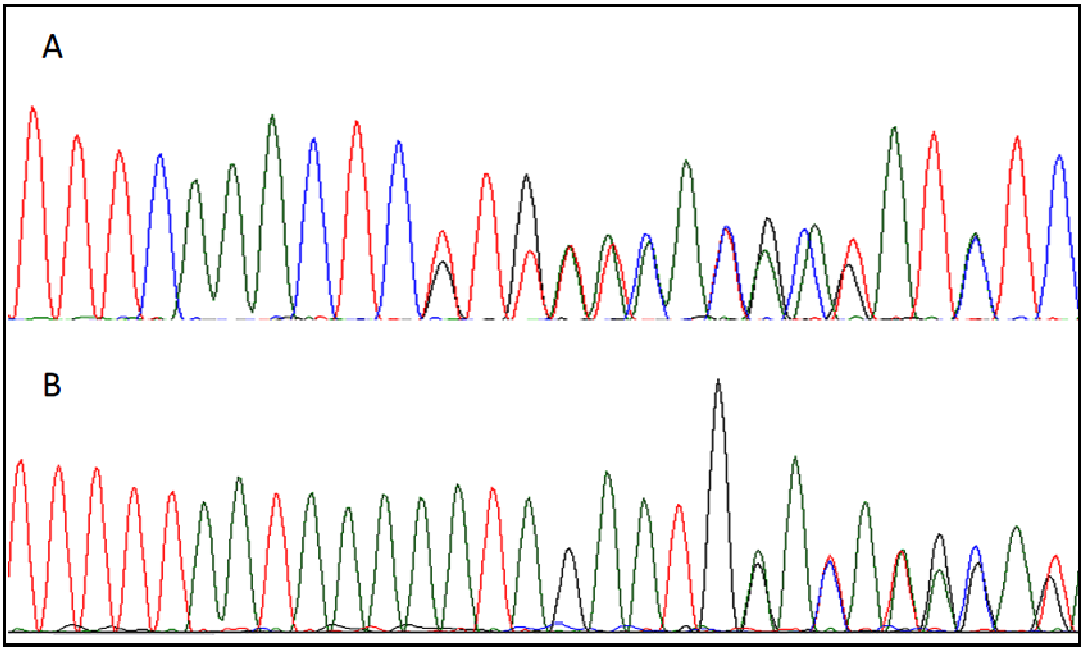

Figure 2.5. A representation of a failed sequence due to a possible indel or frameshift (A) and a homopolymer region indicated by a perfect sequence before the homopolymer region, after which the sequence degrades to multiple peaks (B).

3.5 *De novo* SNP Discovery

Putative SNPs were manually identified from the alignments created through ClustalX in the BioEdit software (Figure 2.6).

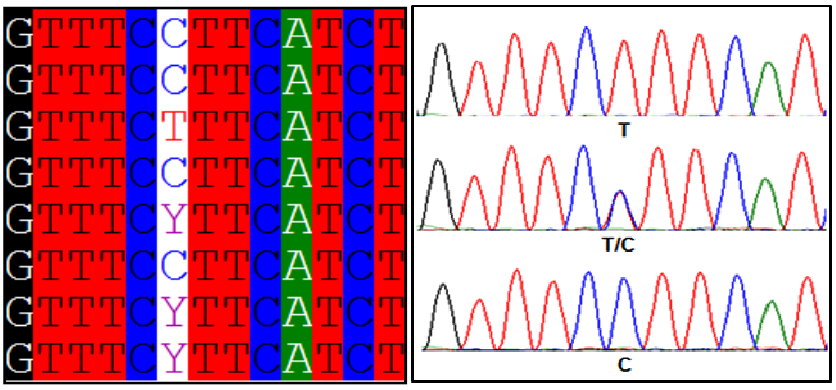

Figure 2.6.A representation of a Single Nucleotide Polymorphism. The left depicts the detection of a SNP in a multiple alignment and verified with the use of electropherograms (right). This particular SNP is a pyrimidine substitution, illustrating individuals homozygous for TT and CC as well as heterozygous individuals (TC), which is illustrated with the double peak or then as a Y in the multiple alignment.

Approximately 9.25 kb of sequence data was used for the discovery of SNPs within the *Haliotis midae* genome. No association was observed between the fragment length and the number of putative SNPs detected. Studies in catfish have found that the frequency of SNP loci increased with the increase of the contig size (Wang *et al*. 2008) and would therefore suggest that the frequency of SNPs would increase in data sets containing larger contig sizes. This observation was not supported in this study as illustrated in figure 2.7 and could be explained by the relatively small quantity of contigs originally used for SNP discovery.

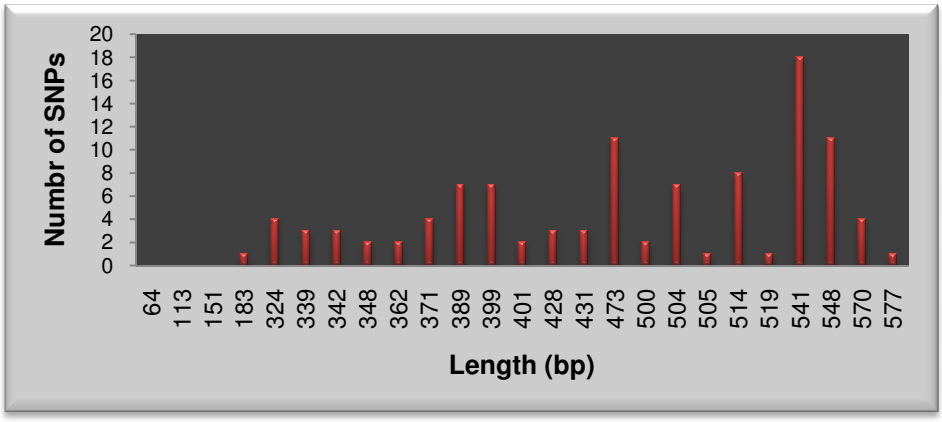

Figure 2.7.SNP density in various lengths of sequenced products.

For this study, a total of 65 (62%) transitions and 38 (36%) transversions were observed, giving an observed ts:tv ratio of 1.71 (Table 2.2). Only two polymorphisms were indicated as tri-allelic SNPs.

Table 2.2: Summary of putative SNP discovery in EST-contigs in *H. midae.*

Number of contigs	58	
SNP-containing fragments	25	
Number of putative SNPs	105	
Transversions		
A/T	15	(14.3%)
A/C	7	(6.7%)
C/G	10	(9.5%)
T/G	6	(5.7%)
Transitions		
A/G	35	(33.3%)
T/C	30	(28.6%)
Other	2	(1.9%)
Average SNP density	1%	

4 Discussion

The development of cDNA libraries and ESTs is a relatively uncomplicated means of generating genetic information in non-model organisms. These ESTs represent transcribed mRNA of the genome that is present in various developmental stages or specific tissue types (Rudd 2003). It allows information to be generated with regards to functional genomic regions, which is especially useful for the identification of markers for the application of linkage and QTL studies for species devoid of genomic information.

A few SNP discovery studies making use of ESTs have previously been conducted on *H. midae.* These studies were based on a targeted EST approach with the use of a cDNA library (Bester *et al.* 2008) and a comparative genome study that relied on *in silico* mining of ESTs from related haliotid species (Rhode 2010). Although ESTs are considered an excellent source for SNP isolation, the yield of these markers is greatly affected by the number of ESTs screened. This is evident from the above-mentioned studies in which only 28 SNPs were

collectively developed from ESTs and illustrates the need for producing more EST sequences.

With the rapid evolution of NGS methodologies, production of high volumes of EST sequences has become more feasible. It has also reduced the need to utilise first generation methods such as Sanger dideoxy sequencing (Sanger *et al*. 1977) for the sequencing of genomes and transcriptomes, which is a labour-intensive and a costly endeavour (Shendure and Ji 2008; Metzker 2010). Next generation sequencing will soon become the preferred method for undertaking genome or transcriptome characterisation studies as it not only generates larger amounts of data in a fraction of the time (Varshney *et al*. 2009) but also provides data for large-scale marker discovery and profiling of gene variation. Although the majority of studies in non-model species rather opt for transcriptome characterisation, genomic studies of these species can be aided with comparative genomics due to the large amount of genome scanning currently underway in model species. This proves useful in applications such as identifying genes of interest and loci under selection as well as researching gene regulation (Quinn *et al*. 2008; Ekblom and Galindo 2010).

This section of the study investigated the utility of ESTs generated by Illumina sequencing-by-synthesis for the development of SNPs in *Haliotis midae*. Due to the limited knowledge of the complete genome of *H. midae*, sequencing of the transcriptome was more feasible than whole genome sequencing, as it promised to provide greater insight into the functional characteristics of this species (non-model) at an affordable rate. Transcriptome sequencing has also shown promising results in previous aquaculture studies including Lake whitefish (*Coregonus* spp.) (Renaut *et al.* 2010), catfish (*Loricaria cataphracta)* (Surget-Groba and Montoya-Burgos 2010) and Atlantic salmon (*Salmo salar*) (Quinn *et al*. 2008). In a previous study, sequencing-by-synthesis yielded over 25 million short reads (75 bp - 100 bp); of which all was submit to the Velvet v0.7.52 for *de novo* assembly. Over 127 000 contigs of no less than 100 bp were generated and used for further annotation and marker discovery. Detailed information and results are discussed in Van der Merwe (2010) and Franchini *et al*. (2011). In the current study, 58 contigs were selected for annotation, of which 88% showed significant similarity to genes of interest in cultured fish and shellfish species. The contigs were selected based on genes of relevant function in aquaculture species, which included a vast selection of

functions ranging from cellular processes to stress response. This extended range of functional genes was chosen to compensate for the limited knowledge of the *H. midae* genome. The largest number (20.7%) of the hits was against Atlantic salmon (*Salmo salar*), illustrating its use as a model organism for genetic research; the frequency of hits correlating with the abundance of publicly available ESTs for this species (498 212) (NCBI's dbEST, 20 October 2011). ESTs in aquaculture species in general are however still underrepresented in comparison to other commercially bred species such as cattle (*Bos taurus*, 1 559 494 ESTs) (NCBI's dbEST, 20 October 2011).

Primers were designed to amplify contigs as well as various shorter segments within longer contigs. For this study a primer success rate (amplifiable primers) of 61% was obtained, which correlates with other aquaculture studies regarding the development of SNPs from ESTs, such as the Eastern oyster (*Crassostrea virginica*) (Zhang and Guo 2010) (69%), Pearl mussel (*Hyriopsis cumingii*) (58%) (Bai *et al*. 2009) and Pacific abalone (*Haliotis discus hannai*) (67.3%) (Qi *et al*. 2009). A recent study in Pacific oyster (Crassostrea gigas) showed a marked (~30%) increase in primer success rate, by designing primers that amplify smaller products (100 bp - 350 bp) of which the forward primer is positioned in the 3'-end of the coding region and the reverse primer in the 3' untranslated region (UTR)(Kim *et al*. 2010). A similar rationale was followed in the current study on the premise that introns are highly infrequent in the 3'-UTR (Scofield *et al*. 2007) and that 3'-end sequences correspond with terminal exons (Krizman and Berget 1993). The amplification of untranslated regions can be beneficial since it is hypothesized that these regions are subject to lower selection pressures and would therefore, yield higher polymorphic amplicons (Chakravarti 1999; Wang *et al*. 2008). In the current study, primers that consistently amplified more than one product were found to have significant similarity with genes that belong to multigene families such as catalase, histones and acyl-CoA dehydrogenase. These families originate from gene duplication events, with the genes retaining a certain degree of similarity. They can either be dispersed within a genome or remain in clusters, act as functional genes or pseudogenes (Ramos-Onsins and Aguadé 1998). The best means to avoid amplifying various members of a gene family is to utilise non-coding regions and 3' and 5'-UTRs of the ESTs for primer design (Horton *et al*. 1997). Due to a lack of genomic information for *H. midae*, further analysis is needed to verify that these fragments/loci are from multigene families. Several

PCR products were noted to be smaller than the expected size; possible reasons for this could be internal deletions or alternative splicing events (Yap 2001; Matsuzaki *et al*. 2005).

Although ESTs are a reliable source for marker development, these sequences only represent a portion of the original transcript (Bouck and Vision 2007). This limitation is coupled with the lack of information regarding gene order, the position of regulatory motifs and introns. The difficulty with having no information regarding the position of introns becomes imperative when designing primers. When a primer is designed over an intron-exon boundary, amplification in genomic DNA will fail due to the primer sequence that cannot bind to its complementary sequence (Figure 2.8).

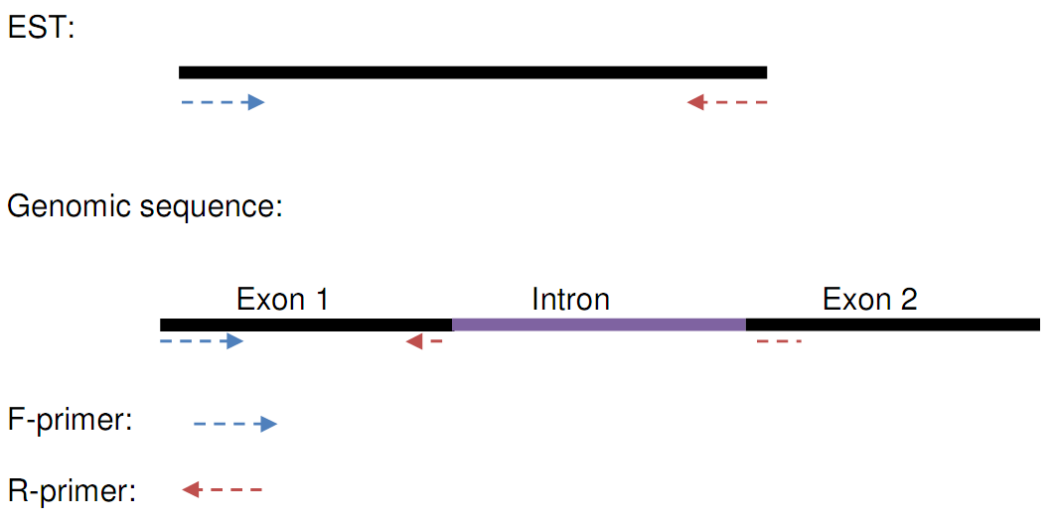

Figure 2.8.A schematic representation of the effect of designing a primer over an intron-exon boundary.

The latter can be circumvented by aligning ESTs to homologous genes (from related species) in which intron positions are known. However, evolutionary divergence may restrict the success of this (Bouck and Vision 2007) and with no complete genomic reference in any Haliotid species; this is still difficult to overcome. The problem is exacerbated in species that are intron-rich and highly polymorphic, for example lophotrochozoans (Raible *et al*. 2005), which includes molluscs such as oysters, mussels and limpets (Kim *et al*. 2010) and various gastropods. Extremely large introns and polymorphisms that occur within priming sites also hampers primer design from EST sequences by prohibiting efficient amplification (Kim *et al*. 2010). Other limitations include the difficulty in identifying alternative splice forms and

different alleles when clustering of ESTs without any reference genome (Wang *et al*. 2004). These problems are limiting factors in itself, but are exacerbated when utilising EST data in non-model organisms such as *H. midae*.

In the current study, a significant proportion (34%) of the sequences could not be used for SNP discovery due to indels or homopolymer regions. The latter occurs when large poly T or A stretches are prevalent within the sequence, causing “slippage” to occur. This mechanism is not fully understood but it is hypothesised to occur when the DNA double helix does not stay paired during *Taq polymerase* polymerisation. This can be circumvented by performing bidirectional sequencing and cloning PCR fragments into vectors prior to sequencing(Krangel and Langdon 2011). Indels are a great source of genetic variation, but pose analytical problems in direct sequencing. These problems manifest as phase-shifted signals that causes overlapping peaks after and the indel region (www.nucleics.com).

A total of 105 putative SNPs were developed from 25 fragments during the current study. This signifies a SNP frequency of ~1% or alternatively, one SNP every 88 bp. Single nucleotide polymorphism frequencies in *H. midae* have previously been reported to range from one SNP every 100 bp - 200 bp {Bester *et al*. 2008 (185 bp); Rhode *et al*. 2008 (113 bp); Rhode 2010 (150 bp)}. The frequency of putative SNPs found in *H. midae* is significantly lower than frequencies found in the Pacific oyster (*Crassostrea gigas*) (1 SNP/ 40bp - Curole and Hedgecock 2005), but similar to reports in other Halitiod species, including the Pacific abalone (*Haliotis discus hannai* - 1 SNP/ 100 bp) (Qi *et al*. 2009), Greenlip abalone (*Haliotis laevigata* - 1 SNP/ 27 bp),Blacklip abalone *(Haliotis rubra* - 1 SNP/ 29 bp)*,* Green abalone(*Haliotis fulgens* - 1 SNP/ 48 bp), *B*lackfoot paua (*Haliotis iris* - 1 SNP/ 45 bp) as well as Disk abalone (*Haliotis discus*) and Red abalone (*Haliotis rufescens* - 1 SNP/ 32 bp) (Kang *et al*. 2010). These as well as the current study suggest that a higher SNP density is present in the genomes of Haliotid species than in other mollusc species for example the Eastern oyster (*Crassostrea virginica* - 1 SNP/ 169 bp) (Quilang *et al*. 2007) and Pearl mussel (*Hyriopsis cumingii* - 1 SNP/ 345 bp)(Bai *et al*. 2009). A study conducted by Fahrenkrug *et al*. (2002) indicated that SNP frequencies are dependent on the locality of the primer pairs and that higher SNP frequencies are possibly associated with primers that are developed from introns and intergenic regions along with consideration to intron-exon boundaries and 3’-end

sequences (Wang *et al*. 2008; Kim *et al*. 2010). Although this has not been substantiated in Haliotid species, the general consensus is that stringent parameters and the knowledge of intron-exon boundaries are important considerations to bear in mind when designing primers(Qi *et al*. 2009; Kang *et al*. 2010). Of the 105 SNPs isolated, two were found to be tri-allelic markers. The presence, although theoretically viable, of a mutation site generating a second mutation is somewhat unexpected. Hodgkinson and Eyre-Walker (2010) hypothesized that three mutational mechanisms could explain the occurrence of such markers. The first explaination could be that the SNP induces a second mutation to compensate for the first as a means of corrective action. The second mechanism could be explained by the presence of hypermutable sites within the genome giving rise to tri-allelic markers. The last mechanism could cause the simultaneous generation of two alleles within a single individual; this could be a result of chemical or radiation exposure.

On a theoretical basis, all substitutions should be equally likely with an expected transition (ts) to transversion (tv) ratio of 1:2. This is based on the fact that twice as many transversions than transitions are likely to occur, but studies across various genomes have indicated this ratio to be biased (Vignal *et al*. 2002; Keller *et al*. 2007; Lynch 2007; Arnheim and Calabrese 2009). This could be attributed to sampling artefacts(Keller *et al*. 2007; Lynch 2007) or evidence of cytosine 5-methylation in the case of increased incidence of T/C transitions (SanMiguel *et al*. 1998; Arnheim and Calabrese 2009). The aforementioned has been illustrated in numerous aquaculture vertebrate and invertebrate species including Lake whitefish (*Coregonus clupeaformis* - 1.65) (Renaut *et al*. 2010), Atlantic salmon (*Salmo salar* - 1.37)(Hayes *et al*. 2007a) and Pacific white shrimp (*Litopenaeus vannamei* - 1.73) (Ciobanu *et al*. 2009). The ts:tv ratio observed in the current study (1.71) is slightly different than previous studies done on *H. midae* {Bester *et al*. 2008 (0.67); Rhode *et al*. 2008 (1);Rhode 2010 (1.5)}. It however conforms to the above-mentioned studies as well as with other mollusc species including the Pacific abalone (*Haliotis discus hannai*) (2.2) (Qi *et al*. 2009), the Eastern oyster (*Crassostrea virginica*) (1.3)(Quilang *et al*. 2007) and the Weathervane scallop (*Patinopecten caurinus*) (2.4) (Elfstrom *et al*. 2005). The ts:tv ratio is indicative of the level of genetic divergence that occurred, where a higher ratio corresponds to a lower level of diversity and vice versa(Riju and Arunachalam 2009).

In summary, research in *Haliotis midae* have mainly been directed at marker development in applications such as linkage mapping (Hepple 2010), QTL analysis (Slabbert 2010) and parentage assignment (Swart 2011), to name a few. To date, microsatellite markers were preferred for the above-mentioned applications due to their high level of polymorphism. Due to the increased popularity of SNPs, research was directed to the development of these markers in the current study, ultimately to be applied in conjunction with microsatellites in these various applications. Transcriptome characterisation with the aid of next generation sequencing technologies has proven to be adequate for the use of marker development in a non-model species such as perlemoen. Transcriptomic information will yield more valuable data since most of the resulting markers should be directly linked to coding regions, thus expressed genes within the genome. The analysis of such markers can be directed to target genes of interest such as those regulating immune response and environmental adaptation as well as gene expression studies (Van der Merwe *et al*. 2011). A prerequisite for the use of putative markers for any of the aforementioned applications is the characterisation of such markers within populations, which will be elaborated upon in the next chapter.

Chapter 3

Marker Characterisation

1 Introduction

The identification of polymorphisms from EST data (transcribed sequences) forms part of a multifaceted approach in using these markers to identify key traits within both commercial and natural populations. Single Nucleotide Polymorphisms became the preferred marker for various studies and applications, due to its abundance and distribution as well as the relative ease of developing these polymorphisms into genetic markers (Fan *et al*. 2003). Although discovery approaches have been streamlined to accumulate a large number of polymorphisms in a limited amount of time, these markers still require characterisation within individuals by means of genotyping (Hyten *et al*. 2010). This formulates the second facet of marker development, namely the characterisation of polymorphisms genotyped within populations to determine the validity of the markers. This requires validation methodologies that are highly accurate and simple, allowing for various degrees of multiplexing at both an efficient time scale and cost. Currently, various platforms exists permitting large-scale genotyping, but only a few are suitable for medium-throughput required for the bulk of the research on non-model species (Garvin *et al*. 2010). The development of SNPs in non-model organisms are however still limited, elevating the need for genotyping strategies that will ultimately adapt to these limitations and give accurate data.

