

Medium-throughput SNP genotyping and linkage mapping in *Haliotis midae*

by
Jana du Plessis

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



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

December 2012

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

December 2012

Copyright © 2012

Stellenbosch University

All rights reserved

Abstract

Haliotis midae (locally also known as perlemoen) is the largest of five endemic species found along the coast of South Africa. It is the only species with commercial value contributing to the exploitation of these animals. Due to declines of natural stocks, farming practices were established during the early 1990s in order to supply the international demand. To facilitate efficient breeding methods and ensure the sustainability of these commercial populations, genetic management, which can be accomplished with the use of molecular markers such as single nucleotide polymorphisms (SNPs), is necessary.

Single nucleotide polymorphisms have become the markers of choice in various applications in aquaculture genetics due to their abundance in genomes, reduction in developmental costs and increased throughput of genotyping assays. Identification of SNPs in non-model species such as *H. midae* can be achieved by *in silico* approaches. *In silico* methods are suitable for *de novo* SNP identification and are both cost- and time-efficient. It is based on the analysis of multiple alignments where mismatches may be reported as candidate SNPs. Various medium-throughput genotyping methods are available to confirm putative SNPs, but the ideal method depends on factors such as cost, accuracy and multiplexing capacity.

Although SNP markers can have various applications within the aquaculture environment the focus for this current study was saturating the linkage map of *H. midae* with additional markers. This would assist in the identification of quantitative trait loci associated with economically important traits, which in turn could ultimately be employed for marker-assisted selection and improved molecular breeding programs.

In order to identify *in silico* SNPs, sequenced transcriptome data from a previous study was used and subjected to a series of criteria: minor allele frequency 10%, minimum coverage 80, 60 bp flanking regions. Selected loci were genotyped using a 192-plex assay with the Illumina GoldenGate genotyping assay with the VeraCode technology on the BeadXpress platform, in individuals from six mapping families. A conversion rate of 69.35% and global success rate of 76.34% was achieved. Polymorphic loci were subjected to linkage analysis using JoinMap[®] v.4.1 to create

sex-average and sex-specific maps and to saturate the current linkage map for *H. midae*. Along with previously developed markers, 54% of the newly developed SNPs could be successfully incorporated into the linkage map of *H. midae*. A total of 18 linkage groups were observed with an average marker spacing of 6.9 cM and genome coverage of 79.1%.

Bioinformatic analyses and setting stringent criteria to identify SNPs from sequenced transcriptomic data proved to be an efficient way for SNP discovery in the current study. Genotyping of the identified loci with the GoldenGate genotyping assay demonstrated a high success rate; providing a genotyping assay adequate for species with little genomic information. The linkage map created in this study illustrated the utility of SNP markers in conjunction with microsatellite markers for linkage map construction and the adequate marker spacing obtained provides a step closer to quantitative trait loci mapping in this species.

Opsomming

Haliotis midae (plaaslik ook bekend as perlemoen) is die grootste van vyf inheemse spesies wat langs die kus van Suid-Afrika aangetref word. Dit is die enigste spesie van kommersiële waarde wat bydraend is tot die uitbuiting van hierdie diere. As gevolg van die afname in hierdie natuurlike hulpbron het boerdery praktyke gedurende die vroeë 1990's ontstaan om in die internasionale aanvraag te voorsien. Ten einde doeltreffende teelmetodes te beoefen en die volhoubaarheid van hierdie kommersiële populasies te verseker is genetiese bestuur, wat bewerkstellig kan word deur die gebruik van molekulêre merkers soos enkel nukleotied polimorfismes (ENPs), baie belangrik.

Enkel nukleotied polimorfismes is gewilde merkers in verskeie toepassings in akwakultuur genetica as gevolg van hul oorfloed in genome, verlaagde ontwikkelingskoste en verhoogde deurset van ENP-genotiperingsstoets. Identifisering van ENPs in nie-model spesies soos *H. midae* kan uitgevoer word deur *in silico* benaderings te gebruik wat geskik is vir *de novo* ENP identifisering en ook tyd- en koste-effektief is. Dit word gebaseer op die analise van veelvuldige inlynstellings waar nukleotiedes wat nie ooreenstem nie as kandidaat ENPs gerapporteer kan word. Om kandidaat ENPs te bevestig, kan verskeie medium-deurset genotiperingsmetodes uitgevoer word, maar die ideale metode word bepaal deur faktore soos koste, akkuraatheid en multipleks kapasiteit.

Alhoewel ENP merkers in verskeie toepassing binne die akwakultuur omgewing gebruik kan word was die fokus van die huidige studie om die koppelingskaart van *H. midae* te versadig. Dit sal bydrae tot die identifisering van kwantitatiewe eienskap lokusse wat gekoppel kan word aan ekonomies belangrike eienskappe wat dan op die beurt weer vir merkerbemiddelde seleksie gebruik kan word en uiteindelik ten opsigte van die verbetering van molekulêre teelprogramme aangewend kan word.

Ten einde *in silico* ENPs te identifiseer is transkriptoombdata van 'n vorige studie gebruik en onderwerp aan 'n reeks kriteria: geringste alleelfrekwensie 10%, minimum dekking 80, 60 bp gebiede weerskante van polimorfisme. Geïdentifiseerde lokus-genotipering is met behulp van 'n 192-pleks toets uitgevoer met die Illumina GoldenGate genotiperingsstoets met die VeraCode tegnologie op die BeadXpress-

platform, in individue afkomsitg vanaf ses karteringsfamilies. 'n Omskakelingskoers van 69.35% en 'n algehele sukseskoers van 76.34% is bereik. Polimorfiese lokusse is onderwerp aan koppelings-analise met behulp van JoinMap[®] v.4.1 om geslags-gemiddelde en geslags-spesifieke kaarte te skep asook om die kaart wat beskikbaar is vir *H. midae* te versadig. Saam met voorheen ontwikkelde merkers is 54% van die nuut ontwikkelde ENPs suksesvol opgeneem in die kaart van *H. midae*. 'n Totaal van 18 koppelingsgroepe is verkry met 'n gemiddelde merker-spasiëring van 6.9 cM en 'n genoomdekking van 79.1%.

Die gebruik van bioinformatiese analises en streng kriteria om ENPs vanaf transkriptoomdata te identifiseer blyk doeltreffend te wees in hierdie studie. Genotipering van die geïdentifiseerde lokusse met die GoldenGate genotiperingstoets dui op 'n hoë suksesyfer en verskaf 'n voldoende genotiperingstoets aan spesies met min genomiese inligting. Die koppelingskaart in hierdie studie het geïllustreer dat die ENP merkers suksesvol saam met mikrosatelliet merkers gebruik kan word vir koppelingskaart konstruksie en dat die voldoende merker-spasiëring verkry 'n stap nader aan kwantitatiewe eienskap lokus kartering in hierdie spesie bied.

Acknowledgements

I would like to thank my supervisor Prof Rouvay Roodt-Wilding for providing me with the opportunity to join the MARG team and for offering guidance and support throughout my time there. I would also like to thank Dr Aletta Bester-van der Merwe for inputs and assistance with my project. To both of them, thank you for your contribution and dedication towards the finalisation of this thesis.

To all my colleagues at MARG, thank you for always answering my questions and helping me when I needed it. I would especially like to thank Jessica, our linkage mapping expert, who had the utmost patience and time for me when I needed linkage mapping advice. I would also like to acknowledge the National Research Fund for providing me with funding during my MSc studies, and also the farms for providing me with animals to conduct my research on.

To all my friends and loved ones, thank you for the constant support, keeping me positive and always being just a phone call away when I needed company. To my family, especially my mom and dad, thank you for your love, encouragement and interest in my work even though you didn't really understand the science behind it. Thank you for teaching me what hard work and dedication is, for without that I would not have been able to accomplish what I have so far.

Publications and Conference Presentations

Part of Chapter Two and Chapter Three on marker development and genotyping has been submitted for publication and was presented at a national conference.

Blaauw, S., Du Plessis, J., Bester-van der Merwe, A.E., Roodt-Wilding, R. SNP marker development and medium-throughput genotyping in the South African abalone, *Haliotis midae*. **Submitted *Aquaculture***

Du Plessis, J.*, Bester-van der Merwe, A.E., Roodt-Wilding, R. Marker development and high-throughput SNP genotyping in *Haliotis midae*. **Poster presented at the South African Genetics Society Conference, September 2012, Stellenbosch**

Contributions: Provided data for poster and publication. Wrote parts of the materials and methods and results sections of the publication. Presented poster at conference.

The work done in Chapter Four on linkage mapping has been accepted for publication and was presented at one national and one international conference.

Vervalle, J., Hepple, J., Jansen, S., Du Plessis, J., Wang, P., Rhode, C., Roodt-Wilding, R. Integrated linkage map of *Haliotis midae* Linnaeus based on microsatellites and SNPs. **In press *Journal of Shellfish Research***

Vervalle, J.*, Jansen, S., Du Plessis, J., Wang, P., Rhode, C., Roodt-Wilding, R. Integrated linkage map of the South African abalone (*Haliotis midae*) based on microsatellites and SNPs. **Poster presented at the South African Genetics Society Conference, September 2012, Stellenbosch**

Vervalle, J.*, Jansen, S., Du Plessis, J., Wang, P., Rhode, C., Roodt-Wilding, R. Integrated linkage map of the South African abalone (*Haliotis midae*) based on microsatellites and SNPs. **Oral presentation at the International Abalone Society Conference, May 2012, Hobart, Tasmania**

Contributions: Provided data for poster, oral presentation and publication.

Table of contents

Declaration	ii
Abstract	iii
Opsomming	v
Acknowledgements	vii
Publications	viii
Table of contents	ix
List of figures	xiii
List of tables	xv
List of abbreviations	xvii
Chapter One - Literature Review	1
1. Abalone in general	2
1.1 Habitat and distribution	2
1.1.1 <i>South African abalone</i>	2
1.2 Taxonomy	2
1.3 Anatomy	3
1.4 Life cycle	4
1.5 Feeding	4
2. Abalone culture	5
2.1 Abalone farming in general	5
2.2 Abalone farming in South Africa	6
3. Application of genetics in aquaculture	8
4. Molecular markers	10
4.1 Microsatellite markers	10

4.2 Single nucleotide polymorphisms	11
4.2.1 <i>SNP discovery methods</i>	12
4.2.2 <i>Medium-throughput genotyping methods</i>	14
4.2.2.1 PCR-free methods	14
4.2.2.2 Pyrosequencing	14
4.2.2.3 Mass spectrometry based genotyping assay	15
4.2.2.4 Bead-based arrays	15
4.2.2.4.1 <i>Illumina GoldenGate genotyping assay with the VeraCode technology on the BeadXpress platform</i>	16
5. Linkage mapping	18
5.1 Application of linkage mapping: QTL analysis	20
5.1.1 <i>QTL mapping in aquaculture species</i>	20
5.1.2 <i>Linkage mapping: a precursor for QTL mapping in abalone</i>	21
6. Marker-assisted selection	22
7. Aims and objectives	22
7.1 SNP marker development	22
7.2 Linkage mapping	23
Chapter Two - Marker Development	24
1. Introduction	25
2. Materials and methods	27
2.1 EST construction and sequence assembly	27
2.2 <i>In silico</i> identification of SNPs	27
3. Results	28
3.1 Sequence assembly	28
3.2 Putative SNP discovery	28
4. Discussion	30

Chapter Three - Genotyping of <i>In Silico</i> Developed SNPs	36
1. Introduction	37
2. Materials and methods	38
2.1 Sample collection and DNA extraction	39
2.2 Preparing and selecting families for genotyping	39
2.2.1 <i>PCR multiplex</i>	39
2.2.2 <i>Genotyping of microsatellite markers</i>	40
2.2.3 <i>Selected families for genotyping</i>	40
2.3 SNP genotyping assay	41
2.4 Data analysis	41
3. Results	42
3.1 Parentage analysis	42
3.2 Genotyping performance	42
3.3 Validation and performance of SNPs	42
4. Discussion	44
Chapter Four - Linkage Mapping	49
1. Introduction	50
2. Materials and methods	52
2.1 Mapping families	52
2.2 Genotyping of gene-linked markers	53
2.2.1 <i>EST-derived SNP markers</i>	53
2.2.2 <i>Genotype data</i>	53
2.3 Analysis of linkage between loci	54
2.4 Map integration	55
2.5 Genome coverage	55
2.5.1 <i>Expected genome length</i>	55
2.5.2 <i>Genome coverage</i>	56
3. Results	56

3.1 EST-derived SNP markers	56
3.2 Linkage mapping	56
3.2.1 <i>Linkage map of family D</i>	57
3.2.2 <i>Linkage map of family DS1</i>	60
3.2.3 <i>Integrated linkage map</i>	65
3.2.4 <i>Linkage group one (LG_1) comparison</i>	69
4. Discussion	70
Chapter Five - Conclusions and Future Considerations	79
1. Introduction	80
2. Marker development and validation	80
3. Linkage mapping	82
4. Conclusions and future considerations	84
References	86
Addendums	107

List of figures

Figure 1.1	A map indicating the distribution of the five endemic abalone species found along the South African coastline (Lindberg 1992).	3
Figure 1.2	A demonstration of the life cycle of abalone (Hepple 2010).	5
Figure 1.3	Location of fishing zones (A-D) along the South African coast (Hauck & Sweijd 1999).	7
Figure 1.4	Annual abalone production from 2000 to 2010 (DAFF 2011).	8
Figure 1.5a	A workflow of the VeraCode GoldenGate assay - hybridisation, extension, ligation, amplification. (http://www.illumina.com/technology/veracode_goldengate_assay.ilmn)	17
Figure 1.5b	A workflow of the VeraCode GoldenGate assay - wash, scan, allele calling. (http://www.illumina.com/technology/veracode_goldengate_assay.ilmn)	18
Figure 2.1	Classification of 139 contigs.	30
Figure 3.1	An illustration regarding the effects of introns on genotyping (Wang <i>et al.</i> 2008).	46
Figure 4.1	Maternal (P1), sex-average (POP) and paternal (P2) maps of family D.	60
Figure 4.2	Maternal (P1), sex-average (POP) and paternal (P2) maps of family DS1.	65

- Figure 4.3 Integrated map [* indicates SNP markers developed in current study, # indicates positive SNP controls developed by Blaauw (2012)]. 68
- Figure 4.4 LG_1: A comparison of marker order and marker density of sex-average maps from family D (SNPs only) and DS1 (SNPs and microsatellites) with the integrated map. 70

List of tables

Table 1.1	Commercial abalone species and their countries of origin.	6
Table 1.2	Types of molecular markers, their characteristics and corresponding applications (Adapted from Sunnucks 2000; Liu & Cordes 2004; Schlötterer 2004; Collard <i>et al.</i> 2005).	11
Table 1.3	Previous studies employing the GoldenGate genotyping assay for medium-throughput SNP genotyping.	16
Table 1.4	Linkage maps consisting mainly of SNPs and microsatellites for some aquaculture species (Adapted from Jansen 2012).	19
Table 2.1	Summary of putative <i>in silico</i> SNP discovery in <i>H. midae</i> .	29
Table 2.2	Comparison of next-generation sequencing platforms (adapted from Ekblom & Galindo 2010).	34
Table 3.1	Parentage verified animals used for genotyping (* families included in previous studies).	41
Table 3.2	Genotyped samples.	43
Table 3.3	Summary of successful and unsuccessful genotypes.	44
Table 3.4	Comparison of genotyping assay success for a well-studied species ^a with species of limited genomic information ^b .	47
Table 4.1	Families used for linkage map construction (* families included in previous studies).	53

Table 4.2	JoinMap [®] v.4.1 genotype data format for CP populations (Van Ooijen 2006).	54
Table 4.3	Number of markers per linkage group, length of linkage groups, average marker spacing and largest interval for maternal (P1), sex-average (POP) and paternal (P2) maps of family D.	58
Table 4.4	Number of markers per linkage group, length of linkage groups, average marker spacing and largest interval for maternal (P1), sex-average (POP) and paternal (P2) maps of family DS1.	61
Table 4.5	Number of markers per linkage group, length of linkage groups, average marker spacing and largest interval for integrated map.	66
Table 4.6	Number of markers per map, lengths of LG_1, average marker spacing and largest interval.	69
Table 4.7	Anchor loci informative in four or more families.	73
Table 4.8	Examples of linkage maps constructed for fish and shellfish species.	77

List of abbreviations

%	Percentage
<	Less than
>	Greater than
°C	Degrees Celsius
µl	Microlitre
µm	Micrometre
µM	Micromolar
®	Registered Trademark
3'	Three Prime
A	Adenine
ADT	Assay Design Tool
AFLP	Amplified Fragment Length Polymorphism
A _s	Average Marker Spacing
ASO	Allele-Specific Oligonucleotide
ATP	Adenosine Triphosphate
BLAST	Basic Local Alignment Search Tool
BLASTX	Basic Local Alignment Search Tool (protein search)
bp	Base Pair
C	Cytosine
CA	California
cDNA	Complementary DNA
cM	CentiMorgan
Contig	Contiguous Sequences
CTAB	Cetyltrimethylammonium Bromide
dCAS	Desktop cDNA Annotation System
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide Triphosphate
dsDNA	Double Strand Deoxyribonucleic Acid
EDTA	Ethylenediamine Tetra-acetate
EST	Expressed Sequence Tag
Fam	Family
G	Guanine

GA II	Genome Analyser II
GC	Genome Coverage
gDNA	Genomic Deoxyribonucleic Acid
$G_{e\text{ ave}}$	Average Expected Genome Size
G_{e1}	Genome Size Estimation 1
G_{e2}	Genome Size Estimation 2
G_o	Observed Map Length
GO	Gene Ontology
INT	Integrated
kg	Kilogram
k_i	Number of Markers on Linkage Group i
km	Kilometre
KOG	Eukaryote Clusters of Genes
LD	Linkage Disequilibrium
LG	Linkage Group
LOD	Logarithm of Odds
LSO	Locus-Specific Oligonucleotide
MAF	Minor Allele Frequency
MALDI-TOF	MS Matrix-Associated Laser Desorption Time of Flight Mass Spectrometry
MAS	Marker-Assisted Selection
mf	Mapping Function
$MgCl_2$	Magnesium Chloride
Min	Minutes
m	Metre
mg	Milligram
ml	Millilitre
ML	Maximum Likelihood
mm	Millimetre
mM	Millimolar
mtDNA	Mitochondrial Deoxyribonucleic Acid
ng	Nanogram
NGS	Next-Generation Sequencing
NHLS	National Health Laboratory Services

P	Probability Value
P1	Maternal Linkage Map
P2	Paternal Linkage Map
PCR	Polymerase Chain Reaction
PFAM	Database of Protein Families and Domains
pH	Concentration of Hydrogen Ions in a Solution
POP	Sex-average Linkage Map
PSV	Paralogous Sequence Variant
QTL	Quantitative Trait Loci
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic Acid
rpm	Revolutions per Minute
s	Seconds
SNP	Single Nucleotide Polymorphism
spp.	Several Species
STR	Short Tandem Repeats
T	Thymine
TAC	Total Allowable Catch
<i>Taq</i>	<i>Thermus aquaticus</i> DNA polymerase
TE	Tris-Ethylenediamine Tetra-acetate
TM	Trade Mark
Tris-HCl	Tris- (hydroxymethyl) Aminomethane Hydrochloric Acid
USA	United States of America
US\$	United States Dollar
v.	Version
v/v	Volume to Volume
w/v	Weight per Volume
ZAR	South African Rand

Chapter One

Literature Review

1. Abalone in general

1.1 HABITAT AND DISTRIBUTION

Abalone are marine snails and have a worldwide distribution in coastal tropical and temperate waters. These molluscs are found along rocky shores and reefs, usually from sea level to up to 30 m deep (Degnan *et al.* 2006). Even though no single abalone species is globally distributed, molecular phylogenetics studies have indicated four discrete areas of endemism namely North America, New Zealand, Australia and South Africa (Lee & Vacquier 1995; Geiger 2000; Estes *et al.* 2005).