The aim of this section of the study was to assess the validity of the SNP markers isolated from next generation sequencing by means of genotyping as well as to compare the informativeness of these markers when compared to markers that were isolated using alternative methods. If proved to be successful, the use of this workbench will allow for a marked increase in marker development and characterisation in *H. midae*.

2 Experimental Design

2.1 Selection of SNPs for Genotyping

Putative SNPs were considered for the Illumina GoldenGate genotyping assay with VeraCode technology on the BeadXpress platform. The assay requires at least 60 bp flanking regions at each variant locus. Tri-allelic SNPs and loci that did not adhere to this criterion were excluded

from the assay. A subset of loci containing SNPs was submitted to the Illumina Assay Design Tool (http://www.illumina.com/support/array/array_software/assay_design_tool.ilmn) to determine the functionality and primer designability score for each SNP locus. Scores of 0.5 - 1.0 are obligatory for a high quality assay. Only markers with a final score above 0.75 were considered for genotyping. Consequently, a 48-plex assay was designed; sufficient for genotyping 480 samples. The 48 SNPS comprised of 24 SNPs developed during the current project (*in vitro* SNPs), four from Bester *et al*., (2008), eight from Rhode, (2010) and 12 novel SNPs developed *in silico* using the SNP discovery application available on the CLC Workbench v4.5. This bioinformatics software package integrates various applications including sequence assembly, primer design and cloning, to name a few (http://www.clcbio.com; “CLC DNA Workbench: Features and Benefits”; [online 2011]). The *in silico* SNPs were selected as a test panel to determine the efficiency and performance of SNPs identified using the CLC workbench. The criteria set were a minimum coverage of a 100 (a correlation with the number of contigs previously used), minor allele frequency (MAF) larger than 10% and two allelic variants.

2.2 Selection of Samples for Genotyping

The samples selected for genotyping included the parents and offspring of four linkage mapping families (Table 3.1) from three commercial farms and individuals from three natural populations sampled from various locations along the South African coastline (Table 3.2). These locations included Saldanha Bay (West coast), Witsand (South coast) and Riet Point (East coast). DNA samples were precipitated according to the CTAB extraction method to increase the amount of DNA (250 ng) available for the assay (Black and Duteau 1997). Thereafter, DNA was diluted in TE Buffer (1 M Tris-HCl, 0.5 M EDTA, water at pH 8.0) or DNase free water (Promega) to a final concentration of 50 ng/µl and placed in a 96-well skirted plate. Two positive controls (individuals with previously determined genotypes (Bester *et al*. 2008) and one negative control were added to each plate to monitor genotyping efficiency and possible contamination.

Table 3.1.Genotyping individuals from commercial populations.

Sample origin	Number of individuals	
	Parents	**Offspring**
Family DS_1	M342 and F617	103
Family DS_2	M456 and F462	94
Family DS_5	W06 and W88	90
Family DS_6	D06 and V06	94

Table 3.2.Genotyping individuals from natural populations.

Natural Population	Number of individuals
Saldanha Bay	23
Witsand	26
Riet Point	26

2.3 SNP Genotyping

The Illumina GoldenGate genotyping assay with the VeraCode technology on the BeadXpress platform is a multifaceted approach that is divided into multiple sections. The first being the GoldenGate genotyping procedure depicted in the pre-PCR section. The post-PCR steps involve the incorporation of the VeraCode technology into the genotyping procedure. This is then transferred to the BeadXpress Reader; after which assessment is concluded with the GenomeStudio™ Software.

2.3.1 Pre-PCR Procedure

The pre-PCR procedure includes seven steps to prepare the samples for genotyping. The first step is the creation of the Single Use DNA (SUD) plate through the biotinylation of the DNA by means of a chemical heat-activation procedure. This process involves the addition of 5 µl of biotinylation reagent (MS1) to each individual DNA sample to activate it. A volume of 5 µl (approximately 250 ng DNA) is then transferred to the SUD plate. This is followed by the addition of 5 µl of precipitation reagent (PS1) and 15 µl of 2-propanol to every well of the SUD

plate for precipitation. Precipitation involves numerous vortex and centrifugation steps according to manufacturer's specifications (Lin *et al*. 2009). After precipitation, a pellet is observed, which is then resuspended in 10 µl of resuspension reagent (RS1).

The following process combines the biotinylated gDNA with the query oligonucleotides and the streptavidin-coated paramagnetic particles for the Allele Specific Extension (ASE) step. A volume of 30 µl of oligonucleotide annealing reagent (OB1) and 10 µl of oligonucleotide pool (OPA) reagent is added to a new 96-well plate, labelled ASE. This is followed by the addition of 10 µl of the template DNA to the ASE plate. The beads are resuspended by means of vortex and incubated on a heat source for up to sixteen hours at 30°C. After incubation of the ASE plate, the beads are captured on a raised-bar magnetic plate and several wash steps are performed with 50 µl of each of the wash buffers (AM1 and UB1). Both the allele specific and ligation reactions are then completed simultaneously by adding 37 µl of Extension/Ligation (MEL) mixture to the ASE plate. The ASE plate is subsequently incubated for 15 min at 45°C. The final step in the pre-PCR protocol is the preparation of the PCR plate. This commences with the addition of 64 µl of Titanium *Taq* DNA polymerase and 50 µl of Uracil DNA glycosylase (UDG) to the PCR master mix (MMP). A volume of 30 µl of MMP is aliquoted into the PCR plate and stored away from any light source due to the light sensitivity of the universal primers. The beads from the ASE plate are then captured after incubation and washed with 50 µl of UB1. The supernatant is discarded before and after the wash procedures. The ASE plate is removed from the magnet and the beads resuspended in 35 µl of Inoc PCR (IP1) reagent. After recapturing of the beads, 30 µl of the supernatant is placed in each corresponding well of the PCR plate. The PCR plate is sealed with PCR plate sealing film and placed in a thermal cycler for the amplification of the template DNA with the fluorescently labelled primers. Thermal cycling times and temperatures are indicated in Table 3.3.

Table 3.3.PCR protocol for Illumina GoldenGate Assay.

	Temperature	Time
	37°C	10 min
	95°C	3 min
34 x	95°C	35 sec
	56°C	35 sec
	72°C	2 min
	72°C	10 min
	4°C	5 min

2.3.2 Post-PCR Procedure

Following PCR, single stranded DNA is generated for hybridisation to the VeraCode beads through removal of the non-fluorescent strands. A total of 20 µl of resuspended paramagnetic bead mixture (MPB) is added to the PCR plate, after which the mixture is transferred to a filter plate. The filter plate is incubated at room temperature for one hour. After incubation, a waste plate is placed underneath the filter plate, both plates are then centrifuged at 1000 xg for 5 min at 25°C, after which 50 µl of UB2 is added to the filter plate. The plates are centrifuged again under the same conditions. During centrifugation 30 µl of hybridisation reagent (MH2) is added to a new 96-well plate, referred to as the INT plate, which replaces the waste plate. Following centrifugation, 30 µl of 0.1 M NaOH is added to the filter plate, which is then centrifuged at the conditions stipulated above. The INT plate now contains the single stranded DNA ready for hybridisation. A solution of MH2 is neutralised by adding 3 ml of MH2 to 3 ml of 0.1 M NaOH, of which 50 µl is pipetted into each well of the INT plate. A 100 µl was taken from each well from the INT plate and added to the corresponding well of the VeraCode bead plate (VBP). The VBP was then incubated for three hours at 45°C. Incubation is followed by washing the VBP with 200 µl of VeraCode bead wash buffer (VW1). The VBP is then transferred to the BeadXpress Reader for scanning of the plate.

2.4 Genotyping Analysis

The GenomeStudio™ Genotyping Module v1.0 software was employed to analyse data generated by the Illumina BeadXpress platform. This software allows for easy assessment of raw data generated by Illumina platform as well as the inference of genotypes and allele frequencies per locus. Assessment of data quality was done using parameters set out by Illumina of a No-Call (GenCall) threshold and GenTrain (clustering algorithm) of 0.25 and an individual call rate of 0.85. Individuals and SNPs for which genotypes could not be inferred were omitted before performing clustering analysis. Clustering was completed and genoplots were generated for each SNP separately. Subsequently, SNPs that illustrated ambiguous clustering were removed prior to further analysis.

Nine SNPs were randomly selected and sequenced in a total of 30 individuals as an additional validation step. Sequencing was performed with the ABI PRISM® BigDye Terminator v3.1 Cycle Sequencing kit in the forward direction using the 3100 Genetic DNA Analyser (Applied Biosystems). Genotypes acquired from sequencing were compared to those generated on the GenomeStudio Genotyping Module.

3 Results

3.1 SNP Performance

A total of 48 markers (Table 3.4) satisfied genotyping prerequisites and were validated with the Illumina GoldenGate genotyping assay with the VeraCode technology on the BeadXpress platform. These consisted of SNPs developed by various means from the *H. midae* genome (Table 3.4). SNPs validated for this assay were found to have a higher functionality score (> 0.80) than the recommended score (> 0.6) when evaluated by the Illumina Assay Design Tool. Although all SNPs selected for genotyping had a final functionality score higher than 0.75, 14.6% (7) of the SNPs failed to cluster and were considered as genotyping failures. Previous studies found that the inclusion of SNPs with functionality score lower than 0.6 reduced the overall success rate of such markers significantly (Pavy *et al*. 2008; Lepoittevin

et al. 2010). Although this is the case, earlier literature (Wang *et al*. 2008) has also found that the functionality score is not the sole determinant of the genotyping success rate.

Table 3.4. SNP loci genotyped with the Illumina GoldenGate genotyping assay with the VeraCode technology on the BeadXpress platform.

Isolation method	SNP name	Variant	SNP position
Bester *et al*. 2008 Targeted EST approach	3B4_2	T/C	561
	3B4_7	A/T	492
	3D10_1	A/G	122
	2H92	A/T	177
Rhode *et al*. 2008 Microsatellite flanking regions	HdSNPc148_820T_C	T/C	-
	HdSNPc106_688C_T	C/T	-
	HmSNPc4_815C_T*	C/T	-
	HaSNPdw500_207C_T*	C/T	-
	HmLCS5M193T_A*	T/A	196
	HmLCS5M479C_T*	C/T	484
	HmLCS55T318G_T	G/T	322
	HmRS36T262T_C	T/C	306
in vitro SNPs current study	SNP101_113	A/C	113
	SNP101_201	C/G	201
	SNP146.2_132	A/G	132
	SNP146.3_123	T/G	123
	SNP149.2_165	A/G	165
	SNP149.4_75	A/G	75
	SNP149.4_341	T/C	341
	SNP210_266	T/G	266
	SNP214_86	T/C	86
	SNP214_434	T/C	434
	SNP17550.3_221	A/T	221
	SNP17550.3_555	A/T	555
	SNP1949_235	A/C	235
	SNP4691_183	A/G	183
	SNP1718_109	A/T	109
	SNP149.1_106	A/C	106
	SNP149.1_374	C/G	374
	SNP449.2_110	A/G	110

	SNP449.2_443	T/C	443
	SNP1833_160*	A/G	160
	SNP1834_464	A/G	464
	SNP1834_76	A/G	76
	SNP17550.1_463	A/G	463
	SNP342.2_537	T/C	537
in silico SNPs current study	SNP48_322*	T/G	322
	SNP67_164	A/G	164
	SNP140_2421	T/C	2421
	SNP229_2772	T/C	2772
	SNP300_1828	A/G	1828
	SNP972_1055	T/C	1055
	SNP1001_388	T/C	388
	SNP2091_264	A/C	264
	SNP3129_923	A/G	923
	SNP5837_204*	T/C	204
	SNP13865_165	T/C	165
	SNP20648_3041	A/G	3041

* SNPs that failed to cluster correctly

- positions not made available by previous author

3.2 Genotyping Performance

GenomeStudio Genotyping Module made allele calls for all loci and thus generated genotypes for individuals for each locus in the form of cluster plots referred to as genoplots. Through the application of fluorescent signals, a genoplot maps a specific individual to a specific allele for each locus. A GenCall score is subsequently generated; representing the clustering of each individual SNP and thus the reliability of each genotyping score. Scores are assigned between 0 and 1; a score closer to 1.0 indicates that the genotype inferred is reliable and can be visually verified with the genoplots (Figure 3.1a). Individuals with lower GenCall scores are less reliable and are distinguished by a clear separation from the centre of the cluster. Failed SNPs could not be assigned to a genotypic cluster due to low GenCall and GenTrain (< 0.25) scores; depicted in Figure 3.1b. Sanger sequencing of randomly selected SNPs was used to confirm the accuracy and reliability of the calls made by the

GenomeStudio Genotyping Module software. Due to technical difficulties, the majority (87.6%) of the commercial population, DS_2, could not be genotyped and thus genotyping results were considered negligible for this population. Apart from identifying failed SNPs, monomorphic (Figure 3.1c) and polymorphic loci (Figure 3.1a) can also be distinguished through variation in the clustering. Only SNPs consisting of a GenTrain score of ≥ 0.80 and a call rate above 80%, as well as a MAF of greater than 0.01 were considered as successfully genotyped SNPs. An average MAF of 0.151 was obtained for all SNPs genotyped, with the majority of the markers having a MAF between 0.001-0.050, as depicted in Figure 3.2 (detailed estimates in Table S2: Appendix C).

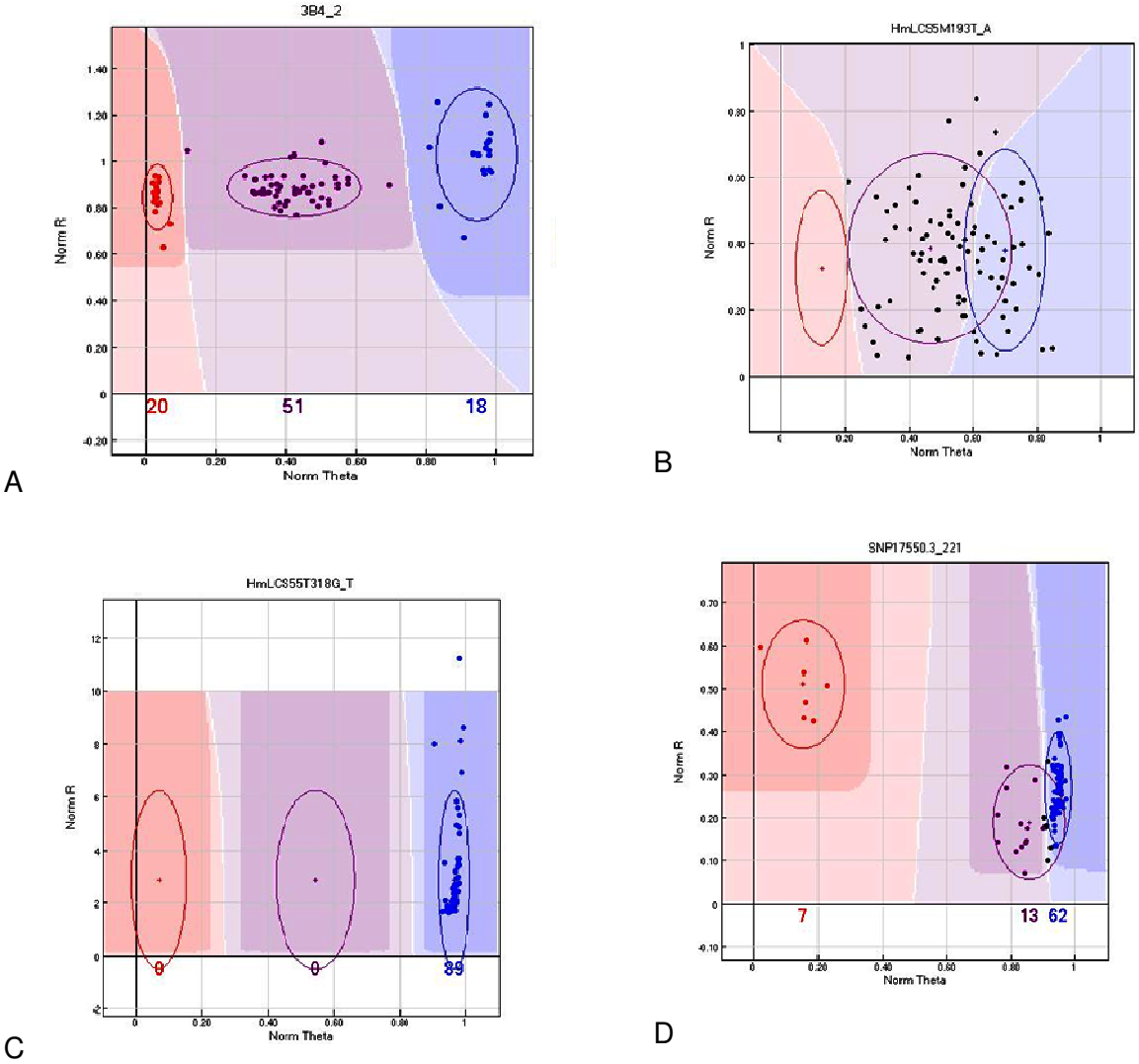

Figure 3.1.GenomeStudio Genotyping Module v1.0 genoplots.The genoplots are represented by the Norm Theta (X-axis) and the Norm R (Y-axis). Norm R is the normalized signal intensity and norm Theta corresponds to the allele frequency (http://www.illumina.com/software/genomestudio_software.ilmn [accessed July 2011]). Shaded areas represent the call zones for the different genotypes (AA, AB and BB) and is defined by the GenCall score. Numbers below each shaded area indicate the number of individuals (dots) in that genotypic group. Black dots represent individuals that could not be genotyped. A: a representation of a successfully genotyped (polymorphic) SNP; B: Failed SNP; C: monomorphic SNP and D: cluster compression.

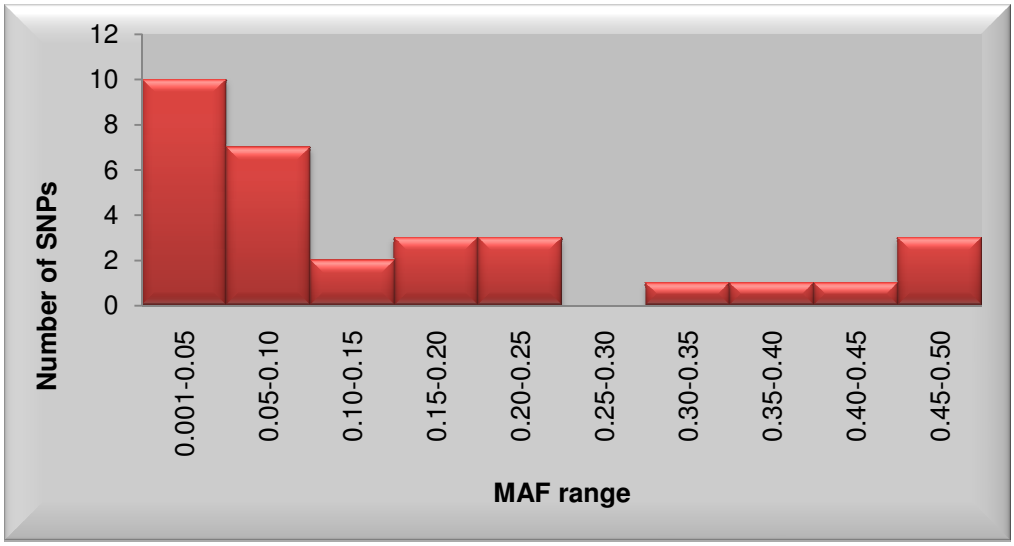

Figure 3.2. A graphical representation of the range of Minor Allele Frequency obtained with the 31 clustered SNPs. All loci that produced a MAF were categorised in the MAF ranges across all populations.

An overall (global) genotyping success rate of 85.4% (41 SNPs) was obtained for the assay; calculated by dividing the number of successfully genotyped loci (polymorphic and monomorphic) by the total number of SNPs. Of all the successfully genotyped loci, 75.6% (31) were found to be polymorphic and 24.4% (10) were monomorphic (Table 3.5).The conversion rate was determined to be 64.5%, as defined by Fan *et al*. (2003); the number of polymorphic SNPs divided by the total number of SNPs. The values obtained in this study showed that the assay is sufficient when typing a limited number of markers in a small number of individuals. Of the 24.4% (10) monomorphic SNPs, 20% (2) were found to be due to a cluster compression where the normalized intensity of genotypes are too clustered on one side of the axis (Figure 3.1d).

Table 3.5.Genotyping success rate.

Categories	Number of SNPs	Av. Quality Score
SNPs genotyped	48	0.60
Successful genotypes	41	0.66
Polymorphic SNPs	31	0.64
Monomorphic SNPs	10	0.70
Failed SNPs	7	0.20

The SNPs genotyped in this study were obtained through various isolation methods. These included SNPs isolated from microsatellite flanking regions (17%), SNPs derived from a targeted EST approach (8%), as well as *in vitro* (50%) and *in silico* (25%) EST-derived SNPs from the *H. midae* transcriptome generated from next generation sequencing. It was found that the majority (57%) of SNPs that failed to genotype were identified from microsatellite flanking regions. A comparison between SNPs isolated from the next generation sequencing revealed that the *in vitro* SNPs had a higher genotyping success rate (95.8%) than the *in silico* SNPs (83.3%). Although, the *in silico* SNPshad a lower genotyping success rate, considering the time and efficiency with which these markers were isolated compared to *in vitro* SNPs, one could argue that *in silico* SNPs would still be a more viable option for SNP isolation in the future.

4 Discussion

Previous studies related to *Haliotis midae* describe the development of over 200 polymorphic microsatellite loci (Bester *et al*. 2004; Slabbert *et al*. 2008, 2010; Hepple 2010; Rhode 2010; Slabbert 2010). Although these markers demonstrated to be highly informative, focus has shifted towards the use of SNPs as markers due to their high coverage, non-homoplasy and easy high-throughput characterisation.

4.1 Genotyping Performance

The Illumina GoldenGate genotyping assay proved to be adequate for the genotyping of SNPs within a non-model species such as *H. midae*. Although the GoldenGate assay has been applied to the genotyping of human polymorphisms (Altschuler *et al*. 2005) and the Bovine HapMap consortium (Sonstegard 2007), little evidence supports its use in non-model species especially in aquaculture, thus genotyping success rates could only be compared to the limited existing data on other non-model species. The assay consists of various parameters that contribute to its success as a genotyping tool. These include, but are not limited to, the preliminary functionality score, GenTrain score and finally the GenCall score. The functionality score measures the SNP adequacy for the assay. This considers repetitive sequences present up to 200 bp both upstream and downstream of the SNP, the general sequence uniqueness and sequence redundancy against the relevant species database (Shen *et al*. 2005), the latter of which is not applicable to the current study. All SNPs in this study were chosen with a functionality score above 0.80 that would ensure a high genotyping success rate as well as to compensate for the lack of sequence data available for *H. midae*. The high functionality score was an acceptable indicator of genotyping success considering the considerable lack of genome data available for *H. midae.* Following genotyping, GenTrain and GenCall scores were assigned to each SNP. The GenTrain scores correspond to the degree of separation between heterozygote and homozygote clusters for a particular SNP, as well as the ease with which individuals are categorised within each cluster allowing for accurate high signal-to-noise ratio (Fan *et al*. 2003; http://www.illumina.com, “Illumina Technology spotlight: Illumina GenCall data analysis software”, [online 2005]; http://www.illumina.com [accessed: June 2011]). This study utilised the recommended GenTrain cut-off value of < 0.25, which indicates a low call rate, and therefore SNPs portraying a value lower than this were excluded from further analysis. The number of SNPs that failed was found to correspond with low GenTrain scores. The final parameter accounted for was the GenCall score, which considers the genotypes called for a particular SNP (Fan *et al*. 2003). A successfully genotyped SNP is indicated by a GenCall rate higher than 80% and a GenTrain score in excess of 0.80.

The success rate of any genotyping method is reflected in what is referred to as the SNP conversion rate. This considers the number of polymorphic markers excluding all monomorphic markers that were successfully genotyped. The conversion rate found for this study (65.4%) is in accordance with conversion rates found for the white spruce (*Picea glauca*) (69.2%, 77.1% for the black spruce (*Picea mariana*)), maritime pine (*Pinus pinaster* - 51%) (Lepoittevin *et al*. 2010) and catfish (*Ictalurus* spp. - 59.8%) (Wang *et al*. 2008), which made use of EST data for the isolation of SNP markers (Pavy *et al*. 2008).The application of the GoldenGate genotyping assay in abalone yielded much lower conversion rates compared to that of the manufacturers' predictions (93%), however it must be noted that these predictions were from species of which the genomes were completely sequenced such as the human genome (Montpetit *et al*. 2006). The various factors attributing for the lower conversion rates may be either the extreme cut-off values for the GenTrain scores, or the lack of knowledge regarding the complexity of the abalone genome. Another possible explanation for the low conversion rates along with the low GenTrain scores are paralogous SNPs. These cause background signal to form a cluster compression, resulting in poor signal-to-noise ratios. Although tedious, this can be circumvented by manually editing SNPs that fail to cluster properly (Shen *et al*. 2005), which would not only allow the utilisation of these SNPs in data analysis, but also improve the overall (global) genotyping success rate.

The global success rate considers all markers (monomorphic and polymorphic) that were successfully typed within the sample group (Lepoittevin *et al*. 2010). The global success rate for this study was found to be 85.4%, corresponding with studies done on the common bean (*Phaseolus vulgaris* - 86%) (Hyten *et al*. 2010), but higher than both catfish (*Ictalurus* spp. 69%) (Wang *et al*. 2008) and maritime pine (*Pinus pinaster* - 66.9%)(Lepoittevin *et al*. 2010). SNPs that failed to cluster and thus failed to genotype could be attributed to the deficient knowledge of the abalone genome, inhibiting genotyping assay design efficiency. This, along with the presence of repetitive elements within the SNP flanking regions, could hinder primer attachment during genotyping (Fan *et al*. 2003; Shen *et al*. 2005; Pavy *et al*. 2008). Another reason for failed genotypes is undetected SNPs that manifest in the priming site due to the use of insufficient sample populations for genotyping, especially when the organisms exhibit high levels of genetic diversity (Eckert *et al*. 2009).

The number of monomorphic SNPs found in this study was significantly lower (24%) than other aquaculture species (Wang *et al*. 2008). Monomorphic SNPs that showed distinct cluster compressions, such as a homozygote subgroup, were also considered insignificant and were subsequently excluded from the data analysis. To date, no literature has been published evaluating whether or not the genotyping calling algorithms can detect cluster compression, or if failed reactions are caused by monomorphic SNPs. Such cluster compressions could be caused by paralogous genes, which match one allele resulting in a compression of both the homozygous cluster by increasing the signal of the allele for both the other genotypes (Lepoittevin *et al*. 2010). Some markers were found to be heterozygous in all individuals typed, which could represent paralogous duplications; a phenomenon already discovered in the sperm lysin gene in *H. tuberculata coccinea* (Clark *et al*. 2007). Unfortunately, due to the limited knowledge of the perlemoen genome, duplication occurrences need to be explored further. A possible reason for these complete heterozygous markers could be cross-amplification of the allele-specific primers (Zhang and Guo 2010). Although the GoldenGate genotyping method has been used for genotyping SNPs in human (Fan *et al*. 2003) and other model species, extensive research needs to be completed to ensure its validity in non-model aquaculture species.

4.2 SNP Performance

Due to the various means in which SNPs utilised in this study were obtained, it seemed necessary to evaluate each in terms of performance. The SNPs were isolated in four different ways: directly from a cDNA library via a targeted EST approach (8%), from the flanking regions of microsatellites (17%), from EST contigs and re-sequencing (referred to as *in vitro* - 50%) and using the CLC workbench version v4.5, as a means of *in silico* SNP isolation (25%). The latter two methods utilised transcriptome data generated from next generation sequencing. Considering that all the techniques, with the exception of the CLC workbench, represented an EST-derived means of isolation, a comparison was made between *in silico* and EST-derived isolation methods.

Evaluation of the various techniques revealed that more (71%) EST-derivedSNPs failed to genotype, of which the majority (57%) were isolated from the flanking regions of existing microsatellite markers (Rhode *et al*. 2008). This could be explained by the fact that microsatellite flanking regions are hyper-variable (Rhode *et al*. 2008), and that secondary SNPs occurring closely to the genotyped SNP were missed, therefore hindering primer binding. Wang *et al*. (2008) found that the avoidance of mutational hotspots were crucial for a high genotyping success rate.

Quality assessment parameters for genotyping *in vitro* SNPs do not exist; therefore, care must be exercised during SNP discovery. The most prominent issue is sequencing errors that pronounce themselves as variations (pseudo-SNPs) within the assembled contigs, low quality SNP flanking regions or, more importantly, the presence of intron-exon boundaries adjacent to the SNP (Wang *et al*. 2008). The latter is a major problem when working with non-model species for which no complete genome exists. These boundaries prohibit primers to bind effectively to the SNPs' flanking regions, which lead to genotyping failure. Another problem is highly similar paralogous sequence variants (PVS), a current issue due to the short reads produced by next generation sequencing (Liu *et al*. 2011). These variants are presented as SNPs after genotyping. These sequence variants are identified by heterozygous excess present for that SNP within a given population and would thus not segregate according to Hardy-Weinberg expectations (Gut and Lathrop 2004). Another important factor to consider is

the number of individuals used for SNP validation. This could limit the number of markers found due to too limited coverage, even when these individuals are randomly chosen.This approach both prohibits the identification of rare SNPs as well as increases the likelihood of calling false SNPs. The high (70%) number of non-informative SNPs genotyped from *in vitro* contig assembly in the current study supports the importance of implementing strict quality assessment parameters.

The success of *in silico* SNP detection methods depends on the species in which it is tested as well as the quality of the EST and/or sequencing data (Wang *et al*. 2008; Eckert *et al*. 2009; Lepoittevin *et al*. 2010). Additionally, care must be taken to ensure that the *in silico* data include a representative population sample in order to avoid SNPs not being detected, or a pronounced increase in the detection of monomorphic SNPs. All the above-mentioned parameters affect the global success rate as well as the conversion rates when using *in silico* SNPs. Another important factor to consider is contig depth, which is highlighted in several studies (Wang *et al*. 2008; Lepoittevin *et al*. 2010). Even though 28% of the *in silico* markers in the current study failed to genotype, it is still a time and cost effective means of isolating SNPs and proves to be adequate for the use in a non-model species, such as *Haliotis midae*.

Chapter 4

Marker Application

1 Introduction

Currently, 18 commercial abalone farms exist in South Africa (Britz *et al*. 2009). These farms could alleviate strain from the diminishing natural *H. midae* populations that reside along the coast. If managed correctly, the farms pose the potential to act as genetic reservoirs through which genetic variation can be maintained and which could also form the bases of genetic improvement programmes. It is therefore imperative to determine the genetic diversity within these commercial populations in terms of heterozygosity levels, which could indicate whether key processes such as genetic drift, inbreeding or genetic bottlenecks are present within these cultured populations (Roodt-Wilding and Slabbert 2006). Knowledge of these processes will provide a better basis for breeding and enhancement programmes (Slabbert *et al*. 2009). The commercial farms were established with the recruitment of animals from the natural populations. Therefore, determining genetic variability and population differentiation of natural populations is also important. This will aid in the management and conservation of populations in terms of restocking and recovery programmes.

Only a few studies have investigated the genetic diversity and population differentiation within and amongst *H. midae* populations. The first assessment of diversity between commercial and natural populations was based on three microsatellite markers en mtDNA (Evans *et al*. 2004a, b). These studies showed a marked loss of diversity from the natural to the commercial populations. However, these findings were contradicted by Slabbert *et al*. (2009) who reported no significant loss of diversity to the commercial environment based on six microsatellite markers. A more recent study done by Bester-van der Merwe *et al*. (2011) further investigated population structure of nine natural populations along the South African coast and found that weak genetic differentiation was present between the populations from the West and East coast mainly attributable to oceanographic barriers.

The main objective of this chapter was to test the utility of the successfully genotyped SNPs in determining genetic diversity and population differentiation in *H. midae*. Therefore, 31 gene-linked and genomic SNPs were collectively analysed in three natural and three cultured populations, respectively. Genetic diversity was assessed using basic descriptive statistics, whereas population differentiation was determined with the use of summary statistics and multivariate analysis.

2 Experimental Design

The SNPs that were successfully genotyped (31) with the GoldenGate platform were selected for statistical analysis to determine the utility of these SNPs in genetic diversity and population analysis. The SNPs were evaluated in three commercial and three natural populations. The samples selected for genotyping were selected from three linkage mapping families from two commercial farms and individuals from three natural populations from Saldanha Bay (West coast), Witsand (South coast) and Riet Point (East coast)(Table 4.1).

Table 4.1.Sample cohort for SNP genotyping.

Sample origin	Nr of samples
Family DS_1	105
Family DS_5	92
Family DS_6	96
Saldanha Bay	23
Witsand	26
Riet Point	26

2.1 Statistical Analysis

2.1.1 Genetic Diversity

For genetic diversity analysis the expected (H_e) and observed (H_o) heterozygosity were determined per locus as well as per population using the Genetic Data Analysis software (GDA 1.1) (Lewis and Zaykin 2001). The inbreeding coefficient (*f*) was determined using GDA. The Exact test, available in GENEPOP v4.0 (Rousset 2008), was performed to determine departure from Hardy-Weinberg Equilibrium (HWE) for each locus in the various populations as well as per population. The Markov chain randomisation method was employed to calculate significance values (P-values) (Guo and Thompson 1992). Linkage disequilibrium (LD) between pairs of SNP loci was determined by Fisher's exact test (1000 permutations) also available in GENEPOP.

2.1.2 Population Differentiation

Population differentiation was determined using the pairwise F_{ST} summary statistic from Weir and Cockerham (θ), (1984) implemented in GENETIX v4.03. Significance was tested with 1000 permutations and the Bonferroni correction method was used to adjust the significance levels for multiple tests (Rice 1989). A multi-dimensional graph was created by submitting allele frequencies to factorial correspondence analysis (FCA) also available in GENETIX. This provided a three-dimensional view of the distribution of genetic variation between individuals of the six genotyped populations.

3 Results

3.1 Genetic Diversity

Genetic diversity indices were calculated for six populations, three of which were commercial and three natural populations distributed along the coast of South Africa (Table 4.1). The observed (H_o) and expected (H_e) heterozygosities across all populations ranged from 0.156 to 0.268 and 0.154 to 0.195, respectively with the averages determined to be 0.20 and 0.168, respectively (Table 4.2). The inbreeding coefficients obtained for all populations indicate that no inbreeding is currently present in any of the commercial or natural populations. The lack of inbreeding (average f = -0.349) in the commercial populations could be partly attributed to the use of representative broodstock i.e. broodstock that were established from various natural populations and thus share very little or no common distribution of alleles. The natural populations had a similar average H_o and H_e (0.158) and an average (across all natural populations) inbreeding coefficient of 3.0 x 10^{-4}. The assessment of H_e and H_o across all loci ranged from 0.003 to 0.497 and 0.003 to 0.788, respectively, with the average H_e and H_o across all loci of 0.248 and 0.225, respectively (Table 4.3). The average observed heterozygosity for all loci is represented in Figure 4.1. The average MAF was relatively low at 15% (0.151) (Chapter 3). In the current study only two of the polymorphic SNPs (*HdSNPc106_688C_T* and *SNP1834_464*) were found to have a significantly ($P < 0.05$) high (≥ 0.7) level of H_o in all populations typed. The high heterozygosity depicted by these SNPs could be an indication of paralogous (duplicated) alleles rather than varying alleles of the same gene (Hubert *et al*. 2010).

Table 4.2. Detailed estimates of the heterozygosities, inbreeding coefficient (*f*) and HWE probability values for all the populations analysed.

Population	Heterozygosity		Inbreeding coefficient	P-value (HWE)
	Observed	Expected		
DS_1	0.268	0.195	-0.375	< 0.01
DS_5	0.214	0.166	-0.290	< 0.01
DS_6	0.237	0.172	-0.382	< 0.01
Average	**0.240**	**0.178**	**-0.349**	**-**
Riet Point	0.156	0.154	-0.016	0.165
Saldanha	0.166	0.157	-0.061	0.234
Witsand	0.152	0.164	0.078	0.062
Average	**0.158**	**0.158**	**3×10^{-4}**	**-**
Overall Average	**0.199**	**0.168**	**-0.188**	-

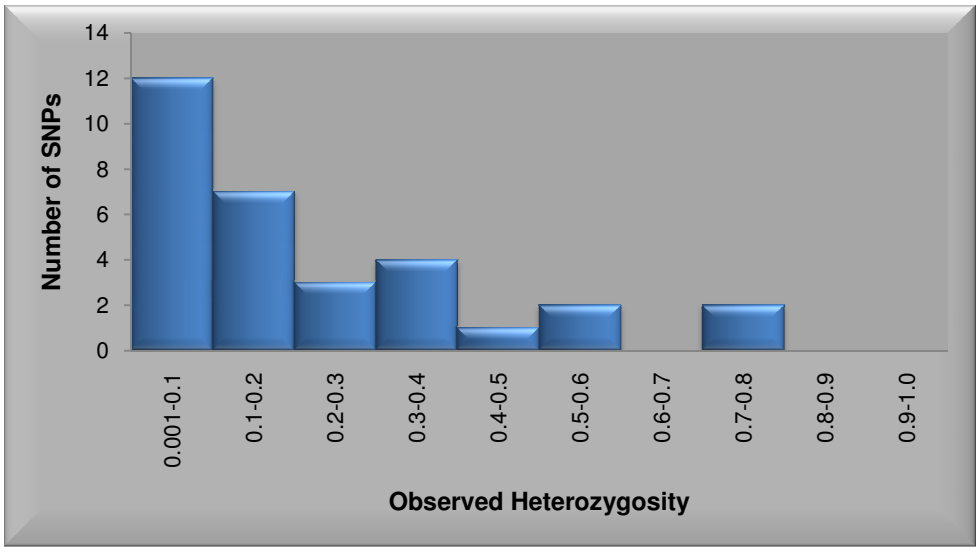

Figure 4.1. The observed heterozygosity for all the polymorphic loci across all populations. SNPs were categorised into the various ranges of observed heterozygosity.

Table 4.3. Detailed estimates of the heterozygosities, inbreeding coefficient (*f*) and HWE probability values for all the loci analysed.

Population	Heterozygosity		Inbreeding coefficient	Probability (HWE)
	Observed	Expected		
HdSNPc106_688C_T	0.721	0.462	-0.563	< 0.01
HdSNPc148_820T_C	0.183	0.172	-0.066	0.000
HmRS36T262T_C[#]	0.009	0.331	0.972	0.002
SNP4691_183	0.138	0.353	0.610	0.025
SNP149.4_75	0.003	0.003	0.000	-
SNP101_113[#]	0.069	0.073	0.049	0.530
SNP1949_235	0.309	0.266	-0.162	0.000

SNP214_86	0.041	0.040	-0.019	0.875
SNP149.1_374	0.025	0.031	0.189	0.799
SNP149.2_165#	0.483	0.478	-0.009	0.001
SNP101_201	0.091	0.126	0.278	0.786
SNP210_266	0.044	0.043	-0.021	1.000
SNP449.2_110#	0.320	0.497	0.357	0.000
SNP342.2_537	0.098	0.111	0.116	0.793
SNP17550.3_221	0.013	0.013	-0.005	-
SNP1834_464	0.788	0.489	-0.614	< 0.01
SNP17550.3_555	0.119	0.112	-0.062	0.081
SNP146.2_132	0.285	0.245	-0.165	0.011
SNP1834_76	0.013	0.012	-0.005	-
3D10_1	0.281	0.472	0.406	0.001
3B4_2#	0.592	0.465	-0.274	< 0.01
3B4_7#	0.537	0.473	-0.134	< 0.01
SNP972_1055#	0.082	0.394	0.794	< 0.01
SNP1001_388	0.280	0.250	-0.119	0.000
SNP20648_3041	0.164	0.166	0.015	0.063
SNP300_1828#	0.195	0.181	-0.075	0.058
SNP2091_264	0.314	0.486	0.354	0.001
SNP67_164	0.182	0.176	-0.033	0.181
SNP3129_923	0.318	0.462	0.312	< 0.01
SNP229_2772	0.099	0.128	0.228	0.880
SNP140_2421#	0.178	0.162	-0.096	0.310
Average	**0.225**	**0.248**	**0.092**	**-**

- SNPs that are in linkage disequilibrium.– No P-value was assigned.

Analysis in GENEPOP v4.0 showed that 16 (~52%) loci deviated significantly ($P < 0.05$) from Hardy-Weinberg equilibrium. The average inbreeding coefficient (*f*) across all populations was determined to be -0.188 (Table 4.2), indicating an excess of heterozygotes. All three commercial populations deviated significantly ($P < 0.05$) from Hardy-Weinberg equilibrium after Bonferroni corrections, whereas the natural populations were in accordance with Hardy-Weinberg equilibrium. Nine of the loci showed significant linkage disequilibrium ($P < 0.05$), diminishing the total number of independent polymorphic loci and information obtained in this study.

Considering the various means of SNP isolation utilised in this study, it was found that a larger number (57%) of EST-derived SNPs deviated from Hardy-Weinberg proportions than *in silico* SNPs (44%) and both categories showed significant deviation in all commercial populations.

3.2 Population Differentiation

Population heterozygosity levels were calculated to determine genetic variation for all the populations. Pairwise $F_{ST}(\theta)$ values ranged from -0.014 to 0.511, with significance (F_{ST} = 0.343, P < 0.05) over most of the populations (Table 4.4). Results indicated limited genetic differentiation between the three natural populations. The negative F_{ST} value obtained between Riet Point and Witsand could be attributed to the small sample sizes (Long 1986) that were taken in these areas or a higher likelihood that the differences exist between two random individuals of the same population rather than two random individuals of various populations. The significant population differentiation observed between Saldanha Bay and Riet Point is expected considering the geographical distance. Testing for isolation by distance and performing AMOVA (Analysis of Molecular Variance) could further substantiate the population differentiation found between these populations. The relatively high values (F_{ST} = 0.447- 0.531) obtained between the natural and commercial populations on the other hand indicated considerable genetic differentiation between natural and commercial abalone.

Table 4.4.Pairwise F_{ST} (θ) of the six populations of *H. midae*.

	DS_5	DS_6	RP	SD	WS
DS_1	0.213*	0.135*	0.503*	0.511*	0.481*
DS_5		0.232*	0.468*	0.483*	0.447*
DS_6			0.507*	0.531*	0.489*
RP				0.015#	-0.014
SD					0.016

* Significance at P < 0.01. # Significance at P < 0.05

RP – Riet Point, SD – Saldanha Bay, WS – Witsand.

Factorial correspondence analysis was performed to obtain a three-dimensional observation of the genetic relationship between the six different populations. The first factor accounted for 69.33% of the genetic variation and the second factor for 10.83% (Figure 4.2). A clear separation is evident between the natural populations (indicated on the left side of the plot) and the commercial populations (indicated on the right). Comparison of the individual genotypes indicated substantial overlap between individuals from natural populations while less overlapping was visible for individuals from different commercial populations. In conclusion, the genetic differentiation between the natural

populations was negligible indicating very low levels of population structure or even panmixia for the natural populations.

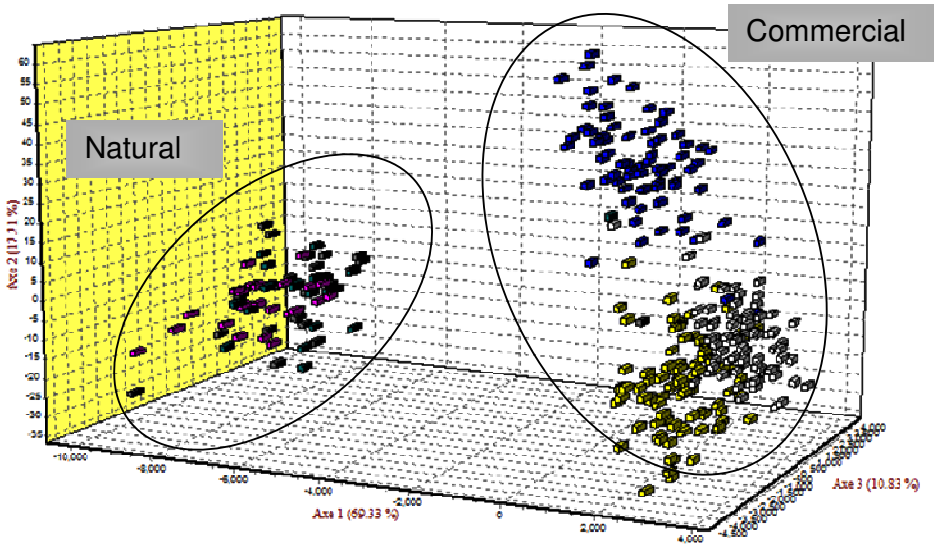

Figure 4.2. Factorial correspondence analysis depicted in a scatter plot of all the individuals in the various populations. Each colour represents the various populations tested; DS_1 (Yellow), DS_5 (Blue), DS_6 (White), Riet Point (Grey), Saldanha Bay (Pink) and Witsand (Green).