1.1.1 South African abalone

Haliotis midae, or perlemoen as it is known locally, is one of 56 abalone species found worldwide (Geiger 2000) and one of five species endemic to South Africa (Figure 1.1), with six species found around the coast of Southern Africa (*H. alfredensis*, *H. midae*, *H. parva*, *H. pustulata*, *H. queketti* and *H. spadicea*). Perlemoen is the largest of the species in South Africa and has a wide coastal distribution ranging from St. Helena Bay (west coast) to Port St. Johns (east coast); a stretch of approximately 1500 km (Lindberg 1992; Geiger 2000) (Figure 1.1). It is also the only species that has commercial value in South Africa (Sales & Britz 2001).

1.2 TAXONOMY

Abalone forms part of the phylum Mollusca that also includes clams, sea slugs, octopuses and squid. These animals have a body that is surrounded by a mantle, a large adductor muscle (also known as the foot) and an anterior head. Another characteristic that these animals are widely known for are their beautifully formed and coloured calcareous shell that is secreted by the mantle (Bunje 2010).

Within this phylum abalone is grouped in the class Gastropoda along with other snails, whelks and sea slugs. Unlike clams, gastropods only have one shell (or none at all), and not two. Within Gastropoda, abalone forms part of the subclass Vetigastropoda, superfamily Haliotoidea, family Haliotidae in the genus *Haliotis* (Geiger 1999).

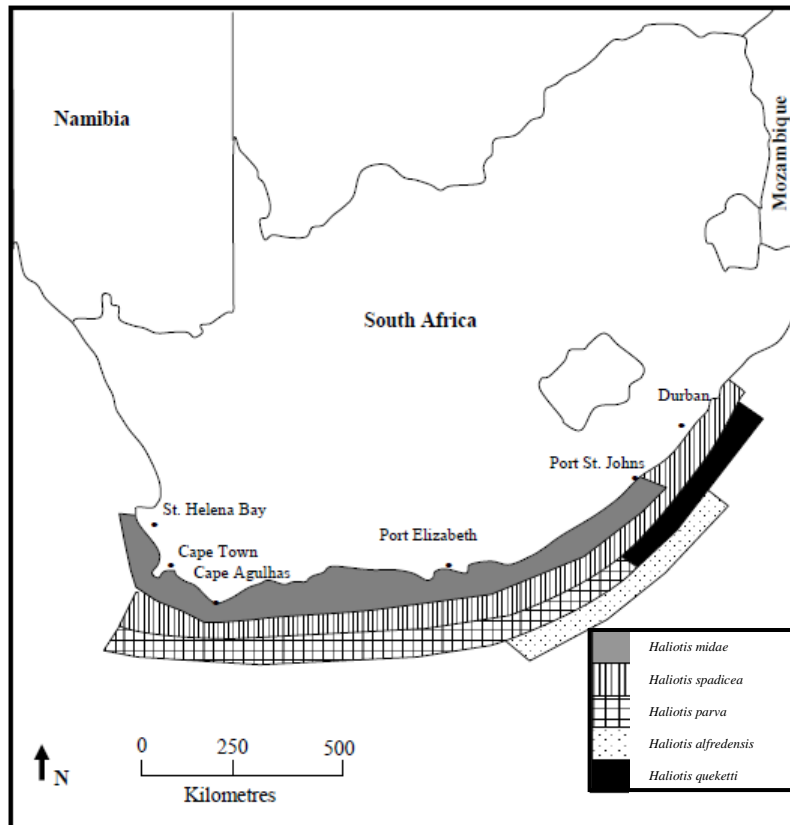


Figure 1.1: A map indicating the distribution of the five endemic abalone species found along the South African coastline (Lindberg 1992).

Shell shape and -size, along with geographical location of the animals are used as the basis of taxonomy and to date four possible subgenera have been identified in *Haliotis* including 1) *Haliotis*, *sensu stricto*; 2) *Nordotis*; 3) *Notohaliotis* and 4) *Sanhaliotis* (Geiger 2000) with *Haliotis midae* residing in *Notohaliotis* according to observations by Lee *et al.* (1995) and Van Wormhoudt *et al.* (2009).

1.3 ANATOMY

The protective shell of the abalone is oval shaped, and is the most conspicuous part of the animal. The exterior is rough with a row of respiratory pores near the outer edge of the shell and could have sponges or different types of algae growing on it. These pores allow for the removal of waste products as well as respiration. The interior of the shell is smooth and pearl-like (Fallu 1991; Landau 1992).

The adductor muscle (the very muscular foot) has strong suction power allowing the animal to clamp tightly to rocky surfaces. The mantle and the epipodium surround the foot, with the latter being a sensory structure comprising of tentacles. The internal organs, of which the gonad is the most prominent, are arranged around the foot under the shell. For females, the colour of the gonad is green or gray, and for males it is cream-coloured (Fallu 1991; Landau 1992).

The abalone also has a pair of eyes, a mouth and an enlarged pair of tentacles. It has a tongue called the radulae, and no obvious brain structure. It has a heart as well as a gill chamber next to the mouth under the respiratory pores (Fallu 1991; Landau 1992).

1.4 LIFE CYCLE

Gametes from the male and female animals are spawned under conditions affected by water temperature, high wave actions or extreme weather conditions, length of day and lunar cycle. The presence of gametes in the water can also affect spawning, and multiple spawning events during one season are possible (Fallu 1991).

The newly spawned eggs will hatch as microscopic, free living larvae which will settle after a week or more. At this stage the abalone are termed spat and will begin developing the adult shell form. They will grow to sexually mature adults, and then the cycle repeats itself (Figure 1.2) (Fallu 1991).

1.5 FEEDING

Abalone are herbivorous, slow-growing, slow-feeding animals. The main source of food for the adults is seaweed, while for the juveniles it is microalgae and diatoms that are found on the surfaces on which they settle. During the larval stage they feed on phytoplankton (Fallu 1991; Elliot 2000).

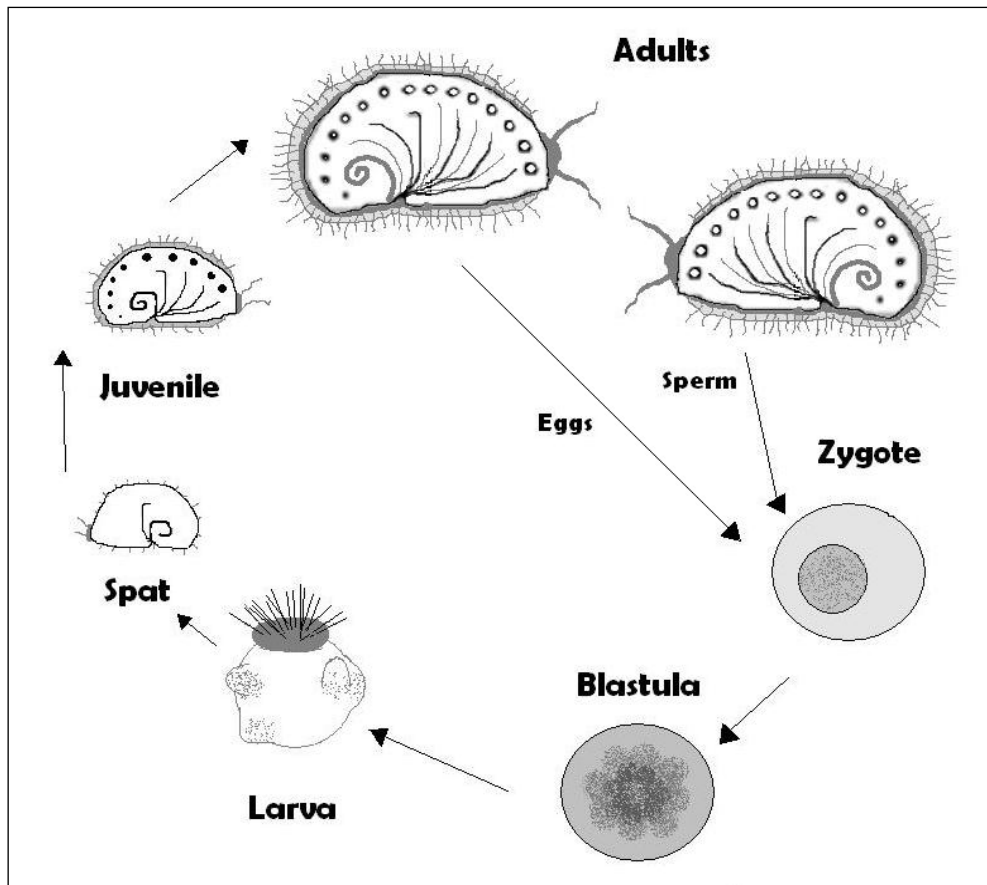


Figure 1.2: A demonstration of the life cycle of abalone (Hepple 2010).

2. Abalone culture

2.1 ABALONE FARMING IN GENERAL

During the 1990s, a rapid development of abalone cultivation took place due to the overexploitation and depletion of populations in the wild. Today, the cultivation of this natural resource is prevalent in many countries in order to supply the world demand including Australia, China, Iceland, Ireland, Japan, Mexico, New Zealand, Taiwan, USA and also South Africa (Gordon & Cook 2001; Cook & Gordon 2010). The major producer of cultivated abalone in the world is China with more than 300 functional farms with the largest one supplying more than 1000 metric tons per year (Cook & Gordon 2010). Outside of Asia, South Africa has become the largest producer of abalone. The abalone industry in South Africa has been dependent on a single commercially exploited species *Haliotis midae* with cultivation mainly being

driven by poaching (resulting in over-exploitation of wild stocks) and high market prices. Other factors that also promoted the fast growth of the South African industry include favourable coastal water quality, cheap labour and infrastructure (Troell *et al.* 2006).

Of the 56 *Haliotis* species, only 15 (not including hybrids) have commercial value as a result of size and growth rate limitations rendering most inadequate for farming purposes (Geiger 2000). These include species from Australia, Europe, Japan, New Zealand, South Africa, Thailand and the USA (Table 1.1) (Tarr 1989; Roodt-Wilding 2007; De la Cruz & Gallardo-Escárate 2011).

Table 1.1: Commercial abalone species and their countries of origin.

Country	Australia	Europe	Japan	New Zealand	South Africa	Thailand	USA
Species	<i>H. rubra</i> <i>H. laevigata</i> <i>H. roei</i>	<i>H. tuberculata</i>	<i>H. discus</i> <i>H. discus hannai</i> <i>H. gigantea</i> <i>H. sieboldii</i> <i>H. diversicolor</i>	<i>H. iris</i>	<i>H. midae</i>	<i>H. asinina</i>	<i>H. rufescens</i> <i>H. corrugata</i> <i>H. fulgens</i>

2.2 ABALONE FARMING IN SOUTH AFRICA

In 1949, commercial harvesting of perlemoen was initiated in South Africa. Abalone harvesting occurred mainly in the Western Cape region with the most intensively fished areas in previous years being zone A - D (Figure 1.3). Up until 1970 no fishing regulations and limitations on abalone harvest were in place, but in 1983 a mass quota system was introduced which led to numbers remaining relatively stable. However, in the 1996/1997 harvesting season, downward adjustments had to be made and total allowable catch (TAC) was decreased by 90% due to over-exploitation in zone C (Hauck & Sweijd 1999).

A biological problem also started to emerge that led to further declines in natural stocks. The areas where poaching had been most widespread started to experience large-scale movements of rock lobster into these areas. This resulted in the increased predation of sea urchins, which in turn lead to further decreases in juvenile abalone numbers, due to sea urchins providing juvenile abalone with shelter against

the elements. This also contributed to the striking decrease in TAC for fisheries between 1996 and 2005 (Hauck & Kroese 2006).

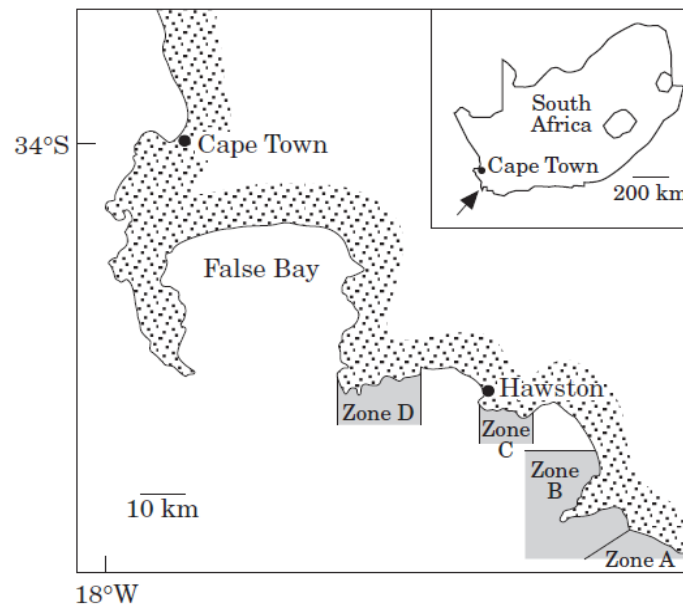


Figure 1.3: Location of fishing zones (A - D) along the South African coast (Hauck & Sweijd 1999).

This ecological disturbance, together with the illegal fishing, eventually led to the 'abalone crisis' which resulted in the complete closure of the fisheries in 2008 but which was opened again in 2010 allowing only commercial fisheries to continue and prohibiting all recreational fishing of abalone (Raemaekers *et al.* 2011).

Due to the increased international demand for the product, in 1981 the first attempts were made to cultivate perlemoen when captured specimens were successfully spawned to produce spat and juvenile abalone (Genade *et al.* 1988). To date, 14 abalone farms have been successfully set up with the majority (11) of these farms located in the Western Cape (DAFF 2011).

From 2000 to 2010, a total of 7208.09 tons abalone was produced on South African farms with 1015.44 tons produced during 2010. This was 101 tons more than in 2009; thus an 11.1% increase was recognised for abalone production (Figure 1.4). Produced abalone is mainly exported to Asia, with the value per kg being US\$34 in 2010. The total exports in 2010 amounted to 1005.29 tons worth ZAR352 million (DAFF 2011).

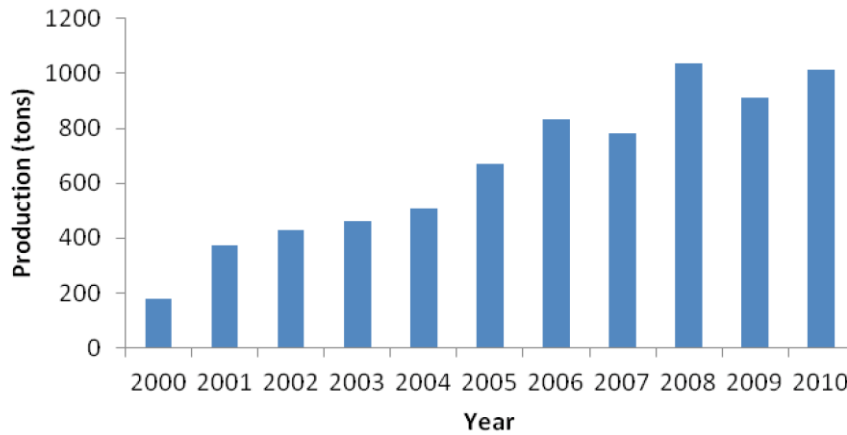


Figure 1.4: Annual abalone production from 2000 to 2010 (DAFF 2011).