4 Discussion

Previously, applications such as diversity and population structure assessment as well as parentage assignment and linkage mapping, to name a few, were addressed with the use of microsatellite markers in *H. midae* due to their neutrality (Bester *et al*. 2004; Slabbert *et al*. 2008, 2009; Rhode 2010; Slabbert 2010; Slabbert *et al*. 2010, Hepple 2010). However, with the marked increase in popularity of SNPs it would be advantageous to consider their use due to easy development, reduced homoplasy and high-throughput genotyping. To date, only a few SNPs have been reported in *H. midae* (Bester *et al*. 2008; Rhode *et al*. 2008; Rhode 2010), of which only 12 SNPs were used in genetic diversity and population structure inference for the species in natural populations (Bester-van der Merwe *et al*. 2011). Therefore, the current study is the first reported usage of SNPs for genetic variation determination in both natural and commercial populations in *H. midae*.

4.1 Genetic Diversity

Analyses indicated that 75.6% of the SNP loci were polymorphic and could be employed for further genetic data analysis. Although monomorphic SNPs were regarded as non-informative and excluded from further analysis for this study, they could still show polymorphism in larger sampling groups. Monomorphic SNPs could be employed in populations where prior estimates of effective population size (N_e) and mutation rates (μ) are available (Morin *et al*. 2004). Unfortunately, due to the limited number of individuals that was genotyped in the current study, the use of monomorphic SNPs can only be substantiated within the populations when more individuals are genotyped. Six populations consisting of three commercial and three natural populations distributed along the coast of South Africa (Figure 4.3) were selected. Genetic variation was assessed across all loci within the six populations.

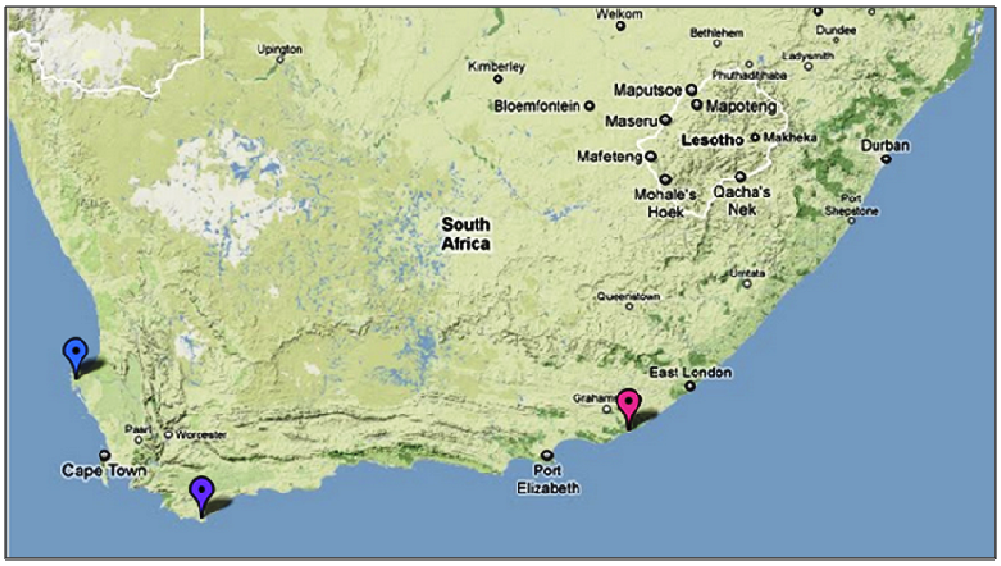

Figure 4.3. A map of South Africa indicating the three sampling locations of the natural populations of *H.midae* included in this study(Blue – Saldanha Bay, Purple – Witsand, Pink – Riet Point).

The low to moderate levels of heterozygosity along with the average MAF observed in the majority of the SNPs are comparable to reports in Pacific abalone, *H. discus hannai* (average MAF: 18.0%, 14.1%, 26.5%, respectively) (Qi *et al*. 2008, 2009; Zhang *et al*. 2010) but differ to reports of SNPs isolated in the South African abalone, *H. midae* {average MAF: 30.5%, 47.0%, 32.0%, respectively (Bester *et al*. 2008, Rhode 2010, Bester-van der Merwe *et al*. 2011)}. Although the aforementioned is indicative of a fairly significant degree of homozygosity within these populations, no inbreeding (average *f* = -

0.188) is supported as indicated by an excess of heterozygotes and one could conclude that inbreeding is not observed more than what is expected from a random mating species (Holsinger and Weir 2009). An excess of heterozygotes could indicate the presence of heterozygote advantage, a phenomenon known as the survival of heterozygote over homozygous individuals due to natural selection (Keller and Waller 2002). Even if $f = 0$, inbreeding could still occur in random mating populations due to differential contribution of the parents and when parents are more closely related than randomly selected individuals leading to pedigree inbreeding. The latter is often seen in a small population where only a few individuals contribute to the gene pool. Unequal contribution of parents could be attributed to factors including the quality and age of the broodstock along with poor physiological conditions (Slabbert *et al*. 2009). Genetic diversity along with MAF showed no loss of genetic variation among either the commercial or natural populations. These values were in accordance with various other marine species including the Weathervane scallop, *Patinopecten caurinus* (Elfstrom *et al*. 2005); the Japanese scallop, *Patinopecten yessoensis*(Liu *et al*. 2011) and the Mediterranean mussel, *Mytilus galloprovincialis*(Vera *et al*. 2010) as well as other studies on *Haliotid* species (Qi *et al*. 2008, 2009; Slabbert *et al*. 2009; Rhode 2010; Zhang *et al*. 2010, Bester-van der Merwe *et al*. 2011). For the current study, a higher level of heterozygosity was observed in the commercial populations than their natural counterparts. The high degree of diversity observed in commercial populations could be attributed to the parents originating from diverse natural populations, thus contributing a higher degree of genetic variability to the commercial population. Management of these populations in terms of spawning and grading practices must be considered in order to maintain this high degree of heterozygosity.

The exact test for HWE showed that all natural populations were in equilibrium, indicating no deviations from Hardy-Weinberg expectations. The H-W law states that allele and genotype frequencies remain constant between generations for a given population. This can only be preserved if a series of assumptions are maintained within the populations such as random mating, no migration or mutation, a large population size and no selection for any of the genotypes (Hardy 1908, Weinberg 1908). Therefore, if the individuals chosen from each population represent the population as a whole, one would expect these populations to be in equilibrium considering the random mating behaviour and high fecundity of *H. midae*(Wood and Buxton 1996). Conversely, the commercial populations deviated ($P < 0.05$) significantly from Hardy-Weinberg expectations. Various factors

contribute to such deviations in a closed breeding system such as these represented by the commercial populations. Non-random mating could cause deviations from Hardy-Weinberg equilibrium in broadcast spawning marine species (Hedgecock 1986), especially when only a few individuals contribute to the next generation's gene pool; inevitably leading to a smaller effective population size. Selection could also attribute to a population deviating from HWE, especially when broodstock individuals are selected for favourable economic traits such as growth. This could lead to a loss of overall genetic diversity.

Apart from SNP markers previously isolated by Bester *et al.* (2008) that showed linkage disequilibrium within the same EST fragment, markers found to be in linkage disequilibrium in the current study showed no homology with each other or a common gene (Table S3: Appendix D). Due to relatively little information available for the *H. midae* genome, homology could however still be the reason for the SNPs being in linkage disequilibrium. Linked markers reduce the coverage of the genome and could in theory diminish their statistical power for population analysisespecially when only a limited number of markers are available (Bester-van der Merwe *et al*. 2011). Emphasis is therefore placed on the SNP discovery procedure to maximise the number of ESTs utilised whilst avoiding ESTs exhibiting homology to similar genes. This would reduce the number of linked SNPs detected and thus could increase the information obtained from these markers. Nielsen *et al.* (2005) found that unlinked SNPs were more attainable in rapidly evolving genes than conserved regions. Although avoiding linked loci is preferable in terms of statistical analysis, Morin *et al.* (2009) however demonstrated that using haplotypes and including multiple SNPs within loci rather than using only unlinked markers could increase the statistical power of the markers when inferring population structure and therefore markers found to be in linkage disequilibrium in this study could still be useful.

Ascertainment bias was compensated for, in this study, by genotyping the loci over large geographically dispersed populations, and using a sufficient sequencing depth for *in silico* SNP identification prior to genotyping (Morin *et al*. 2004; Helyar *et al*. 2011). This is especially important when the unrepresentative group is subjected to next generation sequencing (Seeb *et al*. 2011). In cases where the aforementioned criteria are not adhered to, loci with very low MAF could be missed during detection (Helyar *et al*. 2011), resulting in an underestimated genetic diversity that can lead to biased estimates in linkage disequilibrium, demographic changes, population size and nucleotide diversity, to

name a few (Akey *et al*. 2003; Nielsen and Signorovitch 2003; Rosenblum and Novembre 2007; Storz and Kelly 2008; Guillot and Foll 2009; Moragues *et al*. 2010).

The polymorphic SNPs developed in this study demonstrated suitability for population structure inference based on their levels of heterozygosity and MAF, which are key determinants for the utilisation of loci in population analysis.

4.2 Population Differentiation

Haliotis midae are found along the South African coast from St Helena Bay on the West coast to Black Rock on the East coast from which only three sampling populations were included in this study. Populations included abalone from the West (Saldanha Bay), South (Witsand) and East coast (Riet Point) of South Africa (Figure 4.3). Hemmer-Hansen *et al*. (2007) attributed population structure in marine species to various factors including oceanic barriers, geographical distances and environmental changes.The population structure in *H. midae* is mainly attributed to the major currents that flow along the South African coast contributing to larval dispersal patterns and subsequent structuring (Turpie *et al*. 2000; Evans *et al*.2004b; Bester-van der Merwe *et al*. 2011). At the southern tip of Africa, called Cape Agulhas, the ocean waters are divided by the cold Benguela current flowing upwards to the west and the warm Agulhas current flowing down from the East, causing a biogeographic barrier to dispersal (Figure 4.4)(Dijkstra and De Ruijter 2001; Teske *et al*. 2011) possibly limiting gene flow from the western basin to the eastern one (Teske *et al*. 2006; Von der Heyden *et al*. 2008). The bi-directionality of these currents, along with the upwelling of the Benguela current, results in excessive temperature fluctuations between Cape Point and Cape Agulhas, as opposed to the Algoa Bay region (east of Port Elizabeth) that has a perennial thermal front due to the Agulhas current (Beckley and van Ballegooyen 1992; Lutjeharms and Ansorge 2001). Bester-van der Merwe *et al*. (2011) found that these barriers coincide with genetic discontinuities in *H. midae* especially at the biogeographic break at Cape Agulhas.

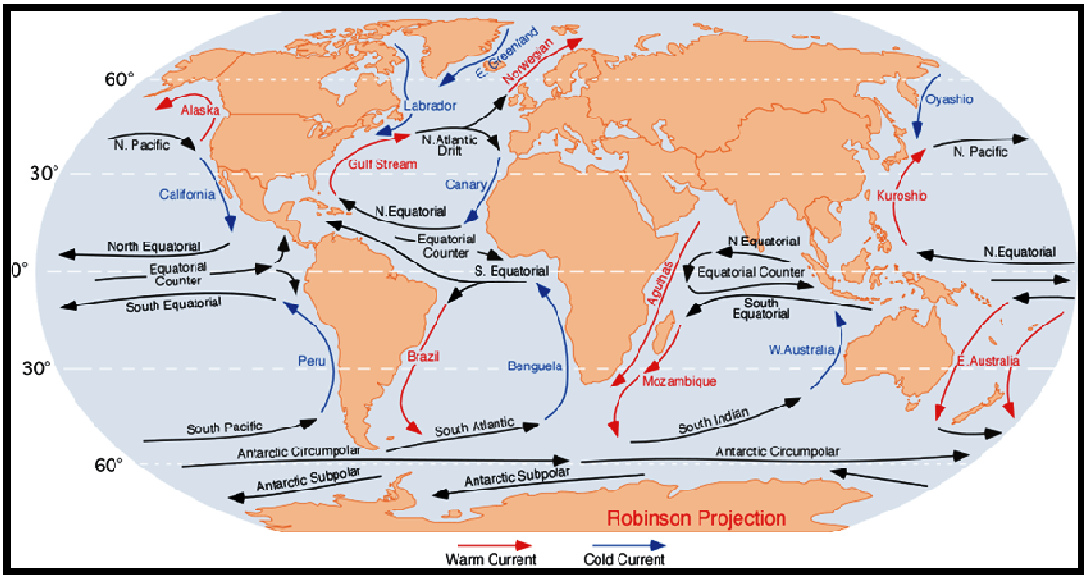

Figure 4.4.Illustration of currents around South Africa (Pidwirny 2006).

Population structure was inferred based on summary statistics (Wrights' F-statistics) as well as multivariate analysis (FCA). F-statistics are the most widely used descriptive statistic to infer population differentiation among and within populations (Holsinger and Weir 2009). They are directly linked to the degree of genetic diversity within populations due to the differences in allele frequencies between individuals within populations and thus to the degree of resemblance within different populations. Population structure analysis indicated that allele fixation (increasing F_{ST}) is evident between commercial and natural populations with significant genetic differentiation between them (Holsinger and Weir 2009). Commercial populations originated from individuals taken from the wild (natural populations) and given that no additional individuals from the natural population were introduced to the commercial populations, it is not unusual to observe varying allele frequencies and thus allele fixation between the natural and commercial populations. This raises concerns regarding the genetic variation within commercial populations as levels of homozygosity increases and ultimately genetic drift takes place. Genetic drift is the change in allele frequencies within a population over generations due to random sampling. This will inevitably lead to alleles "drifting" from the population and thus a decrease in genetic variation. Genetic drift is more pronounced in small populations where only a few individuals serve as the founders of the population; a situation frequently found in cultured populations. Ultimately, the allele frequencies will differ significantly from that of the source

population. This could attribute to the differentiation seen between the commercial and natural populations.

Differentiation within the commercial populations could also be attributed to localised adaptation, which is the selection of favourable loci within certain environments (Beaumont and Hoare 2003). Natural selection for specific alleles can be negatively influenced when these alleles are lost due to localised adaptation. The latter is exacerbated in small populations because genetic drift will occur at an accelerated pace and have a more pronounced effect than in larger populations. An increase in homozygosity (as indicated for the commercial populations in this study) could lead to lower diversity and possible inbreeding. The effects of inbreeding include poor immune responses, slow growth and deleterious phenotypic effects such as reduced fecundity and poor larval and juvenile development (Beaumont and Hoare 2003). This reduction in population fitness is referred to as inbreeding depression and is indicated by an excess of homozygotes.

Small populations with little migration or larval dispersal are highly differentiated from each other as alleles are not carried over to subsequent subpopulations (Holsinger and Weir 2009). After spawning, abalone larvae continue through a series of life stages before reaching the settling stage, where they hold on to rocky surfaces. It is in the stages prior to settling that the larvae remain planktonic and are vulnerable to oceanic currents and dispersal occurs. Therefore, larvae are dispersed over vast distances inducing cross-population gene flow. In the case of *H. midae*, the low F_{ST} values between natural populations indicate that these populations have similar genotypic frequencies and that larval dispersal does exist between populations (Beaumont and Hoare 2003). Due to the oceanic currents along the South African coast, one can hypothesise that gene flow is unidirectional and occurs from the East coast towards the West coast. Also, oceanographic barriers created by the currents may be more permeable than anticipated and therefore no strong barriers to gene flow between *H. midae* populations are expected. However low, the significant genetic differentiation observed here and previously is mainly attributed to the biogeographic break at Cape Agulhas. Additional analyses such as AMOVA could be conducted to substantiate this possible sub-division.

Factorial correspondence analysis was used to determine the genetic relationship among the different natural and commercial populations and supported population structure

inference obtained through summary statistics. It also indicated that a larger degree of genetic differentiation exist between the natural and commercial populations, than within each of these respective groups.

The results obtained with the SNPs applied in this study indicated that commercial populations of *H. midae* still maintains a significant degree of genetic diversity and that natural populations show patterns of weak population structuring. Both the current and previous study (Bester-van der Merwe *et al*. 2011) utilising SNPs in *H. midae* emphasises the ability of only a small number of SNP markers to successfully infer population structure and diversity estimation.

Chapter 5
Concluding Remarks
and
Future Applications

1 Introduction

This study was conducted to determine both the efficiency and reliability of markers developed from next generation sequencing data in a non-model species, *H. midae*. The effectiveness of these markers was then determined by means of high-throughput genotyping with the Illumina GoldenGate genotyping assay with the VeraCode technology. Single Nucleotide Polymorphisms were selected as the markers of choice for this study since marker development conducted on *H. midae* have thus far focused primarily on microsatellite markers. To date, only a small number of SNPs (Bester *et al*. 2008; Rhode *et al*. 2008; Rhode 2010) have been isolated within the species and have shown great potential in applications,including genetic diversity and population differentiation studies (Bester-van der Merwe *et al*. 2011).

2 Integration of next generation technologies for the isolation of molecular markers

In the past, sequencing was largely dominated by the chain-termination concept, including but not limited to, Sanger sequencing (Sanger *et al*. 1977). Various other methodologies have since been introduced despite the eminent success of the Sanger technology;includingpyrosequencing andsequencing-by-synthesis. These improved technologies (termed next generation sequencing) were introduced to improve throughput and efficiency in the generation of sequencing data. NGS constitutes the collaboration of sequencing, imaging, template preparation and assembly. Factors such as amplification and purification of DNA as well as traditional preliminary cloning have been reduced with the use of NGS methodologies (Quinn *et al*. 2008; Metzker 2010). Therefore, the advent of NGS has revolutionised approaches in research, previously only for model species but currently also for non-model species.

A major application of NGS is transcriptome characterisation, especially in non-model species. This includes the annotation of the transcriptome to a public database or genomic reference and is best achieved with a broad representation of the genome. Assorted tissue

types from a number of individuals preferably at different life stages are therefore required for a complete representation of a species' transcriptome (Hahn *et al*. 2009). This characterisation allows for a variety of applications including candidate gene discovery, gene expression profiling and more importantly marker discovery (Varshney *et al*. 2009). The association of markers with functional genes, including large-scale development, makes marker isolation from transcriptomic data an appealing aspect for more targeted research areas. NGS is however not without limitations; these include limited sequence assembly due to the production of short reads that obstructs primer design and data storage. Despite these restrictions, key attributes for example accuracy and high-throughput at an affordable rate increases the popularity of these technologies. Future prospects of NGS aim to minimise current difficulties associated with sequencing errors by the production of longer reads, pair-end sequencing as well as to reduce the amount of genetic material needed (Ekblom and Galindo 2010; Liu, S.*et al*. 2011).

3 Importance of molecular markers in studies of non-model species such as *H. midae*

Previously, SNP discovery in non-model species relied primarily on Sanger sequencing, followed by genotyping of only a small selection of SNPs (Seeb *et al*. 2011). Although the effectiveness of chain-termination methods for non-model species have been demonstrated in species for example Atlantic cod (*Gadus morhua*) (Hemmer-Hansen *et al*. 2011) and Rainbow trout (*Oncorhynchus mykiss*) (Salem *et al*. 2010; Abadìa-Cardoso *et al*. 2011), complete sequence information is still missing for these species. Such limited sequencing information of genomes restricts knowledge regarding the location of gene-associated markers as well as the effects that genetic mechanisms such as alternative splicing and differential expression have on candidate genes. Through the introduction of NGS and the advances that follow, larger data sets can be acquired with increased marker development at a more affordable rate. The Illumina GA II yielded 127 687 nodes from only 19 individuals that were used to construct contiguous sequences, as described in van der Merwe (2010), which indicates the robustness and high-throughput nature of NGS. In the current study, this resource facilitated the design of a small panel (97) of primer pairs. The only major limitation

in the current study presented itself in the number of optimised SNPs (39%). Several reasons could explain to the low optimisation rate, the most prominent being the presence of intron-exon junctions. This can be overcome by techniques such as described by Geraldes *et al*. (2011) and Scofield *et al*. (2007). An advisable solution would be the use of programs such as DNASTAR's Lasergene® GeneQuest™ (http://www.dnastar.com/t-sub-products-lasergene-genequest.aspx [accessed October 2011]) and Genscan (http://genes.mit.edu/GENSCAN.html [accessed October 2011]) to accurately predict the location of these boundaries. Another drawback with the use of transcriptomic data for SNP discovery is the limited number of markers discovered due to coding regions being more conserved. Regardless of this limitation, our finding of a ~1% SNP density of the developed markers provides a good indication of the large number of markers that can be obtained with NGS technologies.

The validity of any marker is determined by the process of genotyping in which the amount of information gained (polymorphism) from that marker is assessed. All genotyping methods have advantages and disadvantages; all dependent on different factors, including accuracy, capacity to multiplex, cost, assay development and throughput (Sobrino *et al*. 2005; Syvänen 2005). The objective of this study was to employ a medium- to high-throughput genotyping platform for the validation of SNPs isolated from *H. Midae*, a non-model species. The GoldenGate genotyping assay coupled with the VeraCode technology provided the ideal means of executing this objective. To our knowledge, this is the first undertaking of medium-to high-throughput genotyping of SNP markers in any Haliotidspecies. The genotyping success rate of 85% demonstrates the efficiency and potential of using ESTs as a resource for SNP discovery as well as the efficiency with which the GoldenGate assay with the VeraCode technology functions. This genotyping platform requires minimal bench-work and pre-processing, is both flexible and scalable; which are all factors highlighting its customisability. Apart from determining the validity of the SNPs isolated in this study, several other SNPs that were isolated by different methods (microsatellite flanking regions, target gene approach and *in silico*) were also evaluated. Due to the high failure rate of genotyped SNPs that were isolated from microsatellite flanking regions, it is strongly recommended that these regions are not used for marker development when considering this platform. The genotyping of these SNPs were unsuccessful due to possible SNP hotspots within these

flanking regions, which could have disrupted primer binding. The recommended use of *in silico* SNPs for future endeavours is supported with the results obtained in the current study. Although the genotyping success rate was more or less equal between the *in vitro* and *in silico* SNPs, the latter marker isolation method was more efficient in terms of time consumption and cost.