Research has shown that growth rates in captivity are significantly higher than in the wild. Where it takes about 30 years for an adult to reach a size of 200 mm in the wild, aquaculture has increased this growth rate to a size of 100 mm in five years in animals reared in a hatchery (Sales & Britz 2001). Over the past 10 years aquaculture production has doubled worldwide and about 30% of fish and shellfish supply are produced through farming (Troell *et al.* 2006; Gjedrem *et al.* 2012). This rapid growth in the farming industry has resulted in an increased awareness regarding management strategies in order to allow sustainability of the industry.

3. Application of genetics in aquaculture

As aquaculture industries are growing at an immense rate, selective breeding programs have been vital in the successful development and ongoing viability of these major enterprises. All breeding initiatives are driven by the same motivation: limit inbreeding, improve genetic constitution of farmed stocks and create a commercially viable industry (Robinson *et al.* 2010). As growth rate is an important factor for determining profitability for abalone aquaculture, this has been a primary trait for improvement in most abalone breeding programs. Another constraint on abalone farming is the animals' susceptibility for disease. They are chronically infected with *Vibrio* spp. and this has damaging effects on the productivity of the industry (Hayes *et al.* 2007a; Robinson *et al.* 2010).

In 1919, the first documented selection program for fish commenced and increased survival against furunculosis in brook trout (*Salvelinus fontinalis*) was selected for. Selective breeding of farmed animals has improved a great deal and many experiments aiming to improve growth rate and disease resistance have since then been conducted (Gjedrem *et al.* 2012). The recent development of selective breeding programs in various countries including Australia, New Zealand, South Africa, China, Korea and Japan has strengthened the culturing of abalone. These programs and the success thereof depend greatly on the maintenance and enhancement of genetic diversity in farm stocks, as well as the consideration of the molecular components linked to important trait loci (such as increased growth rate and disease resistance) (Kang *et al.* 2011).

As mentioned before, it is important for aquaculture breeding programs to employ genetic management practices in order to ensure that inbreeding is limited and that genetic variability is maintained. Other important applications necessary for breeding programs include parentage assignment and linkage mapping studies that includes quantitative trait loci (QTL) mapping and further downstream marker-assisted selection (MAS) (Liu & Cordes 2004; Bowman *et al.* 2011).

It has been reported in previous studies (McAndrew & Napier 2010) that it is imperative to assign parents, especially in a marine species where it is challenging to develop single family rearing, to ensure that the replacement broodstock are not dominated by the offspring of a few individuals. Thus, family assignments are used for 1) taking family breeding value into account and 2) to reduce inbreeding (avoiding mating between related individuals) and increase genetic diversity (Beaumont *et al.* 2010). Determining parentage and estimating genetic diversity can be accomplished with the use of molecular markers.

Another application that also benefits greatly from the use of molecular markers is QTL mapping. High density linkage maps enable the identification of QTLs that are associated with important traits (for example disease resistance and enhanced growth rate; McAndrew & Napier 2010), and using these QTLs in MAS could accelerate the rate of genetic gain in a breeding program even further (Hayes *et al.* 2007a).

4. Molecular markers

Mutations are the main cause of DNA variation resulting in DNA polymorphisms. Various types of mutations exist, including point mutations, insertions, deletions and inversions. Insertions and deletions can cause shifts in sizes of the DNA fragments, and are easier to detect than point mutations which only have base substitutions and do not cause any changes in the fragment sizes.

Some of the first marker types that were used in aquaculture genetics were allozymes and mitochondrial DNA (mtDNA) markers. Recently, more useful markers with higher polymorphic power (the power to reveal genetic variation) such as restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP), markers mined from expressed sequence tags (ESTs), short tandem repeats (STR or microsatellite markers) and single nucleotide polymorphisms (SNPs) have become more popular (Table 1.2) (Liu & Cordes 2004; Varshney *et al.* 2007).

4.1 MICROSATELLITE MARKERS

Microsatellites are highly variable co-dominant markers with high levels of polymorphism and relatively small size (2 - 8 nucleotide repeats). These markers are abundantly distributed throughout genomes and can be easily amplified and rapidly detected (Chistiakov *et al.* 2006). Although microsatellite analysis has been the primary method of choice to use for the past two decades for various types of molecular applications, these markers display some negative characteristics. This includes size homoplasy, complex mutational patterns and being prone to genotyping errors (Glover *et al.* 2010). An additional disadvantage is the high levels of null alleles which lead to ambiguities when performing data analysis (Liu 2007). Due to base pair (bp) variations that occur in the flanking regions, cross-species amplification can be inhibited and screening of microsatellite loci is limited to very closely related species, or only the species in which they were developed (Kang *et al.* 2011). Due to abovementioned drawbacks associated with this marker, focus has shifted in recent years to single nucleotide polymorphisms as marker of choice in

many applications where microsatellite markers were previously predominantly utilised.

Table 1.2: Types of molecular markers, their characteristics and corresponding applications (Adapted from Sunnucks 2000; Liu & Cordes 2004; Schlötterer 2004; Collard *et al.* 2005).

Marker	Prior information required	Type I / Type II*	Polymorphic power	Mode of inheritance	Advantages	Disadvantages	Main application
Allozyme	Yes	I	Low	Mendelian, co-dominant	Relatively cheap, universal protocol	Tissue-specific, environmental factors may play a role, limited number of markers	Linkage mapping, population studies, studies of gene flow
mtDNA	No	-	High	Maternal inheritance	Multiple copies in cells	Only maternally inherited	Maternal lineage, intraspecific phylogeography, systematics
RFLP	Yes	I or II	Low	Mendelian, co-dominant	Robust, reliable	Bi-allelic, laborious and expensive to develop	Linkage mapping, fingerprinting, parentage assignment
AFLP	No	II	High	Mendelian, co-dominant	Simultaneous multiple loci analysis	Complicated methods for detection and analysis	Linkage mapping, population studies
SSR	Yes	I or II	High	Mendelian, co-dominant	Robust, reliable	Laborious and time consuming to develop, high mutation rates	Linkage mapping, population studies, parentage assignment, genetic variability studies
SNP	Yes	I or II	High	Mendelian, co-dominant	High genomic frequency, high-throughput, mutationally stable	Bi-allelic, expensive	Linkage mapping (fine mapping), population studies, cross-study comparisons

* Type I markers: associated with genes of known function; Type II markers: associated with unknown genomic content.

4.2 SINGLE NUCLEOTIDE POLYMORPHISMS

Single nucleotide polymorphisms result from a point mutation (the substitution of one nucleotide for another or the deletion or insertion of one or a few nucleotides) in the genome (Beuzen *et al.* 2000; Artamanova 2007). More recently SNPs have been used in applications such as population genetics and mapping studies (Garvin *et al.* 2010) as these polymorphisms occur frequently throughout the genome (approximately one SNP every 100 - 1000 bp, depending on genomic region as well as species) in both coding and non-coding regions. For gene-related SNP markers, this could lead to the advancement of mapping genes related to specific traits or when identifying genes under selection (Artamanova 2007; Glover *et al.* 2010).

Single nucleotide polymorphisms have become the marker of choice for genetic analyses for various reasons: Unlike microsatellites (or any other polymorphism) SNPs provide more potential markers near or in any locus of interest due to their high prevalence. They are also inherited in a more stable manner than microsatellite

markers, which makes them suited for long term selection markers. Due to the potential association with coding regions, SNPs can directly affect protein function, which might lead to the discovery of polymorphism directly responsible for variation among individuals regarding certain traits. Lastly, high-throughput technologies of SNPs are much more feasible and cost-effective than for any other polymorphisms (Beuzen *et al.* 2000; Sechi *et al.* 2010).

These markers however also have some disadvantages that have to be taken into consideration. Due to the fact that SNPs are generally bi-allelic markers (meaning in a population there are usually only two alleles), the information content per SNP marker is lower than markers that are multi-allelic (such as microsatellites) thus leading to more loci needed for satisfactory levels of statistical power in certain analyses. Depending on the level of heterozygosity, approximately five SNP markers contain the same information as one microsatellite marker; hence 30 - 50 SNPs will be able to equal the information provided by 10 - 15 microsatellites (Beuzen *et al.* 2000; Aitken *et al.* 2004; Rynänen & Primmer 2006).

These caveats mentioned previously can however be easily resolved by increasing the number of SNPs tested (Artamanova 2007). In a study by Glover *et al.* (2010) it was demonstrated that a highly informative set of SNP markers from a larger panel gave considerably more accurate data than any combination of microsatellite loci. Due to the many advantages associated with SNPs, interest in the high-throughput discovery and genotyping of SNPs is rapidly growing. Their abundance in genomes, the reduction in cost and the increased throughput of SNP assays have made these markers attractive for high-resolution genetic mapping, fine mapping of QTLs, linkage-disequilibrium based association mapping, genetic diversity analyses, genotype identification, marker-assisted selection and characterisation of genetic resources (Lepoittevin *et al.* 2010).

4.2.1 SNP discovery methods

For *de novo* SNP discovery, a validation step is required in order to determine if the observed polymorphism is in fact real. For species consisting of a reference genome this is not problematic, but for organisms with little genomic information available, discovering SNPs are slightly more challenging. In a situation where no reference

genome is available there are three main options to pursue in order to identify putative SNPs: 1) whole genome sequencing and assembly; 2) genome complexity reduction and sequencing methods and finally 3) cDNA sequencing. Whole genome sequencing has been performed for a number of species, but this is extremely demanding in terms of bioinformatic capacity and computational power when assembling the sequence scaffolds. For genomic libraries a high level of coverage is required for contig assembly and subsequent SNP identification. However, deep sequencing of cDNA libraries provides the optimal solution for species with limited genomic information content. The high sequence coverage that is needed for *de novo* SNP discovery is gained through transcriptome sequencing and because the SNPs are identified from transcriptomic data, they are directly associated with actual genes (Helyar *et al.* 2012).

Advances in DNA sequencing technology, specifically next-generation sequencing (NGS) has contributed significantly to SNP isolation procedures in non-model organisms. The two most widely used platforms for generating these datasets are the Illumina Genome Analyser II (GA II) and the Roche 454 FLX Titanium (for advantages/disadvantages associated with platforms see chapter two) (Ekblom & Galindo 2010). The use of NGS technologies have not only allowed for the generation of thousands of megabase pairs worth of sequence data, but it has also reduced the time and cost spent associated with DNA sequencing (van Bers *et al.* 2010). According to Renaut *et al.* (2010), the read lengths that are generated by these platforms allow for the satisfactory assembly of contigs for non-model organisms. In the event of SNP detection, the generated ESTs needs to be subjected to cluster analysis and assembly where after SNPs can then be identified by either *in vitro* or *in silico* means (Le Dantec *et al.* 2004). Discovering SNPs *in vitro* requires the re-sequencing of the amplicons in order to identify the variations, whereas *in silico* methods make use of bioinformatic measures resulting in a cheaper and less labour intensive approach to marker discovery (Useche *et al.* 2001). *De novo* SNPs however needs to be validated in order to avoid the identification of false SNPs (pseudo-SNPs) created by sequencing errors. As NGS results in large amounts of data generated it is possible to identify 100s - 1000s of SNPs. Consequently, validation methods are needed that can manage these large

numbers of SNPs in a timely manner. This can be achieved by employing various high-throughput genotyping methods.

4.2.2 Medium-throughput genotyping methods

'Medium-throughput genotyping' can be defined as genotyping >100 SNPs in 100 - 1000 individuals. Methods that are capable of this include Taqman, MALDI-TOF mass spectrometry-based systems, single-base extension-based assays, pyrosequencing and the Invader assay (Tsuchihashi & Dracopoli 2002).

4.2.2.1 PCR-free methods

Most genotyping methods require a pre-amplification step to amplify the genomic region that contains the SNP. Polymerase chain reaction (PCR) is normally used for this step, but a method that does not rely on PCR amplification is the Invader method. It is based on allelic discrimination that involves overlapping probes and an enzyme called Cleavase that specifically recognises the generated 'flap'. Two signal probes and a third invader probe is used. The signal and the invader probes hybridise together, creating a flap that is recognised by Cleavase if the signal probe completely matches the template. The cleaved flap can then either be detected by mass spectrometry or it can be used to generate a fluorescent signal (Twyman 2005). The Invader assay has great sensitivity as well as excellent signal to noise ratio, but the large amount of DNA required for reliable genotyping makes this method not ideal (Tsuchihashi & Dracopoli 2002).

4.2.2.2 Pyrosequencing

Pyrosequencing is a method that was initially employed for DNA sequencing, but due to its limited use in *de novo* DNA sequencing as a result of the relatively short read length, it is now used for genotyping. It works by releasing a pyrophosphate each time a nucleotide is incorporated at the 3'-end by DNA polymerase. Adenosine triphosphate (ATP) sulfurylase then converts the pyrophosphate to ATP, and ATP in turn causes luciferase to oxidise luciferin which leads to the release of a detectable light signal. Pyrosequencing is a very accurate method for genotyping due to its high specificity and ability to read the SNP position as well as its flanking regions, but a

very high degree of multiplexing would likely be difficult to achieve, making it a less suitable method to use (Tsuchihashi & Dracopoli 2002).

4.2.2.3 Mass spectrometry based genotyping assays

This method differs from others in that signal detection is based on the difference in molecular weight of small DNA fragments. Soft ionisation that is achieved by matrix-assisted laser desorption / ionisation time of flight (MALDI-TOF) is required for the DNA analysis by mass spectrometry. This method involves a metal plate, a matrix compound and allele-specific products. The allele-specific products are mixed with the matrix compound, and a short laser pulse is used to heat the mixture. The heat causes the mixture to expand into the gas phase and the application of a strong potential difference leads to ionisation. The ions are accelerated toward the detector and the time it takes for each ion to reach the detector (or the time of flight) is measured and the mass / charge ratio calculated. Alternative alleles in DNA fragments of 3 - 20 nucleotides in length can accurately be distinguished with high-resolution mass spectrophotometers. Advantages associated with this method are high accuracy and the ability to perform thousands of reactions in a single day due to a reaction only taking a fraction of a second (Twyman 2005).

4.2.2.4 Bead-based arrays

Bead-based methods have high multiplexing capabilities. These assays work on the principle of oligonucleotides that are attached to small microbeads (3 - 5 μm in diameter), and determining the identity of each bead. That information is combined with a genotype signal from the bead in order to assign a genotype call to each SNP and individual. A platform invented by Illumina captures the microbeads in solid wells created from optical fibres. The diameter of the beads and the wells are similar to each other, allowing for only one bead per well. Fifty thousand beads can be assembled in a single array and each can be treated as a high-density microarray, pushing the multiplexing potential for this system beyond its limits (Tsuchihashi & Dracopoli 2002).

4.2.2.4.1 Illumina GoldenGate genotyping assay with the VeraCode technology on the BeadXpress platform

The combined GoldenGate Genotyping assay and VeraCode technology has proven to be one of the most flexible, reliable and robust platforms for SNP genotyping. The assay's flexibility includes the ability to genotype SNPs from as few as 48 loci to a maximum of 384 loci on a single well of a standard microplate. The assay delivers consistent performance and utilises minimum cost and time. The assay has been successfully employed in the International HapMap Project (The International HapMap Consortium 2003) where it generated approximately 250 million genotypes (Illumina 2008). For more studies conducted using the GoldenGate genotyping assay see table 1.3.

Table 1.3: Previous studies employing the GoldenGate genotyping assay for medium-throughput SNP genotyping.

Species		SNPs genotyped	Model / Non-model species	Reference
Common name	Scientific name			
Apple	<i>Malus x domestica</i>	1411	Model	Khan <i>et al.</i> 2012
Maritime pine	<i>Pinus pinaster</i>	1536	Non-model	Chancerel <i>et al.</i> 2011
Atlantic cod	<i>Gadus morhua</i>	3072	Non-model	Hubert <i>et al.</i> 2010
Wheat	<i>Triticum</i> spp.	96	Model	Akhunov <i>et al.</i> 2009
Rainbow trout	<i>Oncorhynchus mykiss</i>	384	Non-model	Castaño Sánchez <i>et al.</i> 2009
Catfish	<i>Ictalurus</i> spp.	384	Non-model	Wang <i>et al.</i> 2008
Soybean	<i>Glycine max</i>	384	Model	Hyten <i>et al.</i> 2008

For genotyping a SNP, three oligonucleotide probes are used. Two are allele-specific oligonucleotides (ASO) and one is a locus-specific oligonucleotide (LSO). The allele-specific oligonucleotides are labelled with Cy3 or Cy5 fluorescent dye, and hybridise with their 3'-ends at the SNP site. The locus-specific probe is specific for a certain bead type and binds downstream of the SNP. Genomic DNA is attached to the solid support and then mixed with the three different probes for hybridisation. Any probes that did not bind to the DNA on the solid support are washed away.

Following the wash step is the enzymatic extension of the allele- and locus-specific oligonucleotides and ligation. The ligated strand is used as the template for PCR amplification, and the primers used in the reaction are specific for the ASO and LSO probes. The ASO-specific primer carries a fluorescent tag that is used for allele calling. PCR products are hybridised to the microarray through complementary oligonucleotides on the beads. The Cy3 / Cy5 intensity ratio is used to define the allelic state at a certain SNP position, with 1:1 indicating a heterozygote and 1:0 or 0:1 indicating a homozygote (Shen *et al.* 2005). Genotype calling is performed after clustering of dye intensities and predicts the accuracy of the results obtained (Figure 1.5).

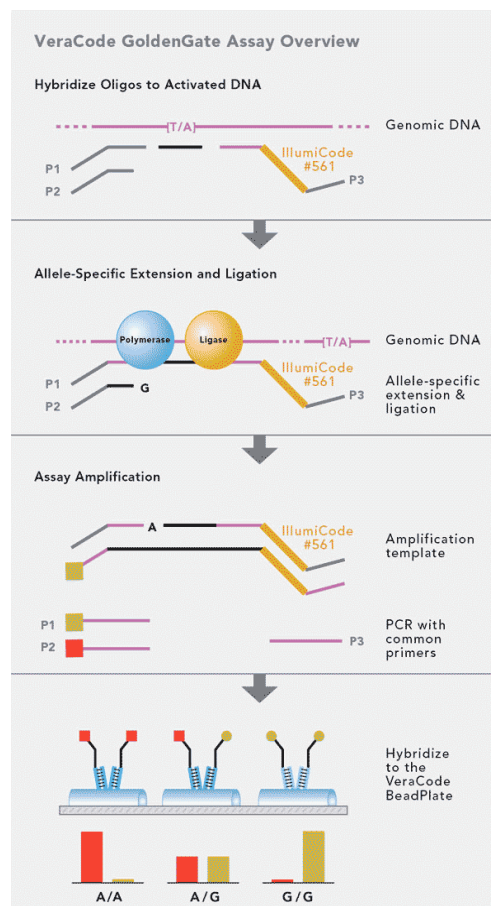


Figure 1.5a: A workflow of the VeraCode GoldenGate assay - hybridisation, extension, ligation, amplification.

(http://www.illumina.com/technology/veracode_goldengate_assay.ilmn)

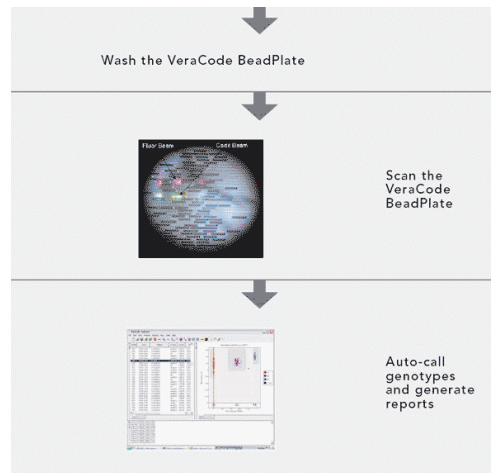


Figure 1.5b: A workflow of the VeraCode GoldenGate assay - wash, scan, allele calling.
 (http://www.illumina.com/technology/veracode_goldengate_assay.ilmn)

5. Linkage mapping

Linkage maps are essential to a wide range of genetic studies and can be utilised for the fine mapping of QTL, comparative analysis of synteny, searching for candidate genes and facilitating genome sequence assembly (Wang *et al.* 2011).

Linkage maps were initially constructed using dominant markers such as AFLPs and RFLPs. Initial linkage maps were called first generation maps and often had very low resolution. In the past few years, maps based on microsatellites and SNPs have started to emerge and are increasingly gaining popularity due to their higher resolution (Xia *et al.* 2010).

For many aquaculture species such as channel catfish (*Ictalurus punctatus*) (Waldbieser *et al.* 2001), rainbow- (*Oncorhynchus mykiss*) (Sakamoto *et al.* 2000; Nichols *et al.* 2003) and brown trout (*Salmo trutta*) (Gharbi *et al.* 2006), Atlantic salmon (*Salmo salar*) (Gilbey *et al.* 2004) and Pacific abalone (*Haliotis discus hannai*) (Liu *et al.* 2006; Sekino & Hara 2007) low resolution genetic linkage maps have been developed. Currently, a first generation linkage map is available for *Haliotis midae* (Hepple 2010; Jansen 2012), but unfortunately the resolution is still not high enough to allow accurate QTL mapping and trait association. Second generation linkage maps span genomes at much higher resolution containing several hundreds of markers and in many instances the markers are usually also developed

from ESTs and therefore associated with candidate genes (Castaño-Sánchez *et al.* 2010). Such maps are available for rainbow trout (Guyomard *et al.* 2006; Rexroad *et al.* 2008), Atlantic salmon (Lien *et al.* 2011) and channel catfish (Kucuktas *et al.* 2009) (Table 1.4).

Table 1.4: Linkage maps consisting mainly of SNPs and microsatellites for some aquaculture species (Adapted from Jansen 2012).

Species	Scientific name	Map length (cM)*	Mapped markers*	Ave. spacing (cM)*	Linkage groups	Reference
Atlantic salmon	<i>Salmo salar</i>	2402.3 / 1746.2	5650	0.4 / 0.3	29	Lien <i>et al.</i> 2011
Asian seabass	<i>Lates calcarifer</i>	2411.5	822	2.9	24	Wang <i>et al.</i> 2011
Japanese flounder	<i>Paralichthys olivaceus</i>	833.8 / 1147.7	1067 / 1167	0.8 / 1.0	24	Castaño-Sánchez <i>et al.</i> 2010
Atlantic cod	<i>Gadus morhua</i>	1421.92	924	1.5	23	Hubert <i>et al.</i> 2010
Grass carp	<i>Ctenopharyngodon idella</i>	1176.1	279	4.2	24	Xia <i>et al.</i> 2010
Blacklip abalone	<i>Haliotis rubra</i>	766 / 621	98 / 102	7.9 / 6.1	20 / 17	Baranski <i>et al.</i> 2006
Brown trout	<i>Salmo trutta</i>	912 / 346	288	3.2 / 1.2	37	Gharbi <i>et al.</i> 2006
Pacific abalone	<i>Haliotis discus hannai</i>	1774 / 1366	119 / 94	15.0 / 14.7	22 / 19	Liu <i>et al.</i> 2006

*If sex-specific maps were created, the female value is given before the male value

In the past, microsatellites were chosen for linkage mapping due to their high levels of heterozygosity and genome-wide distribution (Castaño-Sánchez *et al.* 2010). More recently however, the construction of linkage maps with SNPs has become more preferential due to the abundance as well as accurate, quick and automated genotyping of SNPs (Aslam *et al.* 2010).

With some SNPs being developed from sequences associated with genes (e.g. ESTs) these markers can either be linked to genes, be known to cause differences in gene expression or function, or be associated with certain traits. It is important for the current first generation microsatellite linkage map of *Haliotis midae* to be saturated with additional markers such as SNPs to provide information that can be used for comparative mapping and identifying important traits (Lien *et al.* 2011).

5.1 APPLICATION OF LINKAGE MAPPING: QTL ANALYSIS

Quantitative trait loci (QTL) can be defined as a part or a region of the genome that is associated with having an effect on a quantitative trait. A QTL can either be a cluster of linked genes, or a single gene that affects the trait. The key objectives of QTL mapping are to identify these regions that affect the trait of interest, as well as to analyse the effect of the QTL on the desired trait (Collard *et al.* 2005).

In order to perform QTL mapping two datasets are required; the phenotype of the quantitative trait and genotype of the marker. In a mapping family, parents and offspring are genotyped for a particular marker, where the parent will contain both of the alleles (heterozygous genotype) and segregation of the marker in the progeny evaluated. If the progeny can be sorted into two groups based on the absence or presence of a particular allele (genotype data), and a considerable difference in the mean phenotypic value between the two groups (phenotype data) is observed, the marker has a high probability of being linked to the trait (QTL) of interest.

Alternatively linkage disequilibrium (LD) can be applied where a linkage map with dense markers are available. Linkage disequilibrium can detect markers that occur even closer to a QTL than linkage could, which makes it a much more accurate method to use [linkage: certain genes are inclined to be inherited together because they are located on the same chromosome; linkage disequilibrium: the occurrence of two linked alleles (non-random association) in a population at a frequency higher or lower than expected]. Due to the availability of SNPs and their genotyping becoming more affordable, linkage disequilibrium mapping has become the method of choice for use in QTL mapping studies where linkage maps based on large number of SNP markers are available (Hayes *et al.* 2007a).

5.1.1 QTL mapping in aquaculture species

High density linkage maps are therefore essential for QTL mapping but although various linkage maps have been constructed for aquaculture species (Table 1.4), the maps are not dense enough for QTL mapping. This limits the number of QTLs being discovered. In aquaculture species, only a few QTLs have been identified for traits of economic importance. This is mainly in salmonids, where QTLs associated with

resistance to disease genes have been identified (Sonesson 2007; Gheyas *et al.* 2010). Other examples of QTLs (stress, meat quality, growth and disease resistance) identified in large genomic regions due to lack of a high resolution linkage map in fish species include Asian seabass (*Lates calcarifer*), European seabass (*Dicentrarchus labrax*), Japanese flounder (*Paralichthys olivaceus*), rainbow trout and tilapia (*Oreochromis* spp.). This exemplifies the fact that QTL mapping for aquaculture species is still in its infancy (Wang *et al.* 2011). Nevertheless, as SNP discovery and genotyping methods for non-model species such as *Haliotis midae* is becoming more efficient and inexpensive as technology advances and more genetic information is acquired, the construction of high density linkage maps required for QTL mapping is within reach for research laboratories (Garvin *et al.* 2010).

5.1.2 Linkage mapping: a precursor for QTL mapping in abalone

As mentioned before, high-density linkage maps are needed for QTL analyses, but in the family Haliotidae QTL mapping is still in the initial phases. Progress has however been made regarding type and number of markers developed and mapped for haliotid species. The first linkage map for abalone was constructed by Liu *et al.* (2006) (*H. discus hannai*), and consisted only of AFLP markers. Thereafter the first linkage map to contain microsatellites was constructed by Baranski *et al.* (2006) for *H. rubra*. Since these initial maps, linkage map construction has been expanded to other haliotid species: to date, linkage maps exist for *H. discus hannai* (Liu *et al.* 2006; Sekino & Hara 2007), *H. rubra* (Baranski *et al.* 2006), *H. diversicolor* (Shi *et al.* 2010; Zhan *et al.* 2011) and *H. midae* (Badenhorst 2008; Hepple 2010; Jansen 2012). The map constructed by Jansen (2012) was however the first linkage map to contain SNP markers for any abalone species. Because of the various advantages associated with SNPs, saturating linkage maps with these markers are becoming more prevalent. Due to abalone being artificially cultivated for economic purposes, the identification of markers linked to a QTL that is associated with growth and disease resistance would greatly benefit this farming sector. Previous abalone studies that have successfully identified QTLs for growth-related traits and growth rate include Liu *et al.* (2007) and Baranski *et al.* (2008). In a preliminary study five QTLs associated with growth has been identified for *H. midae*, but these QTLs still have to be validated in other mapping families (Roodt-Wilding & Brink 2011).

6. Marker-assisted selection

In the past, plant and animal breeders mainly relied on pedigree information, phenotypic trait values and/or estimated breeding values in order to breed genetically superior individuals. Recently, with the aid of highly saturated linkage maps, molecular markers and QTL, an improved approach namely marker-assisted selection (MAS) has become available.

Marker-assisted selection is a practice where a marker (DNA-variation based) is used for the selection of a desired trait. Due to the fact that the marker that is associated with the trait is selected and not the gene itself, it can be called an indirect selection process.

A limitation that is currently associated with MAS is the lack of high resolution genetic maps. Denser maps would allow for the detection of markers that are physically closer or maybe even within genes of interest. This would lead to the selection of a favourable trait based on the molecular marker associated with it. An example (specifically for abalone culture) would be if a QTL for improved growth rate which also displayed linkage disequilibrium with a set of markers could be identified. The markers could then be used to select individuals with enhanced growth rates at an early stage, which can ultimately be used for breeding as such marker-selected individuals are likely to produce offspring exhibiting enhanced growth rates. Due to the animals then reaching market size at an earlier age, time and money spent will have been minimised; thereby optimising the farming of the species (Roodt-Wilding & Slabbert 2006).

7. Aims and objectives

The project will consist mainly of two parts:

7.1 SNP MARKER DEVELOPMENT

Aim

To develop and validate single nucleotide polymorphism (SNP) markers from the sequenced transcriptome of *Haliotis midae* with high-throughput technology for use in various genetic applications.

Objectives

Identify putative SNPs in the sequenced transcriptome of *Haliotis midae* using CLC Genomics Workbench and validating it by genotyping six linkage mapping families using the Illumina GoldenGate genotyping assay with VeraCode technology on the BeadXpress platform.

7.2 LINKAGE MAPPING

Aim

To create a high resolution linkage map for *Haliotis midae* using newly developed SNP markers in six linkage mapping families in conjunction with previously developed SNPs and microsatellite markers.

Objectives

Genotype markers in the six different mapping families in order to identify polymorphic loci for linkage map construction. Markers will be placed into linkage groups based on linkage of odds (LOD) analysis and maps will be created using the regression mapping function as well as the maximum likelihood mapping function. Map distances will be calculated using Kosambi's mapping function. JoinMap[®] v.4.1 will be used to conduct the analysis. Due to male and female genomes having different recombination rates and differing in size, sex-specific and sex-average maps will be drawn up separately. Genome size will be estimated in order to determine the degree of genome coverage.

Chapter Two

Marker Development

1. Introduction

Single nucleotide polymorphisms (SNPs) are not only the most widespread type of DNA polymorphism, but are also easy to type due to their bi-allelic nature and have good reproducibility rates. As these molecular markers represent the finest resolution of a DNA sequence they are often referred to as the ultimate genic markers and offer numerous advantages that set them apart from other molecular markers. Except for the abundance of SNPs in genomes, these markers have low scoring error rates and the possibility of high-throughput genotyping makes SNPs well suited for use in various genetic applications including population genetic and mapping studies (Garvin *et al.* 2010; Helyar *et al.* 2011; Singhal *et al.* 2011).

For many non-model organisms little genomic information is available due to insufficient DNA sequence data and unavailability of DNA markers. In many instances genome sequencing is not viable due to the associated costs as well as intensive bioinformatic analysis required, but recently this problem has been overcome by the advent of next-generation sequencing (NGS) (Seeb *et al.* 2011b). This new technology has made sequencing more affordable and resulted in the generation of large amounts of sequence data that can be mined for molecular markers including microsatellites and SNPs. One important aspect for discovering SNPs in organisms without a sequenced genome is however a genome reduction step (Slate *et al.* 2009). One of the most widely used reduction steps for non-model species is transcriptome sequencing (Seeb *et al.* 2011a).

In species with little or no genomic information, the identification and genotyping of polymorphisms are much more complicated than in well-studied organisms. *De novo* SNP mining can either be done by experimental (*in vitro*) or computational (*in silico*) methods. *In vitro* methods are time-consuming and costly, as this requires re-sequencing of amplicons. In contrast, *in silico* methods, although not as successful as *in vitro* methods, are more cost- and time-efficient (Lepoittevin *et al.* 2010; Singhal *et al.* 2011). *In silico* SNP identification makes use of large numbers of sequences present in databases (usually ESTs) and the SNPs are mined using various computer programs without having to perform experimental procedures.

When comparing SNPs discovered from transcriptomic (EST) sequences to SNPs identified from genomic sequences, the former includes advantages such as the ability to identify uncommon sequence variants (Picoult-Newberg *et al.* 1999) as well utility in association analyses for quantitative traits (Liu *et al.* 2011). Due to EST sequences consisting of transcribed sequences, the SNPs mined from this data are associated with actual genes. This permits the use of gene-linked SNPs for mapping as well as comparative genome studies (Wang *et al.* 2008). In a study by Milano *et al.* (2011) where *in silico* SNPs were mined from the transcriptome of the European hake (*Merluccius merluccius*), it was found that although the lack of a reference genome affected the genotyping success rate, it was still an efficient method for large-scale discovery of SNPs in non-model species.

A challenge that needs to be addressed however for computational SNP discovery is identifying sequencing errors which could potentially lead to the identification of false (pseudo) SNPs. Expressed sequence tags are partial sequences of cDNA clones which consist only of single pass reads and have high error rates ranging from 1 - 8% (Liang *et al.* 2000). These sequences however allow for the detection of SNPs in transcribed regions, and setting stringent criteria such as the minimum coverage of a contig and the minor allele frequency can help avoid the identification of false SNPs. The redundancy of reads generated by NGS is beneficial for SNP mining as this helps in identifying putative SNPs. When a contig consists of numerous reads, alignment mismatches can be identified as SNPs and in order to avoid pseudo-SNPs, these mismatches have to occur more than once (Souche *et al.* 2007).

Markers originating from EST sequences will aid in providing functional information that can readily be used for high-resolution genetic mapping, QTL identification, genetic diversity analyses and marker-assisted selection. The aim of this section of the study was to identify putative *in silico* SNPs from the previously sequenced transcriptome of *Haliotis midae*.

2. Materials and methods

The necessity for ethical clearance was clarified with the Stellenbosch University ethical committee and deemed not necessary due to the non-sentient nature of *Haliotis midae*.

2.1 EST CONSTRUCTION AND SEQUENCE ASSEMBLY

A total of 19 individuals from a single family were randomly selected and used for RNA extraction. The animals used were all two-year old siblings with shell sizes ranging between 26 and 64 mm in length. After placing the animals on ice to slow down muscle contraction, all the soft tissue (whole animal) were dissected away from the shell, cut into strips and transferred to a tube containing RNALater solution. Messenger RNA molecules containing poly-A tails were isolated, fragmented and copied into cDNA for high-throughput DNA sequencing on the Illumina Genome Analyser II (GA II). High quality reads were assembled *de novo* using the CLC Genomics Workbench v.4.0 software (CLCbio, Aarhus, Denmark). Sequence annotation was performed by making use of the Desktop cDNA Annotation system (dCAS) v.1.4.3 and Blast2GO v.2.4.4. The databases against which the annotation was completed included the Eukaryote Clusters of Genes (KOG: Tatusov *et al.* 2003), Gene Ontology (GO: Ashburner *et al.* 2000) and the database of Protein Families and Domains (PFAM: Finn *et al.* 2010). All of the above was performed in a previous study and described in detail in Franchini *et al.* (2011).