4 Applications of markers in a non-model species such as *H. midae*

The importance of the number of markers generated from NGS in conjunction with medium- to high-throughput genotyping methodologies lie in downstream applications such as population structure inference, linkage studies,diversity assessments and Marker-Assisted Selection (MAS) (Slate *et al*. 2009; Varshney *et al*. 2009; Ujino-Ihara *et al*. 2010). Traditionally, linkage maps in aquaculture species have been predominantly constructed with microsatellite markers due to their hyper-variability; however, the focus is shifting to the employment of SNPs in aquaculture research. Even though SNPs are less variable than microsatellites and more markers are required, studies have indicated that the use of SNPs is sufficient in accuracy and robustness for this purpose (Slate *et al*. 2009). Therefore, the employment of SNPs along with microsatellites for the development of linkage maps could prove to be quite beneficial in non-model species. This is a future undertaking with the SNPs developed in the current study (Jansen 2011).With the improvements in NGS technology, larger sets op SNPs can be typed which would facilitate genome-wide association mapping without the prerequisite of a linkage map (Slate *et al*. 2009).

The inference of diversity and population structure is an important aspect in the conservation and management of commercial as well as natural populations. Marker development plays an integral role in the execution of these applications as these polymorphisms can be used to detect variations within and among populations. The usefulness of any marker in population studies is measured by the total number of alleles; a markers' discriminatory power. As indicated before, population studies relied predominantly on the use of microsatellite markers due to their high variability; however the use of highly informative SNPs, i.e. SNPs that show

a large allele frequency variation among populations, may exceed the use of microsatellites (Liu *et al*. 2005; Helyar *et al*. 2011). The SNPs found in the current study not only conformed to high information content, but also proved to be efficient for population-related studies. This study indicated that no genetic variation was lost since the establishment of the commercial populations (first generation individuals); however, this does not exclude this from happening in subsequent generations. Although all the populations (natural and commercial alike) share some alleles, it is not recommended to restock natural populations with cultivated juveniles as this can result in either an outbreeding or inbreeding depression (Ward 2006). This is especially important considering the genetic differentiation observed between the western and eastern coast abalone, which concurs with the study done by Bester-van der Merwe *et al*. (2011). Restocking (the restoration of spawning biomass) and stock enhancement (the increasing of the natural supply of juveniles) programmes are one of the few ways to address conservation issues. Unfortunately, natural populations can be negatively affected due to the genetically deleterious effects of hatchery practices (Ward 2006). Therefore, a comprehensive understanding of population structure is required before such programmes are applied.

5 Implications of molecular markers within perlemoen, a non-model species

Haliotis midae is a very important commercial commodity and contributing source to the South African economy. With constraints including poaching, habitat degradation and natural predation, as well as with the eradication of the CITES regulations in 2010, management and conservation of the species requires constant improvement. A complete understanding of intra-population diversity and inter-population structure, with the aid of molecular markers, is crucial for the application of such undertakings. Therefore, the markers developed in this study will be useful resources for applications such as linkage mapping and QTL analysis; applications that will aid in the management of this species.

To date, only two completed linkage maps exist for *Haliotis* midae, of which both lack the integration of SNPs (Badenhorst 2008; Hepple 2010). However, a third map is currently under construction, which combines the use of gene-linked SNPs with microsatellites (Jansen 2011). Although successful, a larger number of SNPs is required for linkage mapping and so, future studies will greatly benefit from the use of NGS as well as high-throughput typing studies to isolate SNPs. This will inevitably facilitate the construction of a high-density linkage map for the species.

This study indicates that the markers characterised are sufficient for the inference of genetic diversification and population structure studies. Coupled with the advances of NGS technologies and high-throughput genotyping, new frontiers are more accessible for studies on non-model species.

REFERENCES

References

Abadìa-Cardoso, A., Clemento, A.J., Garza, J.C. (2011).Discovery and characterization of single nucleotide polymorphisms in steelhead/rainbow trout, *Oncorhynchus mykiss*.*Molecular Ecology Resources, 11 (S1)*, 31-49.

Adams, M.K. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science, 252*, 1651-1656.

Akey, J.M., Zhang, K., Xiong, M., Jin, L. (2003).The effect of single nucleotide polymophism identification strategies on estimates of linkage disequilibrium.*Molecular Biology and Evolution, 20*, 232-242.

Akhunov, E., Nicolet, C., Dvorak, J. (2009). Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay.*Theoretical and Applied Genetics, 119*, 507-517.

Allendorf, F.W., Phelps, S.R. (1980). Loss of genetic variation in a hatchery stock of cutthroat trout.*Transactions of the American Fisheries Society, 109*, 537-543.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). Basic local alignment search tool.*Journal of Molecular Biology, 215*, 403-410.

Altschuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J., Donnelly, P. (2005). International HapMap Consortium.A haplotype map of the human genome.*Nature, 437*, 1299-1320.

Anderson, E.C., Garzam J.C. (2006). The power of single-nucleotide polymorphisms for large-scale parentage inference.*Genetics, 172*, 2567-2582.

Andreassen, R., Lunner, S., Hoyheim, B. (2010).Targeted SNP discovery in Atlantic salmon (*Salmo salar*) genes using a 3'UTR-primed SNP detection approach.*BMC Genomics, 11*, 706.

References

Andrews, C.A. (2010). Natural selection, genetic drift, and gene flow do not act in isolation in natural populations. *Nature Education Knowledge, 1*, 5.

Arnheim, N., Calabrese, P. (2009). Understanding what determines the frequency and pattern of human germline mutations. *Nature Reviews Genetics, 10*, 478-488.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin G.M., Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.*Nature Genetics, 25*, 25-29.

Badenhorst, D. (2008). *Development of AFLP markers for Haliotis midae for linkage mapping.*Unpublished MSc thesis.Stellenbosch University.

Badenhorst, D., Roodt-Wilding, R. (2007).Application of various DNA extraction methodologies on abalone, *Haliotis midae*, larvae and juveniles for fluorescent AFLP analysis.*Aquaculture Research, 38*, 546-549.

Bai, Z., Yin, Y., Hu, S., Wang, G., Zhang, X., Li, J. (2009). Identification of genes involved in immune response, microsatellite, and SNP markers from Expressed Sequence Tags generated from hemocytes of freshwater pearl mussel (*Hyriopsis cumingii*). *Marine Biotechnology, 11*, 520-530.

Baranski, M., Rourke, M., Loughan, S., Austin, C. (2006).Isolation and characterisation of 125 microsatellite DNA markers in the blacklip abalone, *Haliotis rubra.Molecular Ecology Notes, 10*, 1-6.

Baranski, M., Rourke, M., Loughnan, S., Hayes, B., Austin, C., Robinson, N. (2008).Detection of QTL for growth rate in the blacklip abalone (*Haliotis rubra* Leach) using selective DNA pooling.*Animal Genetics, 39*, 606-614.

References

Barany, F. (1991).Genetic disease detection and DNA amplification using cloned thermostable ligase.*Proceedings of the National Academy of Sciences of theUSA, 88*, 189-193.

Barker, G., Batley, J., O'Sullivan, H., Edwards, K.J., Edwards, D. (2003). Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics, 19*, 421-422.

Beaumont, A.R., Hoare, K. (2003). *Biotechnology and Genetics in Fisheries and Aquaculture.* Oxford: Blackwell Publishing, pp. 202.

Beckley, L.E., Van Ballegooyen, R.C. (1992). Oceonographic conditions during three ichtyhoplankton surveys of the Agulhas Current in 1990/1991. *South African Journal of Marine Sciences, 12*, 83-93.

Belfiore, N.M., Hoffman, F.G., Baker, R.J., Dewoody, J.A. (2003). The use of nuclear and mitochondrial single nucleotide polymorphisms to identify cryptic species.*Molecular Ecology, 12*, 2011-2017.

Bensch, S., Akesson, S., Irwin, D.E. (2002). The use of AFLP to find an informative SNP: genetic differences across a migratory divide in willow warblers. *Molecular Ecology, 11*, 2359-2366.

Benzie, J.A., Ballment, E., Forbes, A.T., Demetriades, N.T., Sugama, K., Moria, S. (2002). Mitochondrial DNA variation in the Indo-Pacifc populations of the giant tiger prawn, *Penaeus monodon.Molecular Ecology, 11*, 2553-2569.

Bester, A.E., Slabbert, R., D'Amato, M.E. (2004).Isolation and characterisation of microsatellite markers in the South African abalone (*Haliotis midae*).*Molecular Ecology Notes, 4*, 618-619.

Bester, A.E., Roodt-Wilding, R., Whitaker, H.A. (2008). Discovery and evaluation of single nucleotide polymorphisms (SNPs) for *Haliotis midae*: a targeted EST approach. *Animal Genetics, 39*, 321-324.

References

Bester-van der Merwe, A.E., Roodt-Wilding, R., Volckaert, F.A.M., D'Amato, M.E. (2011). Historical isolation and hydrodynamically constrained gene flow in declining populations of the South African abalone, *Haliotis midae*. *Conservation Genetics, 12*, 543-555.

Bhattramakki, D., Rafalski, A. (2001). Discovery and application of single nucleotide polymorphism markers in plants. In: R. Henry (Ed.), *Plant genotyping: The DNA fingerprinting of plants.* (pp. 179-192). Lismore: CABI publishing.

Black, W.C., Duteau, N.M. (1997). RAPD-PCR and SSCP analysis for insect population genetic studies. In: J.B. Crampton (Ed.), *The Molecular Biology of Insect Disease Vectors: A Methods Manual.* (pp. 361-373) New York: Chapman & Hall.

Black W.C. IV, Vontas, J.G. (2007). Affordable assays for genotyping single nucleotide polymorphisms in insects. *Insect Molecular Biology, 16*, 377-387.

Bolton, J. (2006). Do we have the vision to integrate our marine aquaculture? *South African Journal of Science, 102*, 507-508.

Botes, L., Smit, A.J., Cook, P.A. (2003). The potential threat of algal blooms to abalone (*Haliotis midae*) mariculture industry situated around the South African coast. *Harmful Algae, 2*, 247-259.

Botstein, D., White, R.L., Skolnick, M., Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics, 32*, 314-331.

Bouck, A., Vision, T. (2007). The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology, 16*, 907-924.

Boulding, E.G., Culling, M., Glebe, B., Berg, P.R., Lien, S., Moen, T. (2008). Conservation genomics of Atlantic salmon: SNPs associated with QTLs for adaptive traits in parr from four trans-Atlantic backcrosses. *Heredity, 101*, 381-391.

References

Britz, P. (1991). Prospects for abalone aquaculture. In: R.G.M. Heath (Ed.), *Proceedings of a Joint Symposium convened by the Aquaculture Association of South Africa and Stellenbosch University.* (pp. 110-113). Stellenbosch: Aquaculture '90.

Britz, P. (1996). *The nutritional requirements of Haliotis midae and the development of a practical diet for abalone aquaculture*.Unpublished PhD dissertation.Rhodes University.

Britz, P.J., Lee, B., Botes, L. (2009). *AISA 2009 Aquaculture Benchmarking Survey: Primary Production and Markets.* Grahamstown: AISA report produced by Enviro-Fish Africa (Pty.) Ltd. pp. 118.

Brown, B., Epifanio, J. (2003). Population Genetics: Principles and Applications for Fisheries Scientists. In: E. Hallermann (Ed.), *Nuclear DNA* (pp. 458). Bethesda: American Fisheries Society.

Brumfield, R.T., Beerli, P., Nickerson, D.A., Edwards, S.V. (2003). The utility of single nucleotide polymorphisms in inference of popuation history.*Trends in Ecology and Evolution, 18*, 249-256.

Buetow, K.H., Edmonson, M.N., Cassidy, A.B. (1999). Reliable identification of large numbers of candidate SNPs from public EST data.*Nature Genetics, 21*, 323-325.

Bürgener, M. (2010). Evaluation of the CITES Appendix III listing and delisting of South African abalone *Haliotis midae*. *TRAFFIC Bulletin, 23*, 42-48.

Carlson, C.S., Smith, J.D., Stanaway, I.B., Rieder, M.J., Nickerson, D.A. (2006). Direct detection of null alleles in SNP genotyping data.*Human Molecular Genetics, 15*, 1931-1937.

Cenadelli, S., Maran, V., Bongioni, G., Fusetti, L., Parma, P., Aleandri, R. (2007).Identification of nuclear SNPs in gilthead seabream.*Journal of Fish Biology, 70*, 399-405.

Chagné, D., Gasic, K., Crowhurst, R.N., Han, Y., Bassett, H.C., Bowatte, D.R., Lawrence, T.J., Rikkerink, E.H.A., Gardiner, S.E., Korban, S.S. (2008). Development of a set of SNP markers present in expressed genes of the apple. *Genomics, 92*, 353-358.

Chakravarti, A. (1999). Population genetics - making sense out of sequence.*Nature Genetics, 21*, 56-60.

Chen, K., McLellan, M.D., Wendl, M.C., Kasai, Y., Wilson, R.K., Mardis, E.R. (2007). PolyScan: An automatic indel and SNP detection approach to the analysis of human resequencing data. *Genome Research, 17*, 659-666.

Chen, X., Levine, L., Kwok, P.Y. (1999).Fluorescence polarization in homogeneous nucleic acid analysis.*Genome Research, 9*, 492-498.

Churbanov, A., Vořechovský, I., Hicks, C. (2010). A method of predicting changes in human gene splicing induced by genetic variants in context of cis-acting elements. *BMC Bioinformatics, 11*, 22.

Ciobanu, D.C., Bastiaansen, J.W.M., Margin, J., Rocha, J.L., Jiang D.H., Yu, N.,Geiger, B., Deeb, N., Rocha, D., Gong, H., Kinghorn, B.P., Plastow, G.S., van der Steen, H.A.M., Mileham, A.J. (2009). A major SNP resource for dissection of phenotypic and genetic variation in Pacific white shrimp (*Litopenaeus vannamei*). *Animal Genetics, 41*, 39-47.

Clark, N.L., Findlay, G.D., Yi, X., MaCoss, M.J., Swanson, W.J. (2007). Duplication and selection on abalone sperm lysin in an allopatric population.*Molecular Bioloogy and Evolution, 24*, 2081-2090.

Cockcroft, A.C., van Zyl, D., Huthcings, L. (2008). Large-scale changes in the spatial distribution of South African West Coast rock lobsters: an overview. *African Journal of Marine Science, 30*, 149-159.

Cross, T. (2000). Genetic implications of translocation and stocking of fish species, with particular reference to Western Australia.*Aquaculture Research*, 31, 83-94.

References

Curole, J.P., Hedgecock, D. (2005). *High frequency of SNPs in the Pacific Oyster genome.*Retrieved October, 31, 2011 from http://intl-pag.org/13/abstracts/PAG13_W026.html.

Day, E. (1998). *Ecological interactions between juvenile abalone (Haliotis midae) and sea urchins* (Parechinus angulosus) *off the west coast of South Africa*.PhD dissertation.University of Cape Town.

Day, E., Branch, G.M. (2002). Effects of sea urchins (*Parechinus angulosus*) on recruits and juveniles of abalone (*Haliotis midae*).*Ecological Monographs, 1* , 133-149.

De Waal, S.W.P., Branch, G.M., Navarro, R. (2003). Interpreting evidence of the dispersal by *Haliotis midae* juveniles seeded in the wild. *Aquaculture, 221*, 299-310.

DEAT (2003).*Policy for the allocation of commercial fishing rights in the abalone fishery.* Johannesburg: Department of Environmental Affairs and Tourism.

DEAT (2007).*Marine Living Resources Act (Act No. 18 of 1998). Draft regulations for the protection of abalone* (Haliotis) *(Wild) Government Gazette No. 3054, Notice R. 1141.* Johannesburg.

De La Cruz, F.L., Del Río-Portilla, M.Á., Gallardo-Escárate, C. (2010). Genetic variability of cultured populations of red abalone in Chile: An approach based on heterologous microsatellites. *Journal of Shellfish Research, 29*, 709-715.

Dijkstra, H.A., De Ruijter, W.P.M. (2001). Barotropic instabilities of the Agulhas Current system and their relation to ring formation. *Journal of Marine Research, 59*, 517-533.

Dodgson, J.B., Cheng, H.H., Okimoto, R. (1997). DNA marker technology: A revolution in animal genetics. *Poultry Science, 76*, 1108-114.

Eckert, A.J., Pande, B., Ersoz, E.S., Wright, M.H., Rashbrook, V.K., Nicolet, C.M., Neale, D.B. (2009). High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). *Tree Genetics and Genomes, 5*, 225-234.

Ekblom, R., Galindo, J. (2010). Applications of next generation sequencing in molecular ecology of non-model organisms.*Heredity*, 107, 1-15.

Elfstrom, C.M., Gaffney, P.M., Smith, C.T., Seeb, J.E. (2005). Characterization of 12 nucleotide polymorphisms in weathervane scallop.*Molecular Ecology Notes, 5*, 406-409.

Ellegren, H. (2000). Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics, 16*, 551-558.

Estoup, A., Angers, B. (1998). Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations. *Advances in Molecular Ecology*, 306, 55-89

Evans, B., Bartlett, J., Sweijd, N., Cook, P., Elliot, N.G. (2004a). Loss of genetic variation at microsatellite loci in hatchery produced abalone in Australia (*Haliotis rubra*) and South Africa (*Haliotis midae*). *Aquaculture, 233*, 109-127.

Evans, B.S., Sweijd, N.A., Bowie, R.C.K., Cook, P.A., Elliot, N.G. (2004b). Population genetic structure of the perlemoen*Haliotis midae* is South Africa: evidence of range expansion and founder events. *Marine Ecology Progress Series, 270*, 163-172.

Fahrenkrug, S.C., Freking, B.A., Smith, T.P.L., Rohrer, G.A., Keele, J.W. (2002).Single nucleotide polymorphism (SNP) discovery in porcine expressed genes.*Animal Genetics, 33*, 186-195.

Falconer, D.S., Mackay, T.F.C. (1996).*Introduction to Quantitative Genetics, 4th edn.* Harlow: Longmans Green. pp. 464.

Fallu, R. (1991). *Abalone farming.* Oxford: Fishing News Books. A division of Blackwell Scientific Publications Ltd. pp. 1-120.

Fallu, R., Lang, J. (1994). *All about abalone.* Agmedia, Victoria. Department of Conservation and Natural Resources, Victoria.Department of Agriculture.

References

Fan, J.B., Oliphant, A., Shen, R., Kermani, B.G., Garcia, F., Gunderson, K.L., Hansen, M., Steemers, F., Butler, S.L., Deloukas, P., Galver, L., Hunt, S., McBride, C., Bibikova, M., Rubano, T., Chen, J., Wickham, E., Doucet, D., Chang, W., Campbell, D., Zhang, B., Kruglyak, S., Bentley, D., Haas, J., Rigault, P., Zhou, L., Stuelpnagel, J., Chee, M.S. (2003). Highly parallel SNP genotyping. *Cold Spring Harbour Symposia on Quantitative Biology, 68*, 69-78.

FAO (2004).The State of the World Fisheries and Aquaculture 2004. Rome: FOA.

FAO (2009).*The State of World Fisheries and Aquaculture.*Retrieved September 25, 2010, from Food and Agriculture Organization of the United Nations.

FISHTECH (2010).*Facts about abalone*. Retrieved September 25, 2010, from FISHTECH: http://www.fishtech.com/facts.html

Flicek, P., Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature Methods, 6*, S6-S12.

Franchini, P., van der Merwe, M., Roodt-Wilding, R. (2011).Transcriptome characterization of the South African abalone *Haliotis midae* using sequencing-by-synthesis.*BMC Research Notes, 4*, 59.

Ganal, M.W., Altmann, T., Röder, M.S. (2009).SNP identification in crop plants.*Current Opinion in Plant Biology, 12*, 211-217.

Garvin, M.R., Saitoh, K., Gharrett, A.J. (2010). Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources, 10*, 915-934.

Geiger, D.L. (2000). Distribution and biogeography of the Haliotidae (Gastropoda: Vetigastropoda) world-wide. *Bollettino Malacologico, 35*, 57-120.

Genade, A.B., Hirst, A.L., Smit, C.J. (1988). Observations on the spawning, development and rearing of the South African abalone Haliotis midae Linn.*South African Journal of Marine Science, 6*, 3-12.

Geraldes, A., Pang, J., Thiessen, N., Cezard, T., Moore, R., Zhao, Y., Tam, A., Wang, S., Friedmann, M., Birol, I., Jones, S.J.M., Cronk, Q.C.B., Douglas, C.J. (2011).SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing.*Molecular Ecology Resources, 11 (S1)*, 81-92.

Groenewald, Y. (2009). *SA losing the abalone war.* Retrieved September 26, 2010, from Mail & Guardian: http://www.mg.co.za/article/2009-02-04-sa-losing-abalone-war.

Guillot, G., Foll, M. (2009).Correcting for ascertainment bias in the inference of population structure.*Bioinformatics, 25*, 552-554.

Guo, S.W., Thompson, E.A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles.*Biometrics, 48*, 361-372.

Gurvey, V., Berezikov, E., Malik, R., Plasterk, R.H.A., Cuppen, E. (2004). Single nucleotide polymorphism associated with rat expressed sequences. *Genome Research, 14*, 1438-1443.

Gut, I.G., Lathrop, G.M. (2004).Duplicating SNPs. *Nature Genetics, 36*, 789-790.

Hahn, D.A., Ragland, G.J., Shoemaker, D.D., Denlinger, D.L. (2009). Gene discovery using massively parallel pyrosequencing to develop ESTs for the fleshfly *Sarcophaga crassipalpis.BMC Genomics, 10*, 234.