2.2 *IN SILICO* IDENTIFICATION OF SNPs

More than 25 million short reads were generated by the Illumina Genome Analyser. *De novo* assembly was carried out by CLC Genomics Workbench. SNP detection (Altshuler *et al.* 2000) for the current study was also performed using CLC Genomics Workbench using the mapping functionality with specific criteria of a minimum coverage of 80 as well as a minor allele frequency (MAF) of 10%. To ensure reliable primer design, identified SNPs were checked for flanking sequences of 60 bp in which no other polymorphisms occurred. The sequences containing the selected SNPs were then subjected to BLAST homology searches using Blast2GO v.2.4.4 (Conesa *et al.* 2005) to investigate their potential function in other species (primarily

fish and shellfish species) and also broader where no significant hits to fish and shellfish species were found but important functions could still be conferred. The sequences which adhered to the abovementioned criteria were submitted to Illumina (Illumina, Inc., San Diego, CA) for processing by the Illumina[®] Assay Design Tool (ADT).

3. Results

3.1 SEQUENCE ASSEMBLY

Expressed sequence tags from the sequenced transcriptome were imported into the CLC Genomics Workbench for *de novo* assembly in order to identify putative SNPs. The total number of contigs that resulted from the assembly was 22 761, with an average size of 400.96 reads/contig and an average length of 260.62 bp/contig.

3.2 PUTATIVE SNP DISCOVERY

Of the 22 761 contigs 4 380 contained SNPs; with 11 934 SNPs in total. The average SNP frequency amounted to 1 SNP every 500 bp (Franchini *et al.* 2011). After setting the criteria of a minimum coverage of 80 and a MAF of 10%, 958 assembled contigs containing 3 645 SNPs remained. Of these, 400 SNPs from 256 contigs were identified that adhered to the criteria of 60 bp flanking regions of the SNP. Design rank scores of 0 - 1 were assigned to each SNP. The higher the design rank score of a SNP, the higher probability it has of being successfully converted into a genotyping assay. Scores of 0.5 - 1 are required for a high-quality assay, but only scores above 0.75 were considered for genotyping in the current study. After designability rank scores were assigned to each locus, 186 of the SNPs (from 139 contigs) which had the highest designability rank scores (all of which were 0.75 and higher) were selected for inclusion into the assay. The custom assay, adequate for genotyping 480 samples, comprised of 192 SNPs including six markers from a previous study (Blaauw 2012) to serve as positive SNP controls, and 186 SNPs developed *in silico* as described above. The assay was manufactured by Illumina in California, USA.

Of the 400 putative SNPs that were identified, the observed transitions were 253 (63.3%) and the observed transversions were 145 (36.3%); giving an observed transition to transversion ratio of 1.74. Two tri-allelic SNPs were also observed but were excluded from further analysis (Table 2.1).

Table 2.1: Summary of putative *in silico* SNP discovery in *H. midae*.

Number of contigs	256
Number of putative SNPs	400
Transversions	
A/T	52 (13.0%)
A/C	35 (8.8%)
C/G	15 (3.8%)
T/G	43 (10.8%)
Transitions	
A/G	124 (31.0%)
T/C	129 (32.3%)
Other	2 (0.5%)

To search for significant similarity against genes of known function, a BLASTX (protein BLAST) search using Blast2GO was performed on the 139 selected sequences (Addendum 1). Of the 139 contigs, 110 had significant hits (75 from fish and shellfish species, 35 from organisms not related to fish and shellfish species) and 29 had no hits. The sequences that had significant similarity to non-fish or -shellfish species were still included because they represented important functions in these organisms: *tyrosine 3-monooxygenase / tryptophan 5-monooxygenase activation epsilon polypeptide* (produces 14-3-3 epsilon protein that activates and / or inactivates other proteins involved in cell signaling), *chromosome segregation protein SMC* (involved in chromosome segregation) and *translation elongation factor 2* (involved in protein synthesis). The 110 sequences with significant hits were categorised into Mollusca (53), Chordata (14) and Other (sequences not forming part of the before mentioned groups for example acorn worm, starlet sea anemone, purple sea urchin and wolf spider) (43). The group Mollusca was further divided into Gastropoda and Bivalvia, which had 30 and 20 hits, respectively. Of the remaining

molluscan taxa, one belonged to the Cephalopods and two to the class Polyplacophora (Figure 2.1).

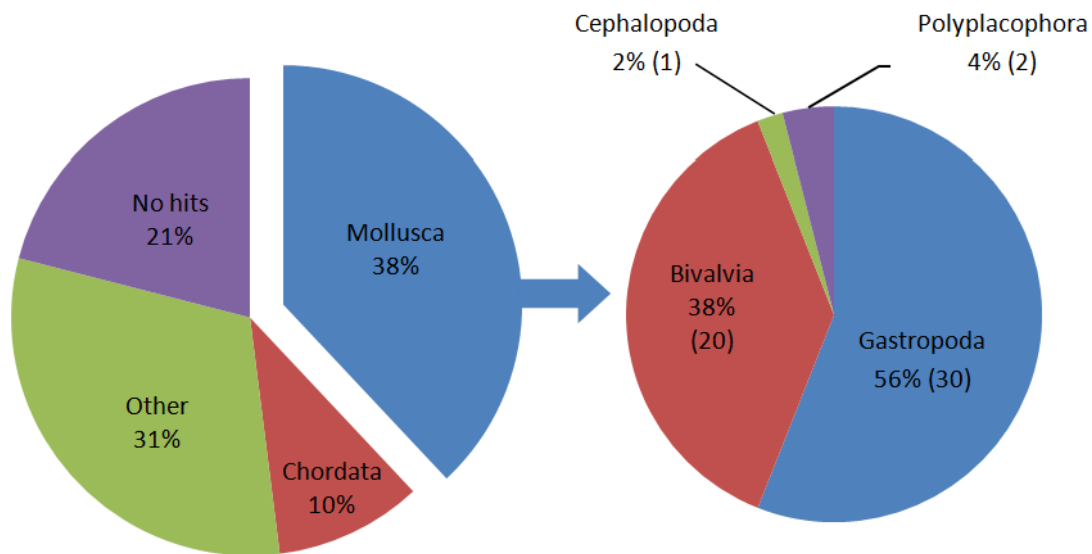


Figure 2.1: Classification of 139 contigs.

4. Discussion

In the current study a total of 400 putative SNPs were identified from EST contigs (produced by transcriptome sequencing) using the CLC Genomics Workbench after setting criteria necessary for genotyping *in silico* SNPs. Of the 400 SNPs, 186 were selected for genotyping based on a designability rank scores of 0.75 and higher. The criteria set for identifying putative SNPs (MAF 10%, coverage 80, 60 bp flanking regions) in the current study proved to be successful and are comparable to other studies that also aimed to identify *in silico* SNPs. One such study was conducted on catfish (*Ictalurus spp.*) (Liu *et al.* 2011) where a MAF of 10%, coverage of 100 and flanking regions of 15 bp (60 bp was not used as a different genotyping system than the GoldenGate Genotyping assay was employed) was used. The study by Liu *et al.* (2011) however identified a great deal more SNPs (1 129 100) than the current study. This can be ascribed to the higher number of contigs used (591 627) for SNP detection as well as the shorter flanking regions (15 bp) without other polymorphisms

utilised. A similar study conducted by Blaauw (2012) on *H. midae* also used a MAF of 10% and a minimum coverage of 100 and because of the validation of the putative SNPs with the GoldenGate Genotyping assay, the criteria of 60 bp flanking regions was also adhered to. The study by Blaauw (2012) identified only 12 *in silico* SNPs, but as that study was intended as a preliminary study to test the success rate of identification and genotyping of *in silico* developed SNPs, it explains the small number of identified SNPs compared to the current study.

When computer analysis is used to screen for polymorphisms (as is done when identifying putative SNPs *in silico*), true polymorphisms needs to be distinguished from sequencing errors to avoid including false SNPs. This problem can be addressed by setting a criterion such as MAF to a particular minimum (Wang *et al.* 2008; Lepoittevin *et al.* 2010; Liu *et al.* 2011) (10% in the current study). When the minor allele is present at least twice in a contig consisting of four or more sequences, it is highly unlikely that a sequencing error will be present in both ESTs at exactly the same base location. The coverage (read depth) is also an important factor to consider when identifying SNPs. According to Wang *et al.* (2008), the validation rates were directly proportional to the coverage: the deeper the coverage, the higher the validation rate. Due to NGS producing shorter DNA fragments, the data might contain short paralogous fragments consisting of paralogous sequence variants (PSV) (genetic changes not due to polymorphisms but due to single bp differences between paralogs; Beckman *et al.* 2007) that could also contribute to the false identification of SNPs (Ho *et al.* 2010; Liu *et al.* 2011). As the segregation of a PSV will be contradictory to that of a normal SNP, deviations from Hardy-Weinberg can help identify PSV (Gut & Lathrop 2004). If a SNP is revealed as a heterozygote in all individuals, it is highly likely that it is a PSV instead of a real SNP. Furthermore, if the *in silico* data is not representative of the populations used in genotyping these markers, the presumed SNPs that are selected may be monomorphic in the populations used for validation which in turn will lead to low conversion rates of the genotyping assay (Useche *et al.* 2001; Andreassen *et al.* 2010). This is known as ascertainment bias where the ascertainment process of a molecular marker is usually conducted on a detection panel of restricted size and therefore resulting in some of the informative loci not displaying variability on the specific panel.

Consequently the marker will not be useful in further data analysis (Guillot & Foll 2009).

Although a number of studies have been performed for *Haliotis* species regarding the identification of SNPs (Bester *et al.* 2008; Qi *et al.* 2008, 2009, 2010; Kang *et al.* 2011), this is the first large-scale *in silico* SNP discovery study. A preliminary study by Blaauw (2012) that included 12 *in silico* developed SNPs for *H. midae* resulted in a high genotyping success rate (83.3%), paving the way for further *in silico* work. In the study by Kang *et al.* (2011), the focus was on SNP markers to aid in phylogenetic analyses of various abalone species. Single nucleotide polymorphisms have also been developed for use or potential use in linkage mapping studies (Qi *et al.* 2010; Jansen 2012) in *H. discus hannai* and *H. midae* with the aim of future marker-assisted selection programs to identify markers associated with quantitative trait loci (QTL) that will promote the rate of genetic gain acquired from selective breeding programs (Franchini *et al.* 2011).

The SNP transition to transversion ratio found in the current study (1.74) correlates well with ratios found in other fish species [Atlantic salmon (*Salmo salar*): 1.37 (Hayes *et al.* 2007b) and Lake whitefish (*Coregonus clupeaformis*): 1.65 (Renaut *et al.* 2010)] as well as mollusc species [Eastern oyster (*Crassostrea virginica*): 1.3 (Quilang *et al.* 2007) and Weathervane scallop (*Patinopecten caurinus*): 2.4 (Elfstrom *et al.* 2005)]. Transition to transversion ratios previously found in *Haliotis* species include 0.67 - 1.71 for *H. midae* (Bester *et al.* 2008; Rhode 2010; Blaauw 2012) and 2.2 for *H. discus hannai* (Qi *et al.* 2009). Transitions are expected to occur twice as much as transversions, and higher transition rates has previously been explained to likely be correlated with the high mutation rate associated with CpG-like repeat units that causes an elevated occurrence of T/C transitions due to cytosine 5-methylation (Vignal *et al.* 2002; Arnheim & Calabrese 2009).

Not all markers are equally valuable for different studies or applications and it depends on for example mode of mutation, location in the genome, type of dominance expression and role in gene expression. SNPs are however applicable to a wide range of studies due to the co-dominant nature of the markers, a simple yet well-defined mutational model, a widespread distribution throughout genomes, easy

allele scoring and also their potential for high-throughput genotyping (Garvin *et al.* 2010).

The identification of SNPs has become very useful in fields such as marker-assisted breeding, conservation, resource management, evolutionary- and ecological studies and aquaculture genetic studies to name a few (Garvin *et al.* 2010). These studies could benefit from rapid and cost effective methods of SNP detection such as the *in silico* method described in the current study. A few recent species in which *in silico* SNPs from transcriptome data were utilised include European hake (*Merluccius merluccius*, Milano *et al.* 2011), Pacific herring (*Clupea pallasii*, Roberts *et al.* 2012) and catfish (*Ictalurus* spp., Liu *et al.* 2011). Single nucleotide polymorphisms detected in abovementioned studies were applied in population genetic studies (population structure), ecological- and evolutionary studies (selection and local adaptation) and management and conservation strategies (study of performance and production traits). A study by Nielsen *et al.* (2012), for example, illustrated that SNPs associated with transcriptomic regions could be successfully employed in the identification of individuals to the population of origin for forensic purposes. They examined gene-associated SNPs of four commercial marine fish [cod (*Gadus morhua*), herring (*Clupea harengus*), sole (*Solea solea*) and hake (*Merluccius merluccius*)] for assigning individuals back to the population of origin in an attempt to address illegal, unreported and unregulated fishing and product mislabeling and were found to have exceptional high levels of accuracy in doing so.

Due to the reduction of natural populations, farming of abalone is of great interest. Their slow growth rate and susceptibility to disease however poses a difficulty for aquaculture practices. Therefore it is important and of economical gain to identify QTLs associated with genes responsible for growth, disease resistance and meat quality (to name a few) in farmed animals (Hayes *et al.* 2007a; Massault *et al.* 2008). The contigs in this study mostly represent genes of relevant functions in fish and shellfish species which include cellular processes and stress response. Genes associated with functions such as heat stress protection (*heat shock protein 90*), protein synthesis (*elongation factor 2*), muscle contraction and motility (*myosin heavy chain*) and calcium cell signaling pathways (*calcium-binding protein*) etc. have been identified. As these functions may be involved in the promotion of growth as

well as disease resistance, these new SNP markers show great potential and could ultimately be employed for marker-assisted selection and breeding of *H. midae*.

During this study sequences generated by the Illumina Genome Analyser II (GA II) was used for SNP detection. Another platform that is also frequently used for high-throughput sequencing is the Roche 454 FLX Titanium. A study conducted by Luo *et al.* (2012) found that despite the shorter read length of Illumina in relation to Roche 454, Illumina gave longer and more accurate contigs. The sequencing errors in the raw ends of the two platforms were comparable, but the costs were not. The cost of the data obtained from the Roche 454 platform was 4 times that of the data obtained from Illumina, as was also indicated in a study by Dames *et al.* (2010). Homopolymer sequencing errors have also been reported for Roche 454 technology (Dames *et al.* 2010) as higher sequencing error rates are associated with A- and T-rich homopolymers (Luo *et al.* 2012). Homopolymer sequencing errors result from non-linear luminescence corresponding to homopolymer length during pyrosequencing (Ronaghi 2001). It is presumed that due to both the high sequence coverage of Illumina that facilitates the resolution of homopolymer ambiguities to a great extent and Illumina's less pronounced sequencing biases that these errors were not observed. Even though Illumina and Roche 454 provided comparable assemblies, there are still instances where Roche 454 will be superior to Illumina (Table 2.2). Due to the significantly longer read lengths that are produced by Roche 454 sequencing it might be more useful when resolving sequences with palindromes or repetitive structures (Luo *et al.* 2012).

Table 2.2: Comparison of next-generation sequencing platforms (adapted from Ekblom & Galindo 2010).

Technology	Sequencing method	Major advantages for studies of non-model species	Major disadvantages for studies of non-model species
Roche 454	Pyrosequencing	Relatively long reads enables assembly of contigs even in the absence of a reference genome.	Relatively few reads result in shallower coverage of sequencing. High error rate, especially in homopolymers.
Illumina	Sequence-by-synthesis	Very deep coverage because of large number of reads gives accurate measurements of gene expression levels.	Short read length means that a reference genome is desirable for assembly.

It is evident that SNPs [most widespread type of sequence variation, Helyar *et al.* (2011)] present exciting new developments for numerous genetic applications in species with little genomic resources. As the development of technology such as NGS is moving towards inexpensive but efficient alternatives, the identification of *in silico* SNPs will become even more rapid, more accurate as well as more cost-effective for species with limited genetic information (Helyar *et al.* 2011). Due to the possibility of producing *in silico* SNPs on a large-scale it will also become increasingly easier to incorporate SNPs into molecular genetic studies (Garvin *et al.* 2010). The progress made regarding SNP discovery methods will aid in saturation of the linkage map available for *H. midae*. As the markers developed in this study are from transcribed regions in the genome, they are associated with genes of interest and can therefore contribute to stock structure analysis, studying selection and local adaptation, QTL mapping and in due course, marker-assisted selection in *H. midae*.

Chapter Three

Genotyping of *In Silico* Developed SNPs

1. Introduction

In the past, genotyping of species for which little genomic information was available was mostly dominated by microsatellite analysis. Regardless of the limitations set by these DNA markers, their superiority persisted due to the successful cross-amplification of primers in sister species that increased research outputs. However, these markers often experience an inability to be replicated among different laboratories due to potential errors in genotyping. This has also contributed to the shift from microsatellites to single nucleotide polymorphisms (SNPs) (Seeb *et al.* 2011a).

As it is possible to streamline SNP identification and genotyping by high-throughput methods, the benefits compared to microsatellite markers are expected to further increase. For instance, a group of several hundred SNPs will have greater power than microsatellites in applications that rely on multilocus estimators of differentiation including population- or parentage assignment. This can be attributed to these markers' lower error rate with regards to genotyping, their higher reproducibility as well as higher genome coverage (Coates *et al.* 2009; Seeb *et al.* 2011a).

Single nucleotide polymorphisms that are developed *in vitro* need to be validated by re-sequencing of amplicons before genotyping can commence (Useche *et al.* 2001). *In silico* developed SNPs however bypass this validation step by being genotyped directly after identification, with the genotyping step acting as the validation of the SNPs. Genotyping of a subsample set of specimens is necessary and important to confirm the existence of putative SNPs and has proven successful in a number of studies to validate the identified *in silico* SNP markers (Wang *et al.* 2008; Castaño Sánchez *et al.* 2009; Lepoittevin *et al.* 2010; Campino *et al.* 2011).

SNP detection and genotyping in non-model species are faced with many challenges such as cost, accuracy, equipment, difficulty of assay, throughput and multiplexing capacity that need to be considered. For example, a large variety of medium- to high-throughput genotyping techniques are available for model organisms, but often their use in non-model species is challenging. However, a method that has proven successful for this purpose is the Illumina GoldenGate genotyping assay with the VeraCode technology on the BeadXpress platform (Lepoittevin *et al.* 2010). This

method can handle from 48- to 384-plex levels, allowing thousands of genotypes to be achieved simultaneously in a short period of time (Illumina 2008).

The highly specific extension and amplification steps of the GoldenGate genotyping assay permits a high degree of loci multiplexing in a single reaction. One of the most noteworthy features of this genotyping assay is that it does not require any prior PCR amplification as it genotypes directly from the genomic DNA. By offering different plex levels that allows the user to either carry out smaller pilot studies or larger scale genotyping studies, the assay provides scalability and flexibility and has demonstrated outstanding performance in terms of call rate, reproducibility and development success rate (Shen *et al.* 2005; Lepoittevin *et al.* 2010; Campino *et al.* 2011). Other medium- to high-throughput methods capable of genotyping larger numbers of SNPs (>100) in 100-1000 individuals such as Taqman, MALDI-TOF mass spectrometry-based systems, single-base extension-based assays, pyrosequencing and the Invader assay were also options available but the GoldenGate assay was preferred for the current study due to obtaining good validation rates of *in silico* SNP genotyping in a preliminary study to test the success of the assay by Blaauw (2012).

Due to markers originating from EST sequences, the SNPs generated in the current study will aid in providing functional information that can be readily used for high-resolution genetic mapping, QTL identification, genetic diversity analyses and marker-assisted selection. The aim of this chapter was to validate 186 newly developed *in silico* SNPs identified from the previously sequenced transcriptome of *Haliotis midae* (see chapter two) using the GoldenGate assay as a medium-throughput genotyping method.

2. Materials and methods

The necessity for ethical clearance was clarified with the Stellenbosch University ethical committee and deemed not necessary due to the non-sentient nature of *Haliotis midae*.

2.1 SAMPLE COLLECTION AND DNA EXTRACTION

One thousand animals, from 10 families (100 animals per family), were collected from the Abagold aquaculture facility and transported to the laboratory at Stellenbosch University on ice. DNA extractions were performed on all 1000 animals using tissue from the adductor muscle. Destructive sampling was done, and the remainder of the animals not used for DNA extraction was stored as a tissue sample for future use. The parental DNA for these 10 families was previously extracted and also included in this study. DNA extractions were performed using the cetyltrimethyl ammonium bromide (CTAB) extraction method (Doyle & Doyle 1987). The tissue was homogenised in 300 µl CTAB lysis buffer (1.4 M NaCl; 20 mM Ethylenediamine tetra-acetate (EDTA [pH 8]); 2% (w/v) CTAB; 100 mM Tris-HCl [pH 6.5] and 0.2% (w/v) β-mercapto-ethanol) to which 0.5 mg/ml Proteinase K was added. The tissue was then incubated overnight in a water bath at 60°C. Equal volumes (300 µl) of chloroform: isoamylalcohol (24:1) were added to the solution. The samples were centrifuged at 12 000 rpm for 5 min using an Eppendorf centrifuge. The supernatant was carefully removed and transferred to a new eppendorf tube. DNA was precipitated by the addition of 2 volumes cold ethanol and incubated at -20°C overnight. The samples were centrifuged at 12 000 rpm for 20 min at 4°C and the pellet washed with 200 µl 70% (v/v) ice cold ethanol, followed by a second centrifugation step at 12 000 rpm for 10 min. The alcohol was removed and the pellet dried in an oven at 55°C. DNA was resuspended in 50 µl TE buffer (10 mM Tris-Cl [pH 7.5]; 1 mM EDTA) and stored at -20°C until further use.

2.2 PREPARING AND SELECTING FAMILIES FOR GENOTYPING

Parentage analysis was conducted using a QIAGEN[®] Multiplex kit. A panel that included seven microsatellite loci was used to validate family assignments (see addendum 2).

2.2.1 PCR Multiplex

A QIAGEN[®] Multiplex kit was used to amplify the target loci following the instructions provided by the manufacturer. The reactions were performed in a final volume of 7 µl as follows: 20 ng of template DNA was added to 3.5 µl 2X QIAGEN Multiplex PCR

master mix (containing HotStart *Taq*[®] DNA Polymerase, Multiplex PCR Buffer with 6 mM MgCl₂ and dNTP Mix) (QIAGEN[®]), 1.1 µl Primer mix (20 µM of each primer). The following PCR cycle was used to amplify the target locus: The cycle is initiated with a 15 min denaturing step at 95°C, followed by 35 cycles of 94°C for 30 s, 57°C for 90 s and 72°C for 90 s. The PCR was completed with an elongation step of 72°C for 10 min.

2.2.2 Genotyping of microsatellites markers

The microsatellite loci were genotyped using the ABI 3730xl DNA Analyser (Applied Biosystems). The lengths of the products were determined by comparing it to the GeneScan[™] 600 LIZ[®] Size Standard (Applied Biosystems). Allele scoring was performed using GeneMapper v.4.1 software (Applied Biosystems) in order to validate family composition.

2.2.3 Selected families for genotyping

Four of the original 10 families (Table 3.1) selected for DNA extraction contained a sufficient number of individuals for linkage analysis (70 and more) and were selected for SNP genotyping. Two additional families (Table 3.1) that were used in previous studies (Blaauw 2012; Jansen 2012) were also included in the SNP genotyping assay since these families contained mapped microsatellite markers; thus assisting with the integration of the SNPs and the microsatellite markers on the consensus map. After family composition was validated, DNA from each individual (offspring and parents) was sent to Inqaba Biotec [Inqaba Biotechnical Industries (Pty) Ltd] for PicoGreen[®] fluorometric dsDNA quantification to determine DNA concentrations. Based on quantification results, dilutions or precipitations were performed to achieve a final concentration of 50 ng/µl. The DNA was then placed in 96-well plates of which each contained three genotyping controls [individuals genotyped in the previous study of Jansen (2012) and with known SNP genotypes] (Table 3.1). The plates were sent to the University of the Witwatersrand and National Health Laboratory Services (NHLS) for SNP genotyping with the Illumina GoldenGate genotyping assay with the VeraCode technology on the BeadXpress platform.

Table 3.1: Parentage verified animals used for genotyping (* families included in previous studies).

Family	Origin	No of animals	
		Parents	Offspring
FamD	Abagold	A22 X B21	72
FamH	Abagold	A35 X B43	71
FamI	Abagold	A17 X B25	81
FamJ	Abagold	A24 X B28	72
FamDS1*	Roman Bay	F617 X M342	70
FamDS2*	Roman Bay	F462 X M456	87
Genotyping controls	-	-	15
Subtotal		12	468
Total		480	

2.3 SNP GENOTYPING ASSAY

A total of 250 ng genomic DNA (gDNA) was used as a template to perform SNP genotyping with the Illumina GoldenGate genotyping assay with the VeraCode technology on the BeadXpress platform. Following the manufacturer's protocol (Shen *et al.* 2005), paramagnetic particles, hybridisation buffer and assay oligonucleotides are combined with the genomic DNA (assay hybridisation step). After the PCR procedure and the down-stream processing of the single-stranded dye-labeled products, these products are hybridised to their complementary bead type by means of their unique address sequences. This in turn allows for the readout of the highly multiplexed SNP genotyping assay with the BeadXpress Reader that is used to analyse fluorescence signals.

2.4 DATA ANALYSIS

For analysing genotyping data generated by the Illumina GoldenGate genotyping assay, the GenomeStudio™ Genotyping Model v.1.0 was employed. The program allows for assessment of the raw data generated by the BeadXpress Reader.

Parameters used to assess the genotypes included a GenTrain score that applies a custom clustering algorithm, as well as a GenCall score that determines a quality score for each genotype called. These scores range from 0 - 1, with 1 being the highest probability of the score being accurate. In the current study a GenTrain score with a cutoff value of 0.45 was applied. Any scores lower than 0.45, or SNPs that did not cluster, were deemed genotyping failures. No cutoff value was applied for the GenCall score as these values were 0.8 and higher which indicated good quality sample genotypes.

3. Results

3.1 PARENTAGE ANALYSIS

Difficulty was experienced obtaining families that contained a sufficient number of individuals (70 and more). Out of the ten initial families, only FamC, D, H, I and J contained 70 and more individuals. FamC was however excluded from further analysis due to many individual samples with low DNA concentrations.

3.2 GENOTYPING PERFORMANCE

Of the 480 samples that were initially sent for genotyping with the GoldenGate assay, only 407 samples were successfully genotyped due to a lack of generated genotypes for some samples (most probably DNA quality) as well as missing genotype data (technical difficulties with the genotyping platform) (Table 3.2).

3.3 VALIDATION AND PERFORMANCE OF SNPS

Out of 400 putative SNPs, 139 contigs containing 186 putative SNPs were selected that had the highest designability rank scores (0.75 or higher) and fulfilled all genotyping prerequisites (see chapter 2). These SNPs were validated with the Illumina GoldenGate genotyping assay with the VeraCode technology on the BeadXpress platform using 480 abalone samples. Six SNPs from a previous study (Blaauw 2012) that served as positive SNP controls for the current study contributed to the total of 192 SNPs on the genotyping assay. After preliminary data analysis

was completed, 44 of the *in silico* developed SNPs were excluded from the study due to either no clustering or GenTrain scores with values below the cutoff value which, for this study, was 0.45. Of the 192 SNPs, 148 (including six positive SNP controls) were successfully genotyped with 133 (including five positive SNP controls) being informative (polymorphic) and 15 (including one positive SNP control) being non-informative (monomorphic) (Addendum 3).

Table 3.2: Genotyped samples.

Families	Individuals	No call	Missing data	Successfully genotyped samples
FamD	72	4	-	68
FamH	71	11	9	51
FamI	81	10	12	59
FamJ	72	9	10	53
FamDS1	70	1	-	69
FamDS2	87	6	-	81
Genotyping Controls	3 (per plate)	-	1	14
Parents	2 (per family)	-	-	12
Total	480	41	32	407

For the genotyping assay a success rate of 76.34% (142 SNPs) was obtained. This was calculated by dividing the number of loci that was successfully genotyped by the total number of SNPs in the assay (the control SNPs were not included in the calculation). As defined by Fan *et al.* (2003), the conversion rate was calculated by dividing the number of polymorphic SNPs by the total number of SNPs (68.82%). Of the 142 successfully genotyped SNPs, 128 (90.14%) were polymorphic and 14 (9.86%) were monomorphic (Table 3.3).

Table 3.3: Summary of successful and unsuccessful genotypes.

Categories	Number of SNPs*
SNPs genotyped	186
Successful genotypes	142
Polymorphic SNPs	128
Monomorphic SNPs	14
Failed SNPs	44

* Controls not included in statistics

4. Discussion

Given that abalone are nocturnal animals that move around extensively, it was necessary to validate family composition before the selected mapping family individuals could be genotyped. The genotyping families also needed to consist of a sufficient number of individuals for performing segregation analysis in order to construct linkage maps; which was the ultimate aim of the current study (chapter four). Of the 10 families used for DNA extraction, only four contained a satisfactory number of individuals.

An 192-plex GoldenGate genotyping assay for *Haliotis midae* was constructed from 186 SNPs screened from ESTs and six positive control SNPs from a previous study conducted by Blaauw (2012). The *in silico* SNPs were selected on the following grounds which proved vital for the validation of these markers: the minor allele frequency (MAF) of the SNP, the quality of the flanking regions and the number of sequences per contig used (coverage) for detection of the SNP. Considering the polymorphic as well as the monomorphic loci, the global success rate of the assay was 76.34%, and considering only the polymorphic loci; a conversion rate of 68.82% was reached. This compares well to other studies that also made use of the Illumina GoldenGate genotyping assay with the VeraCode technology on the BeadXpress platform. Examples include the maritime pine (*Pinus pinaster*): global success rate 66.9% and conversion rate 51% (Lepoittevin *et al.* 2010); rose gum (*Eucalyptus grandis*): success rate 87% and conversion rate 66.1% (Grattapaglia *et al.* 2011); catfish (*Ictalurus spp.*): global success rate 69% and conversion rate 59.8% (Wang

et al. 2008); perlemoen (*Haliotis midae*): global success rate 85.4% and conversion rate 64.5% (Blaauw 2012).

In the study conducted by Blaauw (2012), the conversion rates of *in vitro* versus *in silico* SNPs were also compared. The difference in conversion rates was significantly lower for the *in silico* SNPs compared to the *in vitro* markers. According to Wang *et al.* (2008) the lower conversion rates of *in silico* SNPs could be attributed to sequencing errors that causes the identification of pseudo-SNPs (false SNPs) which leads to genotyping failures in EST-derived SNPs. Other possible causes include low quality flanking sequences which influence primer design or the presence of intron-exon junctions near the SNPs of interest.

When selecting SNPs, all the primers (both allele-specific as well as locus-specific) should be located in the same exon, so that when genomic DNA is amplified there is no intronic region that requires PCR extension across it. Due to the limited ability of the BeadXpress technology to provide adequate primer extension in cases like these, SNP sites involving introns will in all probability fail in genotyping. Also, when a SNP site is situated near an exon-intron boundary it results in the inability of the primers to form base pairs with the DNA from the amplified gDNA (Figure 3.1) (Wang *et al.* 2008).

According to Shen *et al.* (2005), even though the GoldenGate assay can endure DNA degradation to a certain extent, the quality of the genotyping assay is severely compromised when less than 20% of the necessary template DNA is supplied. Insufficient DNA quantity is one of the main reasons for the failure of successful genotypes to be generated by an assay. This could explain the lack of generated genotypes (no calls) for some of the individual samples in the current study (Table 3.2) as some were known to have low DNA concentrations.

As mentioned before the coverage, MAF and the flanking regions of the SNP contribute greatly to the genotyping success of the assay. It is important for the coverage of the sequences to be high enough in order to minimise the effects that pseudo-SNPs have on the success rate of the assay. The higher the coverage, the smaller the chance is that a SNP is present due to a sequencing error. The MAF is equally crucial to the success rate of the assay. For instance, if a gene is sequenced

twice and the minor allele presents once, it is likely that when the gene is sequenced 10 times, the minor allele will present close to half of the times sequenced. However, if a gene is sequenced 10 times and the minor allele only presents once, it is likely that the observed minor allele is due to a sequencing error. Lastly, the regions flanking the SNP of interest play an important role when identifying reliable SNPs. It is essential that SNP hot spots and sequencing errors in the near vicinity of the SNP be avoided, as this will influence the base pairing of the genotyping primers, possibly leading to generation of false SNPs (Wang *et al.* 2008).

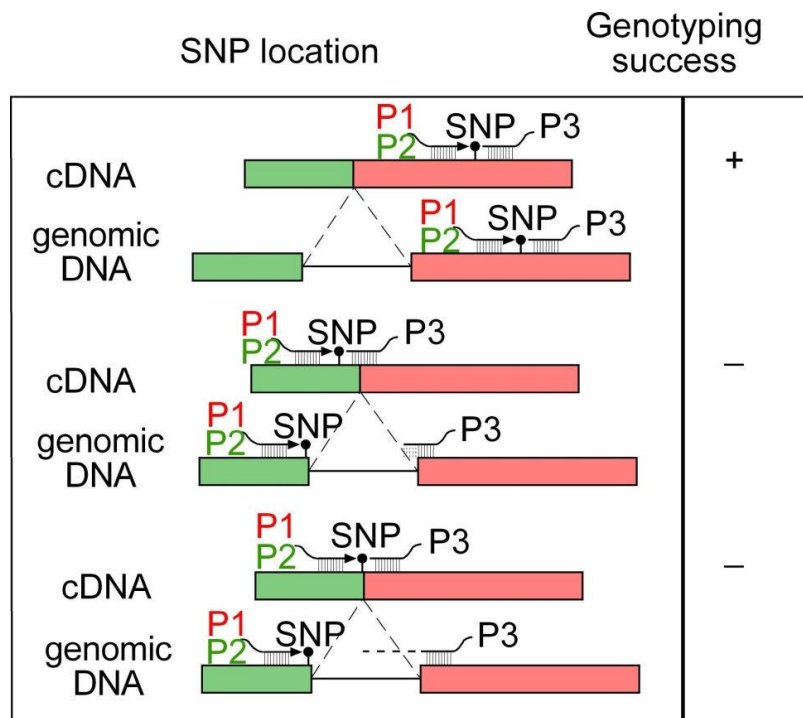


Figure 3.1: An illustration regarding the effects of introns on genotyping (Wang *et al.* 2008).