Hahn, K. (1989). *Handbook of culture of abalone and other marine gastropods.* Boca Raton, Florida: CRC Press. pp. 348.

Hall, T. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series, 41*, 95-98.

Hansson, B., Kawabe, K. (2005). A simple method to score single nucleotide polymorphisms based on allele-specific PCR and primer-induced fragment-length variation. *Molecular Ecology, 5*, 692-696.

Hara, M., Sekino, M. (2007). Parentage testing for hatchery-produced abalone *Haliotis discus hannai* based on microsatellite markers: preliminary evaluation of early growth of selected strains in mixed family farming. *Fisheries Science, 73*, 831-836.

Hardy, H. (1908). Mendelian proportions in a mixed population.*Science, 28*, 49-50.

Harr, B., Turner, L.M. (2010). Genome-wide analysis of alternative splicing evolution among *Mus* subspecies.*Molecular Ecology, 19 (S1)*, 228-239.

Hauck, M., Sweijd, N.A. (1999).A case study of abalone poaching in South Africa and its impact on fisheries management.*Journal of Marine Science, 56*, 1024-1032.

Hayes, B., Laerdahl, J.K., Lien, S., Moen, T., Berg, P., Hindar, K., Davidson, W.S., Koop, B.F., Adzhubei, A., Høyheim, B. (2007a). An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. *Aquaculture, 265*, 82-90.

Hayes, B.J., Nilsen, K., Berg, P.R., Grindflek, E., Lien, S. (2007b). SNP detection exploiting multiple sources of redundancy in large EST collections improves validation rates. *Bioinformatics, 23*, 1692-1693.

Hecht, T. (1994). Behavioural thermoregulation of the abalone, *Haliotis midae*, and implications for intensive culture,.*Aquaculture, 126*, 171-181.

Hedgecock, D. (1986). Is gene flow from pelagic larval dispersal important in the adaptation and evolution of marine invertebrates? *Bulletin of Marine Science, 39*, 550-564.

Helyar, S.J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M.I., Ogden, R., Limborg, M.T., Cariani, A., Maes, G.E., Diopere, E., Carvalho, G.R., Nielsen, E.E. (2011). Application of

SNPs for population genetics of nonmodel organisms: new oppertunities and challenges. *Molecular Ecology Resources, 11 (S1)*, 123-136.

Hemmer-Hansen, J., Nielsen, E.E.G., Grønkjae, P., Loeschcke, V. (2007). Evolutionary mechanisms shaping the genetic population structure of marine fishes; lessons from the European flounder (*Platichthys flesus* L.). *Molecular Ecology, 16*, 3104-3118.

Hemmer-Hansen, J., Nielsen, E., Meldrup, D., Mittelholzer, C. (2011).Identification of single nucleotide polymorphisms in candidate genes for growth and reproduction in a non-model organism, the Atlantic cod, *Gadus morhua.Molecular Ecology Resources, 11 (S1)*, 71-80.

Hepple, J. (2010). An integrated linkage map of *Haliotis midae* (perlemoen).Unpublished MSc thesis.Stellenbosch University.

Hill, W. (2000). Maintenance of quantitative genetic variation in animal breeding programmes. *Livestock Production Scence, 63*, 99-109.

Hodgkinson, A., Eyre-Walker, A. (2010). Human triallelic sites: Evidence for a new mutational mechanism? *Genetics, 184*, 233-241.

Holsinger, K.E., Weir, B.S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting Fst. *Nature Reviews Genetics, 10*, 639-650.

Horton, R.M., Raju, R., Conti-Fine, B.M. (1997).Designing PCR primers to amplify specific members or subgroups of multigene families.*PCR protocols: Methods in Molecular Biology, 67*, 459-479.

Hubert, S., Bussey, J.T., Higgins, B., Curtis, B.A., Bowman, S. (2009). Development of single nucleotide polymorphism markers for Atlantic cod (*Gadus morhua*) using expressed sequences. *Aquaculture, 296*, 7-14.

Hubert, S., Higgins, B., Borza, T., Bowman, S. (2010). Development of a SNP resource and a genetic linkage map for the Atlantic cod (*Gadus morhua*).*BMS Genomics, 11*, 191.

Hyten, D.L., Song, Q., Fickus, E.W., Quigley, C.V., Lim, J.S., Choi, I.Y., Hwang, E.Y., Pastor-Corrales, M., Cregan, P.B. (2010). High-throughput SNP discovery and assay development in common bean. *BMC Genomics, 11*, 475.

Illumina. (2008). *GoldenGate® Genotyping with VeraCode™ Technology: Custom 96-plex and 384-plex Assays.* Retrieved June 28, 2010, from www.illumina.com: http://www.illumina.com/Documents/products/technotes/technote_veracode_goldengate_genotyping.pdf.

Illumina, (2010) *GenomeStudioTM Data Analysis Software.* Retrieved 13 July 2011, from Illumina: www.illumina.com/software/genomestudio_software.ilmn.

Illumina.(2010). *VeraCode Technology*. Retrieved September 24, 2010, from Illumina: www.illumina.com.

Jansen, S. (2011). *Linkage mapping in H. midae using gene-linked markers.*Unpublished MSc thesis.Stellenbosch University.

Jerry, D.R., Preston, N.P., Crocos, P.J., Keys, S., Meadows, J.R.S., Li, Y. (2004).Parentage determination of Kuruma shrimp *Penaeus* (*Marsupenaeus*) *japonicus* using microsatellite markers.*Aquaculture, 235*, 237-247.

Kang, J-H., Appleyard, S.A., Elliot, N.G., Jee, Y-J., Lee, J.B., Kang, S.W., Baek, M.K., Han, Y.S., Choi, T-J., Lee, Y.S. (2010).Development of genetic markers in abalone through construction of a SNP database.*Animal Genetics, 42*, 309-315.

Kawamura, T., Takami, H. (1995). Analysis of feeding and growth rate of newly metamorphosed abalone *Haliotis discus hannai* fed on four species of benthic diatom. *Fisheries Science, 61*, 357-358.

Keller, I., Bensasson, D., Nichols, R.A. (2007). Transition-Transversion bias is not universal: a counter example from grasshopper pseudogenes. *PloS Genetics, 3*, 185-191.

Keller, L.F., Waller, D.M. (2002). Inbreeding effects in wild populations. *Trends in Ecology and Evolution, 17*, 230-241.

Kerstens, H.H.D., Crooijmans, R.P.M.A., Veenendaal, A., Dibbits, B.W., Chin-A-Woeng, T.F.C., den Dunnen, J.T., Groenen, M.A.M. (2009). Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics*, *10*, 479.

Kim, S., Misra, A. (2007). SNP genotyping: Technologies and biomedical applications. *Annual Review of Biomedical Enginering, 9*, 289-320.

Kim, W.-J., Jung, H., Gaffney, P. (2010).Development of type I genetic markers from expressed sequence tags in highly polymorphic species. *Marine Biotechnology*, *13*, 1-6.

Kingdom, G.G., Knight, J. (2003). *Well-being poverty versus income poverty and capabilites poverty.* University of Oxford, Global Poverty Reduction group, Centre for the study of African Economics. Cape Town: Development Policy Research Unit.

Klinbunga, S., Pripue, P., Khamnamtong, N., Puanglarp, N., Tassanakajon, A., Jarayabhand, P., Hirono, I., Aoki, T., Menasveta, P. (2003).Genetic diversity and molecular markers of the Tropical abalone (*Haliotis asinina*) in Thailand. *Marine Biotechnology, 5*, 505-517.

Kondrashov, S. (2002). Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Human Mutation, 21*, 12-27.

Krangel, M., Langdon, S. (2011). *Sequence Troubleshooting*. Retrieved March 12, 2011, from Duke University Health System DNA Analysis Facility: http://www.cancer.duke.edu/DNA/docs/files/file/DNA%20Seq%20Troubleshooting.pdf.

Krizman, D.B., Berget, S.M. (1993). Efficient selection of 3' -terminal exons from vertebrate DNA. *Nucleic Acids Research, 21*, 5198-5202.

Kwok, P. (2001). Methods for genotyping single nucleotide polymorphisms. *Annual Review Genomics Human Genetics, 2*, 235-258.

Kwok, P.Y., Chen, X. (2003). Detection of single nucleotide polymorphism.*Current Issues in Molecular Biology, 5*, 43-60.

Lallias, D., Beaumont, A.R., Haley, C.S., Boudry, P., Heurtebise, S., Lapègue, S. (2007). A first-generation genetic linkage map of the European flat oyster *Ostrea edulis* (L.) based on AFLP and microsatellite markers. *Animal Genetics, 38*, 560-568.

Landegren, U., Nilsson, M., Kwok, P.Y. (1998). Reading bits of genetic information: methods for single nucleotide polymorphism analysis. *Genome Research, 8*, 769-776.

Le Dantec, L., Chagné, D., Pot, D., Cantin, O., Garnier-Géré, P., Bedon, F., Frigerio, J.M., Chaumeil, P., Léger, P., Garcia, V., Laigret, F., de Daruvar, A., Plomion, C. (2004). Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. *Plant Molecular Biology, 54*, 461-470.

Lehoczky, I., Jeney, Z., Magyary, I., Hancz, C., Kohlmann, K. (2005). Preliminary data on genetic variability and purity of common carp (*Cyprinus carpio* L.) strains kept at the live gene bank at Research Institute for Fisheries, Aquaculture and Irrigation (HAKI) Szarvas, Hungary. *Aquaculture, 247*, 45-49.

Lepoittevin, C., Frigerio, J-M., Garnier-Géré, P., Salin, F., Cervera, M-T., Vornam, B., Harvengt, L., Plomion, C. (2010).*In Vitro* vs *In silico* detected SNPs for the development of a genotyping array: What can we learn from a non-model species? *PLos One, 5* ,1-9.

Levinson, G., Gutman, G.A. (1987). High frequency of short framshifts in poly-CA/GT tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12.*Nucleic Acids Research, 15*, 5323-5338.

Lewis, I., Oeser, S., Chen, M., McDaniel, T., Yeakley, J. (2007).*Reliable and Accurate High-Throughput SNP Genotyping using the VeraCode™ GoldenGate® Genotyping Assay.* San Diego: Illumina Inc.

Lewis, P.O., Zaykin D. (2001). Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c).*Free program distributed by the authors over the internet from http://lewsi.eeb.edu/lewsihome/software.html*.

Li, L., Guo, Z. (2004). AFLP-based genetic linkage maps of the Pacific oyster *Crassostrea gigas* Thunberg. *Marine Biotechnology, 6*, 26-36.

Li, L., Xiang, J., Liu, X., Zhang, Y., Dong, B., Zhang, X. (2005). Construction of AFLP-based genetic linkage map for Zhikong scallop, *Chlamys farreri* Jones et Preston and mapping of sex-linked markers. *Aquaculture, 245*, 63-73.

Li, Q., Park, C., Kijima, A. (2003). Allelic transmission ofmicrosatellites and application to kinship analysis in newly hatched Pacific abalone larvae. *Fisheries Science, 69*, 883-889.

Li, Q., Park, C., Endo, T., Kijima, A. (2004). Loss of genetic variation at miocrosatellite loci in hatchery strains of the Pacific abalone (*Haliotis discus hannai*). *Aquaculture, 237*, 207-222.

Liao, P.-Y., Lee, K.H. (2010). From SNPs to functional polymorphism: The insight into biotechnology applications. *Biochemical Engineering Journal, 49*, 149-158.

Lin, C.H., Yeakley, J.M., McDaniel, T.K., Shen, R. (2009). Medium- to high-throughput SNP genotyping using VeraCode microbeads.*Methods in Molecular Biology, 496*, 129-142.

Litt, M., Luty, J.A. (1989). A hypervariable microsatellite revealed by *in vitro* amplification of dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics, 44*, 397-401.

Liu, N., Chen, L., Wang, S., Oh, C., Zhao, H. (2005). Comparison of single nucleotide polymorphisms and microsatellites in inference of population structure.*BMC Genetics, 6 (S1)*, S26.

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

Liu, W., Li, H., Bao, X., He, C., Li, W., Shan, Z. (2011). The first set of EST-derived single nucleotide polymorphism markers for Japanese scallop, *Patinopecten yessoensis*. *Journal of the World Aquaculture Society, 42*, 456-461.

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

Long, J. (1986).The allelic correlation structure of Gianj- and Kalam-speaking people. I. The estimation and interpretation of Wrigth's F-Statistics. *Genetics, 112*, 629-647.

Lucas, T., Macbeth, M., Degnan, S.M., Knibb, W., Degnan, B.M. (2006). Heritability estimates for growth in the tropical abalone *Haliotis asinina* using microsatellites to assign parentage. *Aquaculture, 259*, 146-152.

Lupski, J. (2007). Genomic rearrangements and sporadic disease.*Nature Genetics, 39*, S43-47.

Lutjeharms, J.R.E., Ansorge, I.J. (2001). The Agulhas Return Current.*Journal of Marine Systems, 30*, 115-138.

Lynch, M. (2007). *The Origins of Genome Architecture.* Sunderland, MA: Sinauer Associates.

Marine and Coastal Management.(2010). *Abalone.* Retrieved October 5, 2010, from www.aquarium.co.za: http://www.aquarium.co.za/images/uploads/3b_abalone.pdf.

Martinez, V. (2006).Importance and implementation of molecular markers in selective breeding programs for aquaculture species.*8th World congress on genetics applied to livestock production* (pp. 1-7). Belo Horizonte: CD-ROM communication 09–01.

Matsuzaki, Y., Nakano, A., Jiang, Q-J., Pulkkinen, L., Uitto, J. (2005).Tissue-specific expression of the ABCC6 gene.*Journal of Investigative Dermatology, 125*, 900-905.

May, B.J. (1993). Composite linkage map of salmonid fishes (*Salvelinus*, *Salmo*, and *Oncorhynchus*). In: S. O'Brien (Ed.), *Genetic Maps: Locus Maps of Complex Genomes.* (Vol. 4, pp. 309-317). Cold Spring Harbor, NY: Spring Harbor Laboratory Press.

McGoldrick, D.J., Hedgecock, D. (1997). Fixation, segregation and linkage of allozyme loci in inbred families of the Pacific oyster *Crassostrea gigas*: implications for the causes of inbreeding depression.*Genetics*,*146*, 321-334.

McShane, P. (1989). Early life history of abalone: a review. *Abalone of the World, Biology, Fisheries and Culture, Proceedings of the 1st International Symposium on Abalone.* (pp. 120-138). La Paz, Mexico: Fishing News Books.

McShane, P.E., Black, K.P., Smith, M.G. (1988). Recruitment processes in *Haliotis rubra* (Mollusca: Gasrtopoda) and regional hydrodynamics in southeastern Australia imply localized dispersal of larvae. *Journal of Experimental Marine Biology and Ecology, 124*, 175-203.

Meksem, K., Ruben, E., Hyten, D., Triwitayakorn, K., Lightfoot, D.A. (2001). Conversion of AFLP bands into high-throughput DNA markers. *Molecular Genetics and Genomics, 265*, 207-214.

Metzker, M. (2010).Sequencing technologies - the next generation.*Nature Reviews Genetics, 11*, 31-46.

Mir, K.U., Southern, E.M. (1999).Determining the influence of structure on hybridization using oligonucleotide arrays.*Nature Biotechnology, 17*, 788-792.

Moen, T., Hayes, B., Nilsen, F., Delghandi, M., Fjalestad, K.T., Fevolden, S.E., Berg, K., Lien, S. (2008). Identification and characterisation of novel SNP markers in Atlantic cod: evidence for directional selection. *BMC Genetics, 9*, 18.

Montpetit, A., Nelis, M., Laflamme, P., Magi, R., Ke, X., Remm, M., Cardon, L., Hudson, T.J., Metspalu, A. (2006). An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. *PloS Genetics, 2*, 282-290.

Moragues, M., Comadran, J., Waugh, R., Milne, I., Flavell, A.J., Russel, J.R. (2010). Effects of ascertainment bias and marker number on estimates of barley diversity from high-throughput SNP genotype data. *Theoretical and Applied Genetics, 120*, 1525-1534.

Moreno-Vazquez, S., Ochoa, O.E., Faber, N., Chao, S., Jacobs, M.J., Miason-Neuve, B., Kesseli, R.V., Michelmore, R.W. (2003).SNP-based codominant markers for a recessive gene conferring resistance to corky root rot (*Rhizomonas suberifaciens*) in lettuce (*Lactuca sativa*).*Genome, 46*, 1059-1069.

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

Morin, P.A., Martien, K.K., Taylor, B.L. (2009). Assessing statistical power of SNPs for population structure and conservation studies.*Molecular Ecology Resources, 9*, 66-73.

Morris, D.B., Richard, K.R., Wright, J.M. (1996). Microsatellites from rainbow trout (*Oncorhynchus mykiss*) and their use for genetic study of salmonids.*Canadian Journal of Fisheries and Aquatic Science, 53*, 120-126.

Muller, S. (1986). Taxonomy of the genus *Haliotis* in South Africa.*Transactions of the Royal Society of South Africa, 46*, 69-77.

Narum, S.R., Banks, M., Beacham, T.D., Bellinger, M.R., Campbell, M.R., Dekoning, J., Elz, A., Guthrie, C.M. 3rd, Kozfkay, C., Miller, K.M., Moran, P., Phillips, R., Seeb, L.W., Smith, C.T., Warheit, K., Young, S.F., Garza, J.C. (2008). Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms.*Molecular Ecology, 17*, 3464-3477.

Newman, G. (1967). Reproduction of the South African Haliotis midae.*Investigational Report of the Division of Sea Fisheries, 64*, 1-24.

Nickerson, D.A., Tobe, V.O., Taylor, S.L. (1997). PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research, 25*, 2745-2751.

Nicod, J-C., Largiader, C.R. (2003). SNPs by AFLP (SBA): a rapid SNP isolation strategy for non-model organisms. *Nucleic Acids Research, 31*, e19.

Nielsen, R., Signorovitch, J. (2003). Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium.*Theoretical Population Biology, 63*, 245-255.

Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B.,Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., Sninsky, J.J., Adams, M.D., Cargill, M. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees.*PLoS Biology, 3*, 976-985.

Nucleics (2010). DNA sequencing traces with indels. Retrieved June, 25, 2010 from Nucleics: http://www.nucleics.com.

O'Brien, S. (1991). Molecular genome mapping: lessons and prospects. *Currunt Opinion in Genetics and Development, 1*, 105-111.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research, 27*, 29-34.

Okumuş, I., Çiftci, Y. (2003). Fish population genetics and molecular markers: II - molecular markers and their applications in fisheries and aquaculture. *Turkish Journal of Fisheries and Aquatic Sciences, 3*, 51-79.

Oliveira, E.J., Pádua, J.G., Zucchi, M.I., Vencovsky, R., Vieirra, M.L.C. (2006). Origin, evolution and genome distribution of microsatellites.*Genetics and Molecular Biology, 2*, 294-307.

Pati, N., Schowinsky, V., Kokanovic, O., Magnuson, V., Ghosh, S. (2004). A comparison between SNapshot, pyrosequencing and biplex invader SNP genotyping methods: accuracy, cost and throughput. *Journal of Biochemical and Biophysical Methods, 60*, 1-12.

Pavy, N., Pelgas, B., Beauseigle, S., Blais, S., Gagnon, F., Gesselin, I., Lamothe, M., Isabel, N., Bousquet, J. (2008). Enhancing genetic mapping of complex genomes throught the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce.*BMC Genomics, 9*, 21.

Perez, F.E. (2004). A sex-specific linkage map of the white shrimp *Penaeus* (*Litopenaeus*) *vannamei* based on AFLP markers.*Aquaculture, 242*, 105-118.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., Boyce-Jacino, M. (1999).Mining SNPs from EST databases.*Genome Research, 9*, 167-174.

Pidwirny, M. (2006). Surface and Subsurface Ocean Currents: Ocean Current Map. *Fundamentals of Physical Geography, 2nd Edition*, http://www.physicalgeography.net/fundamentals/8q_1.html.

Pitcher, G.C., Franco, J.M., Doucette, G.J., Powell, C.L., Mouton, A. (2001). Paralytic shellfish poisoning in the abalone *Haliotis midae* on the west coast of South Africa. *Journal of Shellfish Research, 20*, 895-904.

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.

Qi, H., Liu, X., Zhang, G. (2008). Characterization of 12 single nucleotide polymorphisms (SNPs) in Pacific abalone, *Haliotis discus hannai*.*Molecular Ecology Resources, 8*, 974-976.

Qi, H., Liu, X., Zhang, G., Wu, F. (2009). Mining expressed sequences for single nucleotide polymorphims in Pacific abalone (*Haliotis discus hannai*). *Aquaculture Research, 40*, 1661-1667.

Qi, H., Liu, X., Wu, F., Zhang, G. (2010).Development of gene-targeted SNP markers for genomic mapping in Pacific abalone *Haliotis discus hannai* Ino.*Molecular Biology Reports, 37*, 3779-3784.

Qi, L.I., Kijima, A. (2007). Sequences characterization of microsatellite DNA sequences in Pacific abalone (*Haliotis discus hannai*).*Journal of Ocean University of China, 6*, 47-52.

Qi, L., Yanhong, X., Ruihai, Y., Akihiro, K. (2007). An AFLP genetic linkage map of pacific abalone (*Haliotis discus hannai*).*Journal of Ocean University of China (English edition), 6*, 259-267.

Quilang, J., Wang, S., Li, P., Abernathy, J., Peatman, E., Wang, Y., Wang, L., Shi, Y., Wallace, R., Guo, X., Liu, Z. (2007). Generation and analysis of ESTs from the eastern oyster, *Crassostrea virginica* Gmelin and identification of microsatellite and SNP markers.*BMC Genomics, 8*, 157.

Quinn, N.L., Levenkova, N., Chow, W., Bouffard, P., Boroevich, K.A., Knight, J.R., Jarvie, T.P., Lubieniecki, K.P., Desany, B.A., Koop, B.F., Harkins, T.T., Davidson, W.S. (2008). Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome.*BMC Genomics, 9*, 404.

Raemaekers, S.J-P.N., Britz, P.J. (2009). Profile of the illegal abalone fishery (*Haliotis midae*) in the Eastern Cape Province, South Africa: Organised pillage and management failure. *Fisheries Research, 97*, 183-195.