In order to ensure that good quality genotypes are obtained, Illumina has certain criteria which have to be met. Cutoff values for both GenTrain and GenCall scores of 0.25 are suggested, but the relationship between these parameters can only be interpreted within a specific study. The GenTrain scores indicate the level of separation between the homo- and heterozygote clusters for a certain SNP locus and the GenCall scores indicate the degree of reliability for each genotype called (Fan *et al.* 2003; Shen *et al.* 2005). In previous studies GenTrain cutoff values of 0.35 - 0.4 have been used (Wang *et al.* 2008; Khan *et al.* 2012), but as mentioned before these parameters can only be interpreted within a given study, which explains

the differences in cutoff values observed as well as the values used in the current study.

A comparison was made with two studies, one on domesticated apple (*Malus x domestica*, Khan *et al.* 2012) and one on catfish (*Ictalurus* spp., Wang *et al.* 2008) in order to evaluate the difference in success rates of the GoldenGate genotyping assay on a well-studied species (domesticated apple) and species with limited genomic information (catfish and abalone). The study that focused on the domesticated apple utilised a lower GenTrain cutoff value than the current study and also showed less failed genotypes but had a higher occurrence of monomorphic SNPs. The study conducted on catfish, which is a non-model aquaculture species, (as is *Halotis midae*) also made use of a slightly lower GenTrain cutoff value, but had a notably higher number of failed and monomorphic SNPs (Table 3.4).

Table 3.4: Comparison of genotyping assay success for a well-studied species^a with species of limited genomic information^b.

	Khan <i>et al.</i> (2012)^a	Wang <i>et al.</i> (2008)^b	*Du Plessis^b
Contigs	6 888	4 387	958 **
SNPs	37 807	33 594	3 645 **
Flanking sequences > 60 bp			
Contigs	6 525	-	257
SNPs	12 299	-	400
Design score	0.5	0.5	0.8
Contigs	-	-	139
SNPs	10 667	-	186
GenTrain score	0.35	0.4	0.45
Total	1 411	384	186
Failures	197 (14%)	118 (30.7%)	44 (23.7%)
Homozygous	367 (26%)	110 (28.6%)	14 (7.5%)
Polymorphic	847 (60%)	156 (40.6%)	128 (68.8%)

* Current study

** Number of contigs and SNPs after criteria of coverage (80) and MAF (10%) was set

Comparing the failed SNPs, a possible reason for the higher number in the catfish and the perlemoen could be due to the higher cutoff value assigned for the GenTrain

scores. Also, the higher failure rate for the catfish study could be attributed to the fact that the authors wanted to test SNP quality by evaluating different parameters leading to the lower success rate of the overall assay. When taking the polymorphic SNPs into consideration, the catfish study had less polymorphism than the apple study, but due to the high number of failed SNPs it is to be expected that the polymorphic markers (and overall successfully genotyped markers) would be less. Comparing the homozygous SNPs, the apple and catfish studies had quite high numbers in relation to the current study. The catfish monomorphic SNPs could be due to sequencing errors and the high number in the apple genome could be ascribed to the duplicated nature of the genome as a result of a whole genome duplication event occurring millions of years ago (Khan *et al.* 2012). The monomorphic SNPs in the current study are higher when taking each family separately, but on the whole a good success rate was achieved with the assay. Overall, it can be concluded that the assay works equally well for well-studied and non-model species.

It is evident that the GoldenGate genotyping assay can be successfully employed for genotyping *in silico* developed SNPs from sequenced transcriptome data. The high genotyping success rate achieved in the current study can also be attributed to the criteria that were set for identifying the SNPs. A MAF of 10%, minimum coverage of 80 and 60 bp flanking regions proved adequate and contributed to the success of the development and genotyping of SNPs for further downstream applications. Even though SNPs that were developed *in vitro* had higher conversion rates than *in silico* developed SNPs (Lepoittevin *et al.* 2010; Blaauw 2012), the time- and cost benefits associated with *in silico* SNPs as well as using genotyping directly as a validation step makes this method attractive for larger scale marker development studies.

Chapter Four

Linkage Mapping

1. Introduction

Since the first linkage map was constructed by Sturtevant in 1913, genetic studies have greatly benefited from these maps with regards to the ordering of markers on chromosomes or the linear position of genes. Genetic linkage maps however can be utilised in various other applications including evolutionary and comparative genomics studies by offering information on genome-wide recombination rates or insight with regards to inter- and intra-species gene reorganisation between and within chromosomes. However, the ultimate application of linkage mapping would have to be the pursuit of quantitative trait loci (QTL) (Ball *et al.* 2010).

A vast number of aquaculture species exhibit traits that are important to select for in order to increase production. Many of the economically valuable traits are also quantitative in nature. Before QTL mapping, the genetic enhancement of production traits primarily relied on pedigree and phenotypic information. The drawback to this is that pedigree and phenotype information are influenced by environmental factors, making it difficult to detect the genes responsible for the trait. However, the development of genetic markers has made it feasible to discover and select for QTLs associated with certain traits (Wang *et al.* 2011).

Linkage maps have been constructed for several foodfish species including salmon (*Salmo salar*), channel catfish (*Ictalurus punctatus*), tilapia (*Oreochromis* spp.), shrimps (*Caridea* spp.), European seabass (*Dicentrarchus labrax*), Asian seabass (*Lates calcarifer*), grass carp (*Ctenopharyngodon idella*) and Japanese flounder (*Paralichthys olivaceus*) to name a few (Xia *et al.* 2010; Wang *et al.* 2011). In the past, mainly microsatellite markers were used due to their high polymorphism, cross-species transferability as well as being fairly easy and inexpensive to analyse. However, disadvantages such as being prone to scoring errors, complex mutational patterns as well as a high level of null allele occurrence are associated with microsatellites (Glover *et al.* 2010). When selecting markers for linkage mapping purposes, the following needs to be taken into account: 1) even distribution of the markers across the genome, 2) low genotyping error rate, and finally 3) level of polymorphism. These criteria are what have made SNPs popular for the construction of linkage maps. High-throughput SNP discovery and genotyping

methods have also lead to a decrease in time and cost for developing these markers (Ball *et al.* 2010).

In essence, the construction of a linkage map entails finding of a linear arrangement of markers from recombination values. For this purpose, computer packages such as Linkage1 (Suiter *et al.* 1983), GMendel (Echt *et al.* 1992) and MapMaker (Lander *et al.* 1987) are suitable. However, due to the great amount of linkage information becoming available in various organisms for molecular markers, the need for constructing integrated linkage maps has arisen. A program that was specifically created for this purpose is JoinMap (Van Ooijen 2006). Compared to other linkage mapping programs, JoinMap is designed for non-interactive use where the user has no input navigating the process of constructing the maps (in other mapping programs the user has to guide the search, and by inspection find the best fitting order). It performs searches for the best fitting map, while all data are initially considered equally valuable.

One of two mapping algorithms can be selected; the regression mapping algorithm and the maximum likelihood (ML) algorithm (Jansen *et al.* 2001) after linkage groups have been established. These two algorithms should yield more or less the same map order and distances, but the ML algorithm is said to be more robust in the presence of missing data. It is however due to the presence of missing data that the maximum likelihood algorithm can sometimes have inflated map lengths (De Keyser *et al.* 2010). JoinMap does however offer a way to detect doubtfully grouped markers: either by inspecting the Chi-square value after each addition of a new marker to the map (regression mapping) or by examining the plausible positions and "fit and stress" of the markers. When a large jump in the goodness-of-fit value has occurred when a new marker is added, it indicates that the newly added marker may not be part of the specific linkage group it was initially assigned to.

Alternative mapping functions exist that can be used for computing map distances namely Haldane and Kosambi. Stam (1993) defines the map distance between two markers as the mean number of recombination events in that region per meiosis. Map distance is measured in centimorgans (cM), and the relation between recombination frequency and map distance is expressed by a genetic mapping function. Kosambi's mapping function (mf) assumes positive interference, whereas

Haldane's mapping function assumes absence of interference. With positive interference, less double recombinants are expected (Stam 1993). JoinMap uses the map distances computed by either one of these two functions in order to calculate a Chi-square value to determine the goodness-of-fit of the calculated map (Stam 1993).

Currently only a first generation linkage map is available for *Haliotis midae* and it is not yet adequate for QTL mapping due to low marker density as well as uneven marker coverage. The previous linkage map constructed by Jansen (2012) consisted mainly of microsatellites but also contained some SNPs. It comprised of 18 linkage groups, a genome coverage of 65% and an average marker spacing of 9.3 cM. The aim of this study was to use the previously constructed linkage map and to saturate this map with newly developed *in silico* SNPs.

2. Materials and Methods

2.1 MAPPING FAMILIES

Six families were used in the mapping study; four novel families and two of which were previously used for linkage mapping purposes. The four new families originated from the commercial farm Abagold and included FamD, FamH, FamI and FamJ. Linkage maps constructed for these four families consisted only of SNPs developed during the current study (Table 4.1). The two previously used families (Hepple 2010; Blaauw 2012) originated from Roman Bay and included FamDS1 and FamDS2. Linkage maps constructed for these two families consisted of previously developed SNP- and microsatellite markers (Bester *et al.* 2004, Bester *et al.* 2008; Slabbert *et al.* 2008, 2010; Hepple 2010; Rhode *et al.* 2008; Rhode 2010; Blaauw 2012; Jansen 2012; Slabbert *et al.* 2012), as well as SNPs developed during the current study (Table 4.1).

DNA extractions were performed using the CTAB method as described in Chapter 3.

Table 4.1: Families used for linkage map construction (* families included in previous studies).

Family	Origin	Offspring	Marker type
FamD	Abagold	72	SNPs
FamH	Abagold	71	SNPs
FamI	Abagold	81	SNPs
FamJ	Abagold	72	SNPs
FamDS1*	Roman Bay	70	SNPs and microsatellites
FamDS2*	Roman Bay	87	SNPs and microsatellites

2.2 GENOTYPING OF GENE-LINKED MARKERS

2.2.1 EST-derived SNP markers

In total, 192 SNP loci were included in the genotyping assay. Six of these loci were from a previous study (Blaauw 2012) and served as positive SNP controls. Genotyping was performed using a 192-plex Illumina GoldenGate genotyping assay with the VeraCode technology on the BeadXpress platform. The data generated was analysed with GenomeStudio™ Genotyping Model v.1.0. Failed genotypes (no calls or genotypes with GenTrain scores lower than 0.45) as well as monomorphic markers were excluded from downstream analysis.

2.2.2 Genotype data

All SNP genotype data was converted to a JoinMap® v.4.1 format that is suitable for outcrossing populations (CP populations). For SNP markers, three possible genotypes exist (Table 4.2).

Table 4.2: JoinMap[®] v.4.1 genotype data format for CP populations (Van Ooijen 2006).

Code	Description	Possible genotypes
<hkxhk>	Heterozygous in both parents	hh, hk, kk, --
<lmxll>	Heterozygous in first parent*	ll, lm, --
<nnxnp>	Heterozygous in second parent [#]	nn, np, --

* Female parent was chosen as the first parent in families

[#] Male parent was chosen as the second parent in families

-- missing data

2.3 ANALYSIS OF LINKAGE BETWEEN LOCI

Linkage maps were constructed using JoinMap[®] v.4.1. Individuals or markers with more than 20% missing data (failed genotypes etc.) were excluded. The segregation patterns of the informative SNP loci were tested independently in each parent as well as in the offspring from each family by employing the Chi-square goodness-of-fit test in order to establish which markers showed segregation distortion ($p < 0.05$). Markers that displayed segregation distortion were not excluded but were noted when inspecting maps to determine whether these markers clustered together or occurred on the same map location in different families as well as on the integrated map. Sex-specific as well as sex-average maps were constructed.

In order to assign markers to linkage groups, a LOD score (which signifies the likelihood of linkage) of 3 was applied. The LOD, or logarithm of odds, indicates the likelihood that two loci are linked within a set recombination value over the likelihood that they are not linked. A LOD score of 3 indicates a 1000 to 1 odds that loci will be linked for a certain recombination value (meaning the linkage being observed did not occur by chance) (Stam 1993).

For map order construction, regression mapping as well as maximum likelihood mapping algorithms were applied. Regression mapping makes use of a mean Chi-square goodness-of-fit in order to establish whether a good quality map was created. When a jump in the mean Chi-square goodness-of-fit is too large between the additions of new markers, the map was inspected and the marker(s) removed. When using maximum likelihood to create linkage maps, the plausible positions of the markers were examined, as well as the "fit and stress" in order to assess the

quality of the maps (Van Ooijen 2006; 2011). Kosambi's mapping function was used to convert recombination frequencies to mapping distances in centimorgans.

2.4 MAP INTEGRATION

Sex-average linkage maps of the same linkage groups but from different families were combined in order to create integrated linkage maps. Integrated maps were constructed using the regression mapping algorithm as the maximum likelihood mapping algorithm is not yet available for constructing integrated maps.

2.5 GENOME COVERAGE

In order to calculate the map length, the telomeric regions of the linkage group must also be taken into account. With the aim of accomplishing above mentioned, the length of each linkage group can be multiplied by twice the average length from the final marker on a linkage group to the end of the linkage group (Postlethwait *et al.* 1994).

2.5.1 Expected genome length

Average marker spacing (A_S) was calculated by dividing the total length of all the linkage groups by the number of intervals (total number of markers minus total linkage groups) (Fishman *et al.* 2001). Expected genome size was calculated by two equations. The average of the two equations was then used to obtain $G_{e\text{ ave}}$.

Genome Size Estimation 1 (G_{e1})

Genome Size Estimation 1 (G_{e1}) was calculated by multiplying $(k_i+1) / (k_i-1)$ with the length of each linkage group, resulting in this method estimating the average spacing for each chromosome independently (where k_i represents the number of markers at linkage group i) (Chakravarti *et al.* 1991).

Genome Size Estimation 2 (G_{e2})

Genome Size Estimation 2 (G_{e2}) was calculated by adding $2A_S$ (to account for the chromosome ends) to the length of each linkage group resulting in this method estimating the average marker spacing (A_S) of the linkage map on a genome-scale (Fishman *et al.* 2001).

2.5.2 Genome coverage

Genome coverage was subsequently calculated by equation: $GC = G_o / G_{e\text{ ave}}$

Where GC is the genome coverage, G_o is the observed map length and $G_{e\text{ ave}}$ is the average expected genome size.

3. Results

3.1 EST-DERIVED SNP MARKERS

Of the 186 newly developed SNP markers, 44 failed to obtain successful genotypes with the GoldenGate genotyping assay. Of the 142 successful genotypes, 128 were polymorphic and could be used for construction of linkage maps (see Table 3.4). Another 10 of the 128 polymorphic loci had to be excluded due to both parents being homozygotes, making it difficult to determine which alleles of the parents were passed down to the offspring. Of the 118 newly developed *in silico* loci that were used for linkage map construction, only 64 (54.2%) were mapped to the integrated map.

3.2 LINKAGE MAPPING

Sex-average and sex-specific maps were created separately using the 'create population node' and 'create maternal and paternal population node' options in JoinMap[®] v.4.1. All maps were created with the maximum likelihood mapping algorithm. Names of linkage groups were based on the largest (LG_1) to the smallest (LG_18) linkage group when the integrated map was constructed and

names were maintained throughout the separate family maps. When a linkage group resulted in two linkage groups in the integrated maps (caused when markers from the same linkage group could not be linked to each other), "a", "b" or "c" was added to the name to indicate this separation.

Linkage maps and accompanying tables of FamD and FamDS1 are presented in the results section. Linkage maps and accompanying tables of FamH, FamI, FamJ and FamDS2 are presented in addendums 4-11.

3.2.1 Linkage map of family D

Of the 49 markers that were informative in family D, 11 (22.4%) could be mapped to the maternal map (P1), 31 (63.3%) could be mapped to the sex-average map (POP) and 20 (40.8%) could be mapped to the paternal map (P2) (Figure 4.1).

For the sex-average map, the number of markers per linkage group ranged from two to four and the length of the linkage groups ranged from zero to 82.9 cM with an average marker spacing of 13.2 cM (Table 4.3). The genome length, calculated with G_e1 was 654.0 cM and calculated with G_e2 was 765.2 cM. The genome coverage was 47.6%.

The number of markers per linkage group for the maternal map varied between two to three and the length of the linkage groups varied from zero to 11.6 cM with an average marker spacing of 4.2 cM (Table 4.3). The genome length was calculated as 68.2 cM and 70.9 cM with G_e1 and G_e2 , respectively. The genome coverage was computed to be 38.2%.

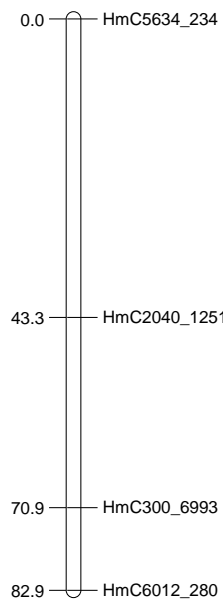
For the paternal map, the length of the linkage groups ranged from zero to 80.7 cM and the number of markers per linkage group ranged from two to four with an average marker spacing of 7.9 cM (Table 4.3). The genome length, determined with G_e1 and G_e2 , respectively was 265.9 cM and 328.2 cM. The genome coverage was 41.9%.

The number of markers that could not be grouped or mapped to the linkage groups of the sex-average, maternal and paternal maps was 15, 35 and 26, respectively.