Raemaekers, S., Hauck, M., Bürgener, M., Mackenzie, A., Maharaj, G., Plagányi, É.E., Britz, P.J. (2011). Review of the causes of the rise of the illegal South African abalone fishery and consequent closure of the rights-based fishery. *Ocean and Coastal Management, 54*, 433-445.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics.*Current Opinion in Plant Biology, 5*, 94-100.

Raible, F., Tessmar-Raible, K., Osoegawa, K., Wincker, P., Jubin, C., Balavoine, G., Ferrier, D., Benes, V., de Jong, P., Weissenbach, J., Bork, P., Arendt, D. (2005). Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*.*Science, 310*, 1325-1326.

Ramos-Onsins, S., Aguadé, M. (1998). Molecular evolution of the cecropin multigene family in *Drosophila*: Functional genes vs. pseudogenes. *Genetics, 150*, 157-171.

Renaut, S., Nolte, A.W., Bernatchez, L. (2010). Mining transcriptome sequences towards identifying adaptive singel nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. *Salmonidae*). *Molecular Ecology, 19*, 115-131.

Rhode, C. (2010). Development of gene-linked molecular markers in South African abalone (*Haliotis midae*) using an *in silico* mining approach. Unpublished MSc thesis.Stellenbosch University.

Rhode, C., Slabbert, R., Roodt-Wilding, R. (2008). Microsatellite flanking regions: a SNP mine in South African abalone (*Haliotis midae*).*Animal Genetics, 39*, 329.

Rice, W. (1989).Analyzing tables of statistical tests.*Evolution, 43*, 223-225.

Riju, A., Arunachalam, V. (2009). *Interspecific differences in single nucleotide polymorphisms (SNPs) and indels in expressed sequence tag libraries of oil palm Elaeis guineensis and E. oleifera.* Retrieved 12 February 2011, from Nature Proceedings: http://hdl.handle.net/10101/npre.2009.3593.1.

Robison, K. (2011). Semiconductors charge into sequencing. *Nature Biotechnology, 29*, 805-807.

Roodt-Wilding, R., Slabbert, R. (2006).Molecular markers to assist the South African abalone industry.*South African Journal of Science, 102*, 99-102.

Rosenblum, E.B., Novembre, J. (2007). Acertainment bias in spatially structured populations: a case study in the eastern fence lizard. *Journal of Heredity, 98*, 331-336.

Rousset, F. (2008). GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Research, 8*, 103-106.

Rudd, S. (2003). Expressed sequence tags: alternative or complement to whole genome sequence? *Trends in Plant Science*, *8*, 321-329.

Ruivo, N. (2007). Microsatellite genotyping of contributing broodstock and selected offspring of *Haliotis midae* submitted to a growth performance recording scheme. Unpublished MSc thesis.Stellenbosch University.

Salem, M., Rexroad, C.E. III, Wang, J., Thorgaard, G.H., Yao, J. (2010).Characterization of the rainbow trout transcriptome using Sanger and 454-pyrosequencing approaches.*BMC Genomics, 11*, 564.

Sales, J., Britz, P.J. (2001).Research on abalone (*Haliotis midae* L.) cultivation in South Africa.*Aquatic Research, 32*, 863-874.

Sanger, F., Nicklen, S., Coulson, A.R. (1977).DNA sequencing with chain-terminating inhibitors.*Proceedings of the National Academy of Science of the United States of America, 74*, 5463-5467.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., Bennetzen, J.L. (1998). The paleontology of intergene retrotransposons of maize.*Nature Genetics, 20*, 43-45.

Sarropoulo, E., Nousdili, D., Magoulas A.G.K. (2008). Linking the genomes of non-model teleosts through comparative genomics.*Marine Biotechnology, 10*, 227-233.

Scofield, D.G., Hong, X., Lynch, M. (2007). Position of the final intron in full-length transcripts: determined by NMD? *Molecular Biology and Evolution, 24*, 896-899.

Seeb, J.E., Carvalho, G., Hauser, L., Naish, K., Roberts, S., Seeb, L.B. (2011).Single nucleotide polymorphism (SNP) discovery and applictions of SNP genotyping in nonmodel organisms.*Molecular Ecology Resources, 11 (S1)*, 1-8.

Seeb, L.W., Templin, W.D., Sato, S., Abe, S., Warheit, K., Park, J.Y., Seeb, J.E. (2011). Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. *Moleular Ecology Resources, 11*, 184-206.

Sekino, M., Hara, M. (2007). Linkage maps for the Pacific abalone (Genus *Haliotis*) based on microsatellite DNA markers. *Genetics, 175*, 945-958.

Selvamani, M.J.P., Degnan, S.M., Degnan, B.M. (2001). Microsatellite genotyping of individual abalone larvae: parentage assignment in aquaculture. *Marine Biotechnology, 3*, 478-485.

Shen, R., Fan, J.B., Campbell, D., Chang, W., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Wickham-Garcia, E., McBride, C., Steemers, F., Garcia, F., Kermani, B.G., Gunderson, K., Oliphant, A. (2005).High-throughput SNP genotyping on universal bead arrays.*Mutation Research, 573*, 70-82.

Shendure, J., Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology, 26*, 1135-1145.

Slabbert, R. (2010). Identification of growth related quantitative trait loci within the abalone *Haliotis midae*, using comparative microsatellite bulked segregant analysis. Unpublished PhD dissertation.Stellenbosch University.

Slabbert, R., Bester, A.E., D'Amato, M.E. (2009). Analyses of genetic diveristy and parentage within a South African hatchery of the abalone *Haliotis midae* Linnaeus using microsatellite markers.*Journal of Shellfish Research, 28*, 369-375.

Slabbert, R., Hepple, J., Venter, A., Nel, S., Swart, L., Van den Berg, N.C., Roodt-Wilding, R. (2010) Isolation and inheritance of 44 microsatellite loci in the South African abalone *Haliotis midae* L. *Animal Genetics*, *41*, 332-333.

Slabbert R., Ruivo N.R., Van den Berg N.C., Lizamore D.L., Roodt-Wilding R. (2008) Isolation and characterisation of 63 microsatellite loci for the abalone, *Haliotis midae*. *Journal of the World Aquaculture Societ*y, *39,* 429-435.

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

Smith, C.T., Elfstrom, C.M., Seeb, L.W., Seeb, J.E. (2005). Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon.*Molecular Ecology, 14*, 4193-4203.

Sobrino, B., Brión, M., Carracedo, A. (2005). SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Science International, 154*, 181-194.

Sokolov, B. (1990). Primer extention technique for the detection of single nucleotide in genomic DNA.*Nucleic Acids Research, 18*, 3671.

Sonstegard, T. (2007).Making a cow jump over the moon: development of a bovine SNP assay (Presentation at conference).*Plant Animal Genome Conference XV.* San Diego.

Stapley, J., Reger, J., Feulner, P.G.D., Smadja, C., Galindo, J., Ekblom, R., Bennison, C., Ball, A.D., Beckerman, A.P., Slate, J. (2010). Adaptation genomics: the next generation. *Trends in Ecology and Evolution, 25*, 705-712.

Steinberg, J. (2005). *The illicit abalone trade in South Africa.* Institute for Security Studies Paper 105.

Storz, J.F., Kelly, J.K. (2008). Effects of spatially varying selection on nucleotide diversity and linkage disequilibrium: insights from deer mouse. *Genetics*, *180*, 367-379.

Subasinghe, R.P., Curry, D., McGladdery, S.E., Bartley, D. (2003). Recent Technological Innovations in Aquaculture.*Review of the State of World Aquaculture.* (pp. 59-74). Rome: FAO Fisheries Circular No. 886, Rev 2.

Surget-Groba, Y., Montoya-Burgos, J.I. (2010).Optimization of *de novo* transcriptome assembly from next-generation sequencing data.*Genome Research, 20*, 1432-1440.

Swart, L. (2011). *Microsatellite markers as a tool in genetic enhancement and husbandry of Haliotis midae: a South African case study.*Unpublished MSc thesis.Stellenbosch University.

Syvänen, A. (2001). Accessing genetic variation: Genotyping single nucleotide polymorphisms.*Nature Reviews Genetics, 2*, 930-942.

Syvänen, A. (2005). Toward genome-wide SNP genotyping.*Nature Genetics, 37*, 5-10.

Tarr, R. (1987). Biology and management of South African perlemoen.*Underwater, 3*, 41-42.

Tarr, R. (2003). *Perlemoen: The South African Abalone*. Retrieved February 16, 2009, from Department of Environmental Affairs, Republic of South Africa: http://www.environment.gov.za//Documents/Documents/2003Jun2_1/abalone_article_20062003.html.

Tarr, R.J.Q., Williams, P.V.G., MacKenzie, A.J. (1996). Abalone, sea urchins and rock lobster: a possible ecological shift may affect traditional fisheries. *South African Journal of Marine Science, 17*, 319-323.

Tassanakajon, A., Pongsomboon, S., Jarayabhand, P., Klinbunga, S., Boonsaeng, V.V. (1998). Genetic structure in wild populations of black tiger shrimp (*Penaeus monodon*) using randomly amplified polymorphic DNA analysis. *Journal of Marine Biotechnology, 6*, 249-254.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics, 4*, 41.

Tautz, D. (1989). Hypervariability of simple sequences as a general source of polymorphic DNA markers.*Nucleic Acids Research, 17*, 6463-6471.

Tegner, M.J., Levin, L.A. (1982). Do sea urchins and abalones compete in California kelp forest communities? In: J. Lawrence (Ed.), *Proceedings of the International Echinoderms Conference* (pp. 265-271). Tampa Bay: Balkema.

Teske, P.R., McQuaid, C.D., Froneman, P.W., Barker, N.P. (2006) Impacts of marine biogeographic boundaries on phylogeographic patterns in three South African estuarine crustaceans. *Marine Ecology Progress Series*, 314, 283-293

Teske, P.R., Von der Heyden, S., McQuaid, C.D., Barker, N.P. (2011) A review of marine phylogeography in southern Africa. *South African Journal of Science*, 107(5/6), 1-11 (DOI:10.4102/sajs.v107i5/6.514)

The International HapMap project (2003). *Nature, 426*, 789-796.

Thompson, J.D., Higgins, D.G., Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research, 22*, 4673-4680.

Toth, A.L., Varala, K., Newman, T.C., Miguez, F.E., Hutchison, S.K., Willoughby, D.A. Simons, J.F., Egholm, M., Hunt, J.H., Hudson, M.E., Robinson, G.E. (2007). Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science, 318*, 441-444.

Trape, S., Blel, H., Panfili, J., Durand, J.-D.(2009). Identification of tropical Eastern Atlantic Mugilidae species by PCR-RFLP analysis of mitochondrial 16S rRNA gene fragments.*Biochemical Systematics and Ecology, 37*, 512-518.

Troell, M., Robertson-Andersson, D., Anderson, R.J., Bolton, J.J., Maneveldt, G., Halling, C., Probyn, T. (2006). Abalone farming in South Africa: An overview with perspectives on

kelp resources, abalone feed, potential for on-farm seaweed production and socio-economic importance. *Aquaculture, 257*, 266-281.

Turpie, J.K., Beckley, L.E., Katua, S.M. (2000). Biogeography and the selection of priority areas for conservation of South African coastal fishes.*Biological Conservation, 92*, 59-72.

Ujino-Ihara, T., Taguchi, Y., Moriguchi, Y., Tsumura, Y. (2010). An effective method for developing SNP markers based on EST data combined with high resolution melt (HRM) analysis. *BMC Research Notes, 3*, 51.

Usesche, F.J., Gao, G., HanaFey, M., Rafalski, A. (2001). High-throughput identification, database storage and analysis of SNPs in EST sequences.*Genome information, 12*, 194-203.

Van Bers, N.E.M., van Oers, K., Kerstens, H.H.D., Dibbits, B.W., Crooijmans, R.P.M.A., Visser, M.E., Groenen, M.A.M. (2010). SNP detection in the great tit, *Parus major* using high throughput sequencing.*Molecular Ecology, 19*, 89-99.

Van den Berg, N., Roodt-Wilding, R. (2010). Parentage assignment in *Haliotis midae* L.: a precursor to future genetic enhancement programmes for South African abalone. *Aquaculture Research, 41*, 1387-1395.

Van der Merwe (neé Bester), A. (2009). *Population genetic structure and demographical history of South African abalone, Haliotis midae, in a conservation context*.Unpublished PhD dissertation.Stellenbosch University.

Van der Merwe, M. (2010).*Growth-related Gene Expression in Haliotis Midae.*Unpublished PhD dissertation. Stellenbosch Universtiy.

Van der Merwe, M., Franchini, P., Roodt-Wilding, R. (2011).Differential growth-related gene expression in abalone (*Haliotis midae*).*Marine Biotechnology*, *13*, 1-15.

Vandeputte, M., Rossignol, M-N., Pincent, C. (2011). From theory to practice: Empirical evaluation of the assignment power of marker sets for pedigree analysis in fish breeding. *Aquaculture, 314*, 80-86.

Van Tassel, C.P., Smith, T.P.L., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., Sonstegard, T.S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods, 5*, 247-252.

Varshney, R.K., Chabane, K., Hendre, P.S., Aggarwal, R.K., Graner, A. (2007). Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys.*Plant Science, 173*, 638-649.

Varshney, R.K., Nayak, S.N., May, G.D., Jackson, S.A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding.*Trends in Biotechnology, 27*, 522-530.

Vera, M., Alvarez-Dios, J.A., Millán, A., Pardo, B.G., Bouza, C., Hermida, M., Fernández, C., de la Herráan, R., Molina-Luzón, M.J., Martínez, P. (2011).Validation of single nucleotide polymorphism (SNP) markers from an immune Expressed Sequence Tag (EST) turbot, *Scophthalmus maximus*, database.*Aquaculture, 313*, 31-41.

Vera, M., Pardo, B.G., Pino-Querido, A., Alverdez-Dios, J.A., Fuentes, J., Martìnez, P. (2010).Characterization of single-nucleotide polymorphism markers in the Mediterranean mussel, *Mytilus galloprovincialis.Aquaculture Research, 41*, e568-e575.

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 and Evolution*, *34*, 275-305.

Vrijenhoek, R. (1994). Genetic diversity and fitness in small populations. In: V.T. Loesheke (Ed.), *Conservation Genetics* (pp. 37-53). Boston: Birkhauser.

Von der Heyden, S., Prochazka, K., Bowie, R.C. (2008). Significant population structure and asymmetric gene flow patterns amidst expanding populations of *Clinus cottoides* (Perciformes, Clinidae): application of molecular data to marine conservation planning in South Africa. *Molecular Ecology*, 17, 4812-4826.

Wallace, R.B., Shaffer, J., Murphy, R.F., Bonner, J., Hirose, T., Itakura, K. (1979). Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Research, 6*, 3543-3557.

Wang, P.J.Z., Lindsay, B.G., Leebens-Mack, J., Cui, L., Wall, K., Miller, W.C., de Pamphilis, C.W. (2004).EST clustering error evaluation and correction.*Bioinformatics, 20*, 2973-2984.

Wang, S., Peatman, E., Abernathy, J., Waldbieser, G., Lindguist, E., Richardson, P., Lucas, S., Wang, M., Li, P., Thimmapuram, J., Liu, L., Vullaganti, D., Kucuktas, H., Murdock, C., Small, B.C., Wilson, M., Liu, H., Jiang, Y., Lee, Y., Chen, F., Lu, J., Wang, W., Xu, P., Somridhivej, B., Baoprasertkul, P., Quilang, J., Sha, Z., Bao, B., Wang, Y., Wang, Q., Takano, T., Nandi, S., Liu, S., Wong, L., Kaltenboeck, L., Quiniou, S., Bengten, E., Miller, N., Trant, J., Rokhsar, D., Liu, Z., Catfish Genome Consortium. (2010). Assembly of 500 000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies. *Genome Biology, 11*, R8.

Wang, S., Sha, Z., Sonstegard, T.S., Liu, H., Xu, P., Somridhivej, B., Peatman, E., Kucuta, H., Liu, Z. (2008).Quality assesment parameters for EST-derived SNPs for Catfish.*BMC Genomics, 9*, 450.

Ward, R. (2006). The importance of identifying spatial population structure in restocking and stock enhancement programmes. *Fisheries Research, 80*, 9-18.

Weinberg, W. (1908).*On the demonstration of heredity in man*. Englewood Cliffs, NJ.: Prentice Hall.

Weir, B.S., Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure.*Evolution*, *38,* 1358-1370.

Weir, B.S., Anderson, A.D., Hepler, A.B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics, 7*, 771-780.

Wielgoss, S., Barrick, J.E., Tenaillon, O., Cruveiller, S., Chane-woon-ming, B., Médigue, C., Lenski, R.E., Schneider, D. (2011). Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3: Genes/Genomes/Genetics, 1*, 183-186.

Wirgin, I., Kovach, A.I., Roy, L.M., Roy, N.K., Waldman, J., Berlinsky, D.L. (2007). Stock identification of Atlantic cod in U.S. waters using microsatellite and single nucleotide polymorphism DNA analyses. *Transactions of the American Fisheries Society, 136*, 375-391.

Wood, A.D., Buxton, C.D. (1996). Aspects of the biology of the abalone *Haliotis midae* (Linne, 1758) on the east coast of South Africa. 2. Reproduction. *South African Journal of Marine Science, 17*, 69-78.

Wright, J. (1993). DNA fingerprinting in fishes. In: P.W. Hochachka, T. Mommsen (Eds.), *Biochemistry and Molecular Biology of Fishes.* (pp. 58-91). Amsterdam: Elsevier.

Wynne, J. W., O'Sullivan, M. G., Cook, M. T., Stone, G., Nowak, B. F., Lovell, D. R., Elliott, N. G. (2008). Transcriptome analyses of amoebic gill disease-affected Atlantic salmon (*Salmo salar*) tissues reveal localized host gene suppression. *Marine Biotechnology, 10*, 388-403.

Yap, E. (2001). *Frontiers in human genetics: diseases and technologies.* London: World Scientific Publishing Co. Pte. Ltd.

You, F.M., Huo, N., Gu, Y.Q., Luo, M.C., Ma, Y., Hane, D., Lazo, D.R., Dvorak, J., Anderson, O.D. (2008). BatchPrimer3: A high throughput web application for PCR and sequencing primer. *BMC Bioinformatics*, *9*, 253.

Yu, Z., Guo, X. (2006). Identification and mapping of disease-resistance QTLs in the eastern oyster, *Crassostrea virginica* Gmelin.*Aquaculture, 254*, 160-170.

Zeng, S., Gong, Z. (2002). Expressed sequence tag analysis of expression profiles of zebrafish testis and ovary. *Gene, 294*, 45-53.

Zerbino, D.R., Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*, 821-829.

Zhan, A., Hu, J., Hu, X., Hui, M., Wang, M., Peng, W., Huang, X., Wang, S., Lu, W., Sun, C., Bao, Z. (2009). Construction of microsatellite-based linkage maps and identification of size-related quantitative trait loci for Zhikong scallop (*Chlamys farreri*). *Animal Genetics, 40*, 821-831.

Zhan, X., Fan, F., You, W., Yu, J., Ke, C. (2011).Construction of an integrated map of *Haliotis diversicolor* using microsatellite markers.*Marine Biotechnology*,1-8,(DOI:10.1007/s10126-011-9390-7).

Zhang, L., Guo, X. (2010). Development and validation of single nucleotide polymorphism markers in the eastern oyster *Crassostrea virginica* Gmelin by mining ESTs and resequencing.*Aquaculture, 302*, 124-129.

Zhang, Z., Zhan, A., Liu, X. (2010). A panel of genic single nucleotide polymorphism (SNP) markers for the Pacific abalone, *Haliotis discus hannai.Conservation Genetic Resources, 2*, 133-135.

Appendices

Appendix A

Composition of reagents:

2% (w/v) Agarose Gel:

- 3g of agarose for every 150ml 1X TBE
- 7µl of Ethidium Bromide

Cresol loading dye:

- 0.02% (w/v) Cresol Red
- 0.35% (w/v) Sucrose

Appendix B

Table S1: Primer information and optimisation conditions.

Contig	Primer Name	Primer Sequence (5'-3')	Amplicon size (bp)	Tm (°C)	Optimisation Tm (°C)	$MgCl_2$ (mM)
30	HmidSNP30.1F	TCTACAATGGCCCAGAAACC	693	60	50 - 65	1.5 - 2.5
	HmidSNP30.1R	CAGCTACACAGGCCACCTTC		61		
	HmidSNP30.2F	TGGCATCAAGAGAGCAACTG	712	60	50 - 65	1.5 - 2.5
	HmidSNP30.2R	GTCAGGTTTGAAGGGTCCTG		60		
149	HmidSNP149.1F	TGATGGTGATGAAACCTGAGA	693	59	55	2
	HmidSNP149.1R	GAACTACAAGGCGGAGGATG		60		
	HmidSNP149.2F	CGATCTTTCTGCTTCTCATC	693	59	54	2
	HmidSNP149.2R	TGACATAGTGCTGGTTGGAG		61		
	HmidSNP149.3F	TGTAGATCCTCCAACCAGCA	539	59	55	2
	HmidSNP149.3R	AACCAACTGCCGCAGCTAT		61		
	HmidSNP149.4F	CGGCAGTTGGTTCATTGAT	536	60	60	2
	HmidSNP149.4R	ATCTGGGGACCACCTACTCA		59		
1030	HmidSNP1030.1F	ACAATGGCTGGTGGTTGACT	693	60	56	2.5
	HmidSNP1030.1R	TGGTTTGCAACAGCACTACG		61		
	HmidSNP1030.2F	CGTTGTCCTTCAGTTTGACG	709	59	54	1.5 - 2.5
	HmidSNP1030.2R	AGCTCTGGCTTCCGAGACTT		61		
1570	HmidSNP1570.1F	CGCAATCCCGAAGTAAAGTT	693	59	53	1.5 - 2.5
	HmidSNP1570.1R	AGGCCACCTTGTTGATGTCT		60		
	HmidSNP1570.2F	CGCTGGGTATGGTCCTACAG	693	61	53	2
	HmidSNP1570.2R	CCGGCCTTGATTCCATTAG		60		

	HmidSNP1570.3F	GAGCAGCCATTTTCACCAAT	693	60	50 - 65	1.5 - 2.5
	HmidSNP1570.3R	ATTGCCGCCCATCAGTAGTA		60		
	HmidSNP1570.4F	TGACTGAATCTGCTGTTTGGT	550	58	53	2
	HmidSNP1570.4R	GCACAAACTACCAGCCCAGA		61		
3112	HmidSNP3112.1F	AGCACTAGTCTTAGCGACACCTC	616	59	58	2
	HmidSNP3112.1R	AGCCCGTATCAACACTTTGC		60		
	HmidSNP3112.2F	CAAAGTGTTGATACGGGCTTA	770	58	50 - 65	1.5 - 2.5
	HmidSNP3112.2R	ACTGGCGACGTGACCAGATA		62		
	HmidSNP3112.3F	TGATCTTGAAAGGGCACTCC	770	60	50 - 65	1.5 - 2.5
	HmidSNP3112.3R	CACCAGGTGTCCTTCCTCTT		59		
	HmidSNP3112.4F	CCGTTCATGTGGCGGTAG	510	61	50 - 65	1.5 - 2.5
	HmidSNP3112.4R	CCTAAAGGACCTGTGTTGCTG		60		
449	HmidSNP449.1F	AAGAAGGCGAAATCCGAGAC	693	61	55	2
	HmidSNP449.1R	AGGTCTGACAGGCAGGAAGA		60		
	HmidSNP449.2F	CCATACGTGTCCACAATGAT	462	60	55	2
	HmidSNP449.2R	AGTGTATGCCTCTCACCTTG		59		
	HmidSNP449.3F	ACCAAGGTGAGAGGCATACA	516	58	50 - 65	1.5 - 2.5
	HmidSNP449.3R	CTACAGCCTCATCTGGCAATC		60		
21867	HmidSNP21867.1F	GTTAAGGGATGGCATCATTG	552	58	50 - 65	1.5 - 2.5
	HmidSNP21867.1R	TCTCTCAAACCTGCAGACAA		57		
	HmidSNP21867.2F	TTGTCTGCAGGTTTGAGAGA	581	57	65	2
	HmidSNP21867.2R	ACACTCAGACATCGTCATGC		57		
210	HmidSNP210F	CTGGTTCACCAGGTTCATCT	106	58	65	2
	HmidSNP210R	TCTCCTGGTTGTCCCATCT		58		
101	HmidSNP101F	AGCTTGCAGTTCCTCCTTC	256	57	60	2
	HmidSNP101R	GTCCAAGGTGAAGGACCTC		57		

17550	HmidSNP17550.1F	CCTTGATCACGTGGCTAGTT	691	58	55	2
	HmidSNP17550.1R	TGACCTGAATCACTCTTGGA		57		
	HmidSNP17550.2F	TCCAAGAGTGATTCAGGTCA	636	57	50 - 65	1.5 - 2.5
	HmidSNP17550.2R	TGATCCAGGTCATGAGAGAA		57		
	HmidSNP17550.3F	CCTTGCCGATGTATTCTCTC	642	57	53	2
	HmidSNP17550.3R	CCAAAGTTGCAAACATAGAGC		58		
	HmidSNP17550.4F	GTTTGCAACTTTGGTGACAA	493	57	50 - 65	1.5 - 2.5
	HmidSNP17550.4R	ATACCAGGGTGCTGTGGAT		58		
	HmidSNP17550.5F	CTTCCATGTTTTTGTGCTGT	527	56	50 - 65	1.5 - 2.5
	HmidSNP17550.5R	CATCAGCCATCTCTGTTTTG		57		
20089	HmidSNP20089F	CAATCTGTGCCATGATTCTC	246	57	50	2
	HmidSNP20089R	CAGAACAGTTGAACGATGATG		57		
237	HmidSNP237.1F	AAACTCTGGAAGAACCAGCA	675	58	54	1.5
	HmidSNP237.1R	AATGGAAGATCCCTTCATCA		57		
	HmidSNP237.2F	CTCTGATGAAGGGATCTTCC	680	56	50 - 65	1.5 - 2.5
	HmidSNP237.2R	CATTTCCACATCCAAACAGA		57		
16644	HmidSNP16644.1F	GACTGTGTTTGGACCTTTGA	708	56	50 - 65	1.5 - 2.5
	HmidSNP16644.1R	AATGAAGCAAACCAGTCTCC		57		
	HmidSNP16644.2F	GGAGACTGGTTTGCTTCATT	492	57	50 - 65	1.5 - 2.5
	HmidSNP16644.2R	ATGTCCTTCAATTTGCTTGG		58		
	HmidSNP16644.3F	GACCAAGCAAATTGAAGGAC	520	57	55	2
	HmidSNP16644.3R	CGTGTTGTAAACCACTGCAC		58		
14549	HmidSNP14549F	ATGGAAAGGTCACATGAGGT	447	57	50 - 65	1.5 - 2.5
	HmidSNP14549R	ACTCATCGAGCCTACTTTGG		57		
5470	HmidSNP5470F	CTCCAGGCTTCTGTTTAAGG	714	57	50 - 65	1.5 - 2.5
	HmidSNP5470R	CATTGCAGCTGTAGTCAAGG		57		

13064	HmidSNP13064F	CGGAGAAGGAAGATCAAAAG	323	57	51	1.5
	HmidSNP13064R	AAGGTTTGGGTAGGAAGGAG		57		
379	HmidSNP379F	GAGACCAGGTTCACCGATAC	696	57	58	2
	HmidSNP379R	TCTCTTGGCATCAACAGGTA		57		
1827	HmidSNP1827F	CACCACCAGCAACAATACC	256	57	50 - 65	2
	HmidSNP1827R	GTGTACTGCGGAAAGAAAGC		58		
1095	HmidSNP1095.1F	CTTGATGCATTCCTGTGAAC	503	57	50 - 65	1.5 - 2.5
	HmidSNP1095.1R	AGAACCTGCTGAAGACCAAC		57		
	HmidSNP1095.2F	ATGTTGGTCTTCAGCAGGTT	578	57	50 - 65	1.5 - 2.5
	HmidSNP1095.2R	GAGAACATCCAATCCATCAAG		57		
	HmidSNP1095.3F	GCTTGATGGATTGGATGTTC	646	58	50 - 65	1.5 - 2.5
	HmidSNP1095.3R	AAGATGGCGACTCATCTGTT		57		
1881	HmidSNP1881F	ATCATGGCATCCATTGAGAG	205	58	54	2.5
	HmidSNP1881R	AAGGAACCTTTGCCCATT		57		
4691	HmidSNP4691F	GGCTTGTCTGACACTGAAGA	466	56	55	2
	HmidSNP4691R	TCCACTGTCTCTCGGTATTG		56		
3516	HmidSNP3516.1F	AGCAACCAAATCAAGTGGAG	396	58	50 - 65	1.5 - 2.5
	HmidSNP3516.1R	ACACCATAGGCACACATCAA		57		
	HmidSNP3516.2F	TGTGTGCCTATGGTGTAACAG	716	57	50 - 65	1.5 - 2.5
	HmidSNP3516.2R	GGCATCTCTCATCAGCTTCT		57		
	HmidSNP3516.3F	GGGTAATGGATTCAACAGTGA	691	57	50 - 65	1.5 - 2.5
	HmidSNP3516.3R	TCAGCATCAGAAGAAAGACG		57		
18559	HmidSNP18559F	TCTACTAAGCAGCCTGAGGAA	676	57	50 - 65	1.5 - 2.5
	HmidSNP18559R	CAGAGTACGGCATTGAACTG		57		
216	HmidSNP216F	GGGGTAAACCTCAAGGAACT	257	57	60	2.5
	HmidSNP216R	TCCCTTGTTGGGTCTGTACT		57		

783	HmidSNP783F	CCTCGGTTTGTTCTGTTCTT	696	57	50 - 65	1.5 - 2.5
	HmidSNP783R	GAGACTTCTCGCCATTTTGT		57		
21833	HmidSNP21833F	AATCTTGGGGTCAAGAAACA	690	57	50 - 65	1.5 - 2.5
	HmidSNP21833R	TCATTCACACAACCTCTCCA		58		
22644	HmidSNP22644.1F	CCGCTTTCTATATCCACACC	705	57	50 - 65	1.5 - 2.5
	HmidSNP22644.1R	GACAGAAACCAAGGGAACAG		57		
	HmidSNP22644.2F	ATGATATCCAAGGGGCAAG	685	57	50 - 65	1.5 - 2.5
	HmidSNP22644.2R	ATTACAGTGGCCTTGAGGAA		57		
2962	HmidSNP2962F	ACACAGGGAACTTGTGCAG	547	57	50 - 65	1.5 - 2.5
	HmidSNP2962R	GCCTAGCAGAAGAGAGGAAA		57		
17023	HmidSNP17023F	CACCATCTGGGGTTCTGTA	166	57	54	1.5
	HmidSNP17023R	GATCATTGTCACACGGAAAG		56		
1659	HmidSNP1659F	CGAATTCCTTTGGTTCAGAG	138	57	55	2
	HmidSNP1659R	AGTGTATTTGACACGCCACA		58		
1949	HmidSNP1949F	GGCGGTGTTCACAAACTT	368	57	54	2
	HmidSNP1949R	GGAAGGTAGCCTGTATTGGA		57		
5553	HmidSNP5553.1F	AGGGACAGTCTGATTTGACC	707	57	50 - 65	1.5 - 2.5
	HmidSNP5553.1R	GTTGCCGTATGACCAGTACA		57		
	HmidSNP5553.2F	AGTGCCAGGATAACTGGAAT	696	56	51	2
	HmidSNP5553.2R	GATCAACTTCACCGAGAACA		56		
516	HmidSNP516F	GGAAAGACTCAACAGCAGGA	261	58	50 - 65	1.5 - 2.5
	HmidSNP516R	CAGAGCAGTGATACCGAGATT		57		
1264	HmidSNP1264F	AAACGCTGGAGAGGTAAAAG	184	56	50 - 65	2
	HmidSNP1264R	CCTACTTCAACGCTCCAATC		57		
17935	HmidSNP17935F	TTGGAAGTATGGAGGCACTT	560	57	50 - 65	2
	HmidSNP17935R	TGTTTGTAGAACACGGACCA		58		

7544	HmidSNP7544F	ATGGTTAGAAGAGGCGGATT	152	58	52	2
	HmidSNP7544R	CGGAGGGTTTAGGTTGTTT		57		
1718	HmidSNP1718F	CCAAATAAGGTTGGGTTGAT	189	56	53	2
	HmidSNP1718R	ACTAAAATGCTCCGTCAAGC		57		
393	HmidSNP393F	CGATCTCTCGTAACCGTCTT	614	57	50 - 65	2
	HmidSNP393R	GGATGGAGTTGTAAGGCTCA		58		
2050	HmidSNP2050F	CCTGAAGGAAGAGCACAAAG	679	58	50 - 65	1.5 - 2.5
	HmidSNP2050R	GTTGGGTACCGGTGAGTTAC		57		
15845	HmidSNP15845F	ACAGATGCTGAGTGACAATGA	688	57	50 - 65	1.5 - 2.5
	HmidSNP15845R	GGCATCAAAGCAACAAATG		58		
1833	HmidSNP1833F	AGATGACTGGGGCATAGGTA	552	57	52	2
	HmidSNP1833R	TTACAGGAAAGGAAGATGCTG		57		
1527	HmidSNP1527F	AACAACACGCTACCTCTTCC	711	57	50 - 65	1.5 - 2.5
	HmidSNP1527R	AAGTTTGGCAACCAGTACG		56		
54	HmidSNP54F	CCCAGACTCCTCAATCTGAC	695	57	50 - 65	1.5 - 2.5
	HmidSNP54R	AAGCTTTGGAACTGGTTGAC		57		
569	HmidSNP569F	TTCATGCTATTCCACTACGC	700	56	50 - 65	1.5 - 2.5
	HmidSNP569R	TCCTCTGTCACACCCAAAC		57		
214	HmidSNP214F	AAGCCGATAGCACCAATATC	694	57	53	2
	HmidSNP214R	GCAACGACGGTAAATCTCTC		57		
11848	HmidSNP11848.1F	CAAGATGTCTGCTTCTCTGC	653	56	50 - 65	1.5 - 2.5
	HmidSNP11848.1R	TCTGCTCTGCTGTCTTCATC		57		
	HmidSNP11848.2F	GCGCTGTATGTGGCTCAG	664	59	50 - 65	1.5 - 2.5
	HmidSNP11848.2R	AGTTATTGGCACGCTCGAT		58		
	HmidSNP11848.3F	ATGTCCCAGGATATCGACAA	542	58	50 - 65	1.5 - 2.5
	HmidSNP11848.3R	TAGGAAATGGGATCACCAGA		58		

	HmidSNP11848.4F	AACTACACAACGCCTCCAGT	355	57	53	2
	HmidSNP11848.4R	ATACCAACAACTTCACCCATC		56		
146	HmidSNP146.1F	GCCGAAGAGACGAGAGTAGA	680	57	55	2
	HmidSNP146.1R	GTCGTAGAAGCCGAAGTAGC		57		
	HmidSNP146.2F	CATCTCTTGCGTAGGAATCA	626	56	52	2
	HmidSNP146.2R	ACGGTAAGGTCAAAGTTCGT		56		
	HmidSNP146.3F	CACGAACTTTGACCTTACCG	460	58	52	2
	HmidSNP146.3R	ATACGGCCCATTTTCTTTGA		59		
5153	HmidSNP5153.1F	CTGGTTCCAAACATCCTGAC	503	58	58	2
	HmidSNP5153.1R	AGACCCTGGAGCGTACAGA		58		
	HmidSNP5153.2F	ACTGCCTGGTTCAGCTTG	679	57	50 - 65	1.5 - 2.5
	HmidSNP5153.2R	CCGAGATGGGAGCTTATGT		58		
342	HmidSNP342.1F	GGTAGCAAGCCCAAACTACA	545	58	55	2
	HmidSNP342.1R	GTGCTGCTGGACAGAGTTC		57		
	HmidSNP342.2F	CACGGAGGCCAGTATAGAAA	728	58	54	2
	HmidSNP342.2R	ACACCATTCTGTCAATGTCG		57		
	HmidSNP342.3F	CAGAATGGTGTCTGAAGCTG	344	57	53	2
	HmidSNP342.3R	AGTCCATCTCCTCCACAGAG		57		
5879	HmidSNP5879.1F	CGGTGCAAAGTTCATACAAG	395	57	50 - 65	1.5 - 2.5
	HmidSNP5879.1R	ACGACATCTGCCTTCTCCT		57		
	HmidSNP5879.2F	AGGAGGAGAAGGCAGATGT	551	56	50 - 65	1.5 - 2.5
	HmidSNP5879.2R	ACATACACACAGAGCGCAAC		57		
11960	HmidSNP11960.1F	GTTGTTCCTGAAAAGGAGGA	548	57	50 - 65	1.5 - 2.5
	HmidSNP11960.1R	TCAAAGTGGGTGAGTGGTAA		57		
	HmidSNP11960.2F	TTACCACTCACCCACTTTGA	672	57	50 - 65	1.5 - 2.5
	HmidSNP11960.2R	GTTATTTTTGTAGCCCTTCTCG		57		

1834	HmidSNP1834F	ACGGCTCTCTGGACCTTT	627	57	55	2
	HmidSNP1834R	GCTTGAGTTCTCGATCTACCC		58		
7614	HmidSNP7614F	TTGATCTTGTCTCCCATGC	630	57	50 - 65	1.5 - 2.5
	HmidSNP7614R	CCTGGATGTCCTCATTGTC		56		
8704	HmidSNP8704F	GGCACACACTCGTACACACT	688	57	50 - 65	1.5 - 2.5
	HmidSNP8704R	ATGATCTCTCCGACCCAAT		57		
153	HmidSNP153F	TGTTCAAGTACGACACAGTCC	559	56	50 - 65	1.5 - 2.5
	HmidSNP153R	CCGGTCAACTTTCCGTTA		57		
3566	HmidSNP3566F	CTCTCTGTTCCTCCATCGAC	677	57	50 - 65	1.5 - 2.5
	HmidSNP3566R	GTTGGAAAGCACGATTGTTT		58		
4384	HmidSNP4384F	AAGCCTACAACAAACGAAGG	730	57	50 - 65	1.5 - 2.5
	HmidSNP4384R	CTCCCCAAGCATCAATCTAC		57		
2187	HmidSNP2187F	AAGGGTGAGATGAGGTGATG	654	57	50 - 65	1.5 - 2.5
	HmidSNP2187R	GAGAGAGCAGCCGGTTATT		57		
984	HmidSNP984F	ACCACAAACAAGTTCCCTTC	691	57	55	1.5 - 2.5
	HmidSNP984R	AGCCCAACCAATGACAATAC		57		
23086	HmidSNP23086F	CGTAAGGATGAAGACGGTTC	699	57	50 - 65	1.5 - 2.5
	HmidSNP23086R	TGTCCATCAGCTTCTCACAT		57		

Appendix C

Table S2: Illumina GoldenGate genotyping assay information.

Row	Locus_Name	Illumicode Name	#No Calls	#Calls	Call Freq	A/A Freq	A/B Freq	B/B Freq	Minor Freq	Gentrain Score
1	SNP4691_183	10	38	331	0.897	0.003	0.127	0.87	0.066	0.7764
2	SNP1718_109	23	39	330	0.894	0	0	1	0	0.6127
3	SNP149.4_75	33	41	328	0.889	0	0.073	0.927	0.037	0.641
4	HmLCS55T318G_T	51	39	330	0.894	0	0	1	0	0.5587
5	SNP101_113	54	41	328	0.889	0.976	0.021	0.003	0.014	0.4921
6	3D10_1	56	40	329	0.892	0.04	0.304	0.657	0.191	0.6997
7	SNP972_1055	67	39	330	0.894	0.218	0.018	0.764	0.227	0.4675
8	SNP1949_235	101	40	329	0.892	0.647	0.35	0.003	0.178	0.7562
9	SNP1833_160	102	369	0	0	0	0	0	0	0.2299
10	SNP1001_388	113	43	326	0.883	0.015	0.328	0.656	0.179	0.794
11	SNP214_86	119	39	330	0.894	0.003	0.021	0.976	0.014	0.7326
12	SNP20648_3041	130	40	329	0.892	0.003	0.17	0.827	0.088	0.7289
13	SNP149.1_374	153	45	324	0.878	0.006	0.065	0.929	0.039	0.6924
14	SNP300_1828	154	43	326	0.883	0	0.141	0.859	0.071	0.7494
15	SNP13865_165	162	39	330	0.894	0.042	0.306	0.652	0.195	0.786
16	SNP149.2_165	178	40	329	0.892	0.331	0.505	0.164	0.416	0.8041
17	SNP2091_264	186	39	330	0.894	0.064	0.373	0.564	0.25	0.799
18	HdSNPc106_688C_T	209	48	321	0.87	0.003	0.897	0.1	0.452	0.3621
19	HmLCS5M193T_A	215	369	0	0	0	0	0	0	0.1884

20	SNP67_164	233	44	325	0.881	0	0.123	0.877	0.062	0.3456
21	HdSNPc148_820T_C	234	39	330	0.894	0.997	0.003	0	0.002	0.7926
22	SNP101_201	265	39	330	0.894	0	0.003	0.997	0.002	0.8149
23	SNP210_266	300	39	330	0.894	0	0.106	0.894	0.053	0.7128
24	SNP149.1_106	331	39	330	0.894	0	0	1	0	0.8084
25	SNP3129_923	398	38	331	0.897	0.592	0.384	0.024	0.216	0.7701
26	SNP214_434	430	41	328	0.889	1	0	0	0	0.8086
27	SNP449.2_110	442	39	330	0.894	0.191	0.406	0.403	0.394	0.7927
28	SNP17550.1_463	480	38	331	0.897	0	0	1	0	0.5132
29	SNP149.4_341	515	38	331	0.897	0.091	0.32	0.589	0.251	0.7793
30	SNP146.2_132	559	38	331	0.897	0	0.266	0.734	0.133	0.752
31	2H92	585	44	325	0.881	0	0	1	0	0.6015
32	SNP146.3_123	607	43	326	0.883	1	0	0	0	0.796
33	SNP229_2772	614	43	326	0.883	0.936	0.064	0	0.032	0.697
34	SNP342.2_537	619	54	315	0.854	0.879	0.111	0.01	0.065	0.3571
35	SNP1834_76	629	38	331	0.897	0.997	0.003	0	0.002	0.8036
36	SNP140_2421	634	43	326	0.883	0.776	0.209	0.015	0.12	0.7558
37	HmSNPc4_815C_T	651	369	0	0	0	0	0	0	0.1731
38	SNP17550.3_221	656	42	327	0.886	0	0.031	0.969	0.015	0.4364
39	SNP1834_464	674	41	328	0.889	0.003	0.957	0.04	0.482	0.3756
40	HaSNPdw500_207C_T	682	369	0	0	0	0	0	0	0.2406
41	SNP48_322	718	369	0	0	0	0	0	0	0.0834
42	3B4_2	720	39	330	0.894	0.2	0.676	0.124	0.462	0.7056
43	SNP449.2_443	732	39	330	0.894	0	0.024	0.976	0.012	0.7446
44	3B4_7	742	42	327	0.886	0.422	0.502	0.076	0.327	0.4918
45	HmLCS5M479C_T	747	369	0	0	0	0	0	0	0.1973

46	HmRS36T262T_C	753	39	330	0.894	0.012	0.061	0.927	0.042	0.569
47	SNP5837_204	756	369	0	0	0	0	0	0	0.0951
48	SNP17550.3_555	764	38	331	0.897	0.861	0.139	0	0.069	0.8158

Appendix D

TableS3: SNP pairs that are in linkage disequilibrium.

Locus pair		df	Probability
S101_113	S149.2_165	12	Highly sign.
RS36	3B4_7	8	Highly sign.
S149.2_165	3B4_7	14	Highly sign.
3B4_2	3B4_7	10	Highly sign.
S449.2_110	S972_1055	8	Highly sign.
3B4_7	S972_1055	12	Highly sign.
S149.2_165	S300_1828	8	Highly sign.
3B4_7	S140_2421	12	Highly sign.

df: degree of freedom