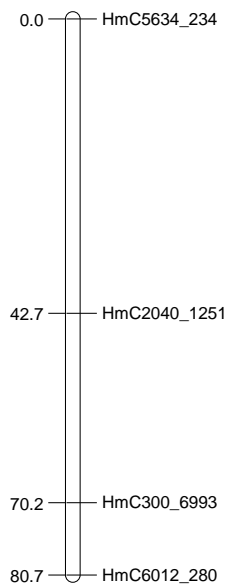
Table 4.3: Number of markers per linkage group, length of linkage groups, average marker spacing and largest interval for maternal (P1), sex-average (POP) and paternal (P2) maps of family D.

Linkage group	No. of markers			Length (cM)			Ave. spacing (cM)			Largest interval (cM)		
	P1	POP	P2	P1	POP	P2	P1	POP	P2	P1	POP	P2
1	-	4	4	-	82.9	80.7	-	27.6	26.9	-	43.3	42.7
2	2	3	2	7.5	41.2	7.5	7.5	20.6	7.5	7.5	33.8	7.5
4	-	4	2	-	80.3	10.6	-	26.8	10.6	-	38.0	10.6
5	-	2	-	-	14.3	-	-	14.3	-	-	14.3	-
6	-	2	2	-	0.1	0.1	-	0.1	0.1	-	0.1	0.1
7	-	2	-	-	15.1	-	-	15.1	-	-	15.1	-
8	3	4	2	11.6	76.1	0.5	5.8	25.4	0.5	8.0	62.6	0.5
11	2	2	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13a	2	2	-	3.0	3.0	-	3.0	3.0	-	3.0	3.0	-
13b	-	2	2	-	10.8	10.8	-	10.8	10.8	-	10.8	10.8
16	-	2	2	-	9.8	9.8	-	9.8	9.8	-	9.8	9.8
17	2	2	2	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5
Total	11.0	31.0	20.0	26.6	338.1	124.5	20.8	158.0	70.7	23.0	235.3	86.5
Average	2.2	2.6	2.2	5.3	28.2	13.8	4.2	13.2	7.9	4.6	19.6	9.6

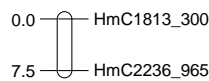
POP_LG_1



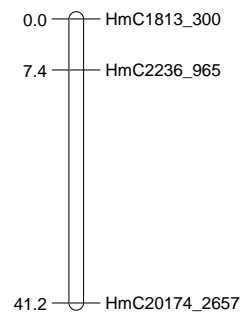
P2_LG_1



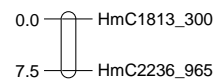
P1_LG_2



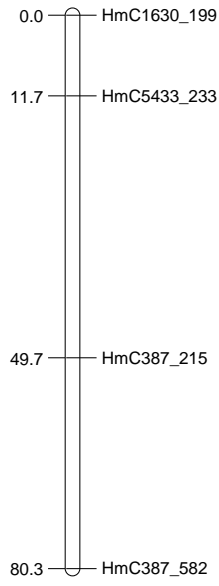
POP_LG_2



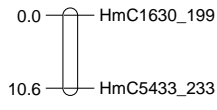
P2_LG_2



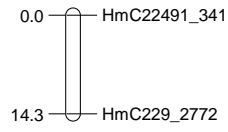
POP_LG_4



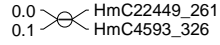
P2_LG_4



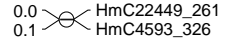
POP_LG_5



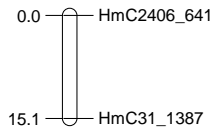
POP_LG_6



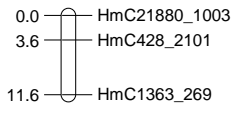
P2_LG_6



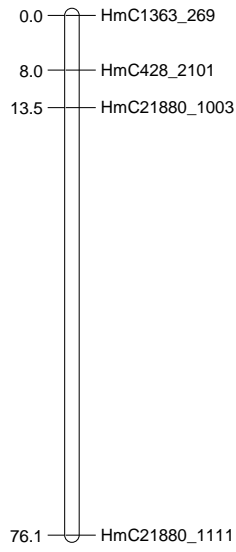
POP_LG_7



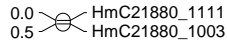
P1_LG_8



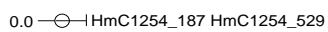
POP_LG_8



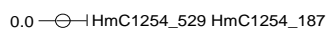
P2_LG_8



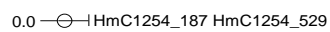
P1_LG_11



POP_LG_11



P2_LG_11



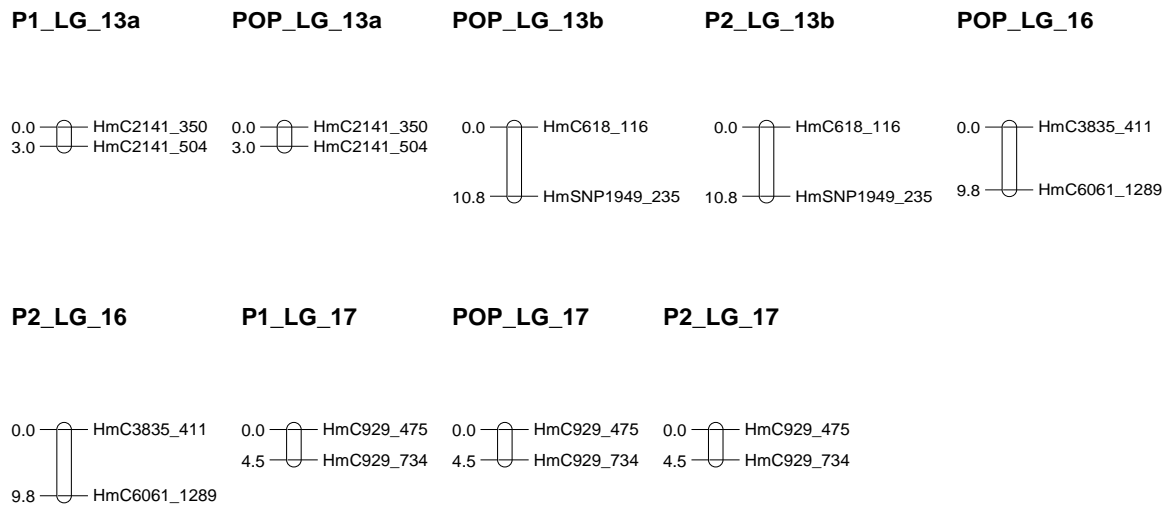


Figure 4.1: Maternal (P1), sex-average (POP) and paternal (P2) maps of family D.

3.2.2 Linkage map of family DS1

Of the 165 markers (66 of which were newly developed *in silico* SNP markers) that were informative in family DS1, 88 (53.3%) could be mapped to the maternal map (P1), 145 (87.9%) could be mapped to the sex-average map (POP) and 111 (67.3%) could be mapped to the paternal map (P2) (Figure 4.2).

Considering the sex-average map, the number of markers per linkage group and the length of the linkage groups ranged from two to 18 and 4.5 cM to 274.2 cM, respectively. An average marker spacing of 14.4 cM was determined (Table 4.4). The genome length, calculated with G_e1 was 2128.9 cM and calculated with G_e2 was 2039.5 cM. The genome coverage was estimated to be 73.1%.

Taking the maternal map into account, the number of markers per linkage group varied from two to 12, the length of the linkage groups varied from 3.1 cM to 169.5 cM and an average marker spacing of 14.0 cM was computed (Table 4.4). The genome length, calculated with G_e1 and G_e2 was 1431.2 cM and 1360.2 cM, respectively. The genome coverage was 58.5%.

For the paternal map, the number of markers per linkage group ranged from two to 15 and the length of the linkage groups ranged from 4.6 cM to 117.5 cM with an

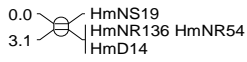
average marker spacing of 8.5 cM (Table 4.4). The genome length, calculated with G_{e1} and G_{e2} was 961.4 cM and 900.5 cM, respectively. The genome coverage was 68.5%.

The number of markers that could not be grouped or mapped to the linkage groups of the sex-average, maternal and paternal maps were 21, 78 and 55, respectively.

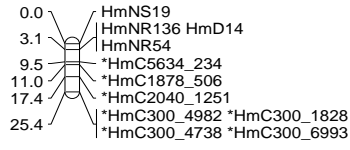
Table 4.4: Number of markers per linkage group, length of linkage groups, average marker spacing and largest interval for maternal (P1), sex-average (POP) and paternal (P2) maps of family DS1.

Linkage group	No. of markers			Length (cM)			Ave. spacing (cM)			Largest interval (cM)		
	P1	POP	P2	P1	POP	P2	P1	POP	P2	P1	POP	P2
1	4	11	11	3.1	25.4	39.8	1.0	2.5	4.0	3.1	8.0	13.9
2	6	6	3	169.5	167.5	5.2	33.9	33.5	2.6	100.0	103.5	5.2
3	3	4	3	37.4	44.1	35.6	18.7	14.7	17.8	30.8	28.6	25.5
4	4	12	6	44.9	274.2	58.2	15.0	24.9	11.6	24.7	136.7	27.6
5	12	18	14	130.2	102.7	71.4	11.8	6.0	5.5	32.2	20.7	18.4
6	5	8	8	81.5	92.8	76.0	20.4	13.3	10.9	63.0	32.2	43.6
7a	4	7	4	29.0	101.2	17.3	9.7	16.9	5.8	15.5	34.9	8.2
7b	2	-	-	3.2	-	-	3.2	-	-	3.2	-	-
8a	6	17	15	4.7	58.5	39.6	0.9	3.7	2.8	2.3	27.8	12.6
8b	5	-	-	9.2	-	-	2.3	-	-	4.4	-	-
9	2	3	3	25.4	27.5	16.1	25.4	13.8	8.1	25.4	24.4	14.6
10a	3	5	9	20.6	101.6	117.5	10.3	25.4	14.7	17.5	67.6	42.9
10b	3	6	-	7.8	30.8	-	3.9	6.2	-	7.8	18.6	-
10c	2	-	-	51.4	-	-	51.4	-	-	51.4	-	-
12	5	6	3	54.0	164.3	26.6	13.5	32.9	13.3	18.8	111.6	17.1
13	8	12	8	43.3	29.0	7.7	6.2	2.6	1.1	25.2	13.5	3.1
14	-	4	4	-	16.2	12.4	-	5.4	4.1	-	8.1	6.2
15a	-	4	3	-	52.7	10.8	-	17.6	5.4	-	29.1	9.6
15b	-	-	2	-	-	29.9	-	-	29.9	-	-	29.9
16	2	4	4	3.0	53.6	15.6	3.0	17.9	5.2	3.0	40.6	11.1
17	-	3	3	-	40.1	13.7	-	20.1	6.9	-	27.9	10.7
18a	2	8	6	23.0	89.7	39.3	23.0	12.8	7.9	23.0	27.6	20.4
18b	3	-	-	9.0	-	-	4.5	-	-	5.2	-	-
22	2	-	-	23.5	-	-	23.5	-	-	23.5	-	-
23	3	3	-	32.9	37.8	-	16.5	18.9	-	23.9	29.7	-
24	2	2	-	9.5	9.3	-	9.5	9.3	-	9.5	9.3	-
25	-	2	2	-	4.5	4.6	-	4.5	4.6	-	4.5	4.6
Total	88.0	145.0	111.0	816.1	1523.5	637.3	307.6	302.7	162.0	513.4	804.9	325.2
Average	4.0	6.9	5.8	37.1	72.5	33.5	14.0	14.4	8.5	23.3	38.3	17.1

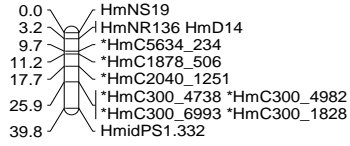
P1_LG_1



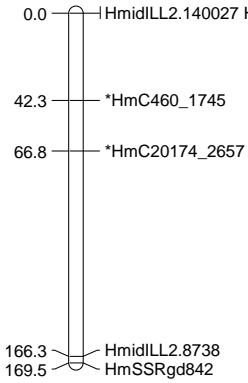
POP_LG_1



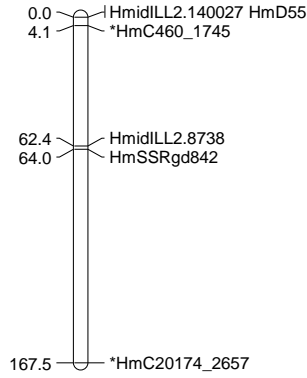
P2_LG_1



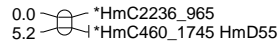
P1_LG_2



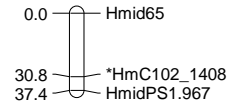
POP_LG_2



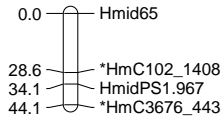
P2_LG_2



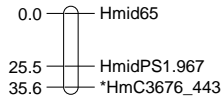
P1_LG_3



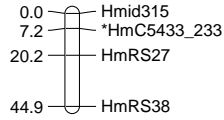
POP_LG_3



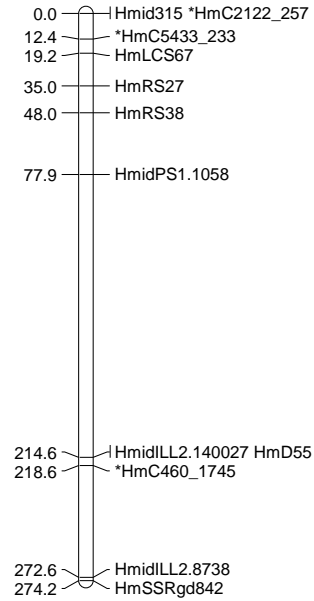
P2_LG_3



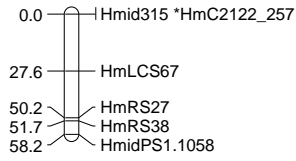
P1_LG_4



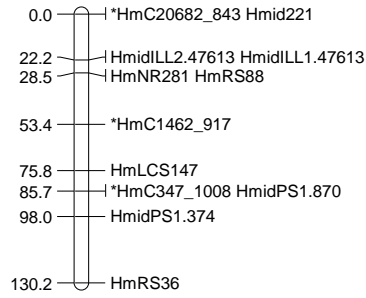
POP_LG_4



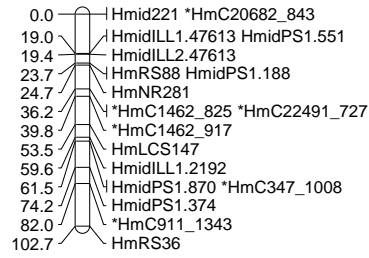
P2_LG_4



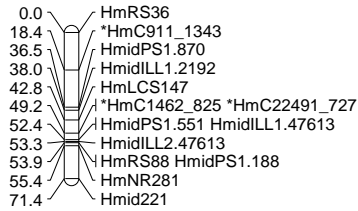
P1_LG_5



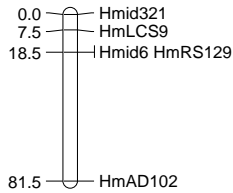
POP_LG_5



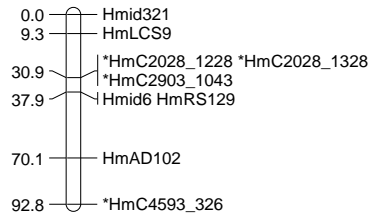
P2_LG_5



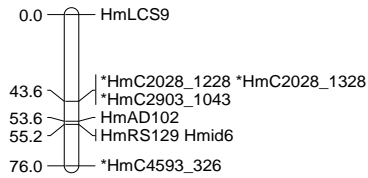
P1_LG_6



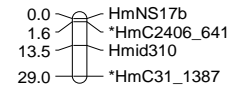
POP_LG_6



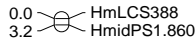
P2_LG_6



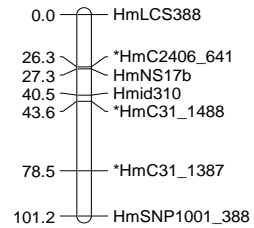
P1_LG_7a



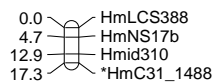
P1_LG_7b



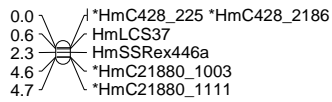
POP_LG_7a



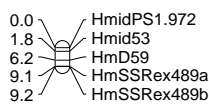
P2_LG_7a



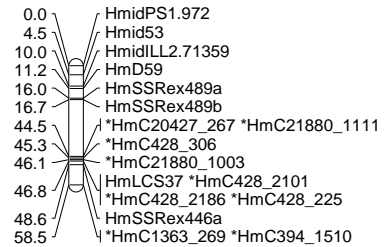
P1_LG_8a



P1_LG_8b



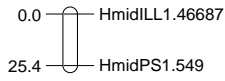
POP_LG_8a



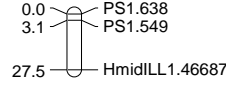
P2_LG_8a



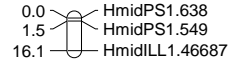
P1_LG_9



POP_LG_9



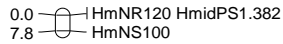
P2_LG_9



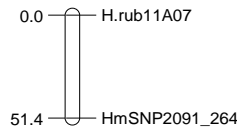
P1_LG_10a



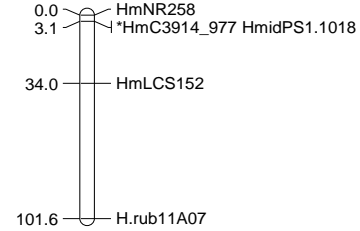
P1_LG_10b



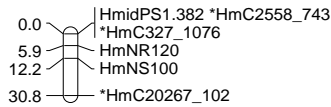
P1_LG_10c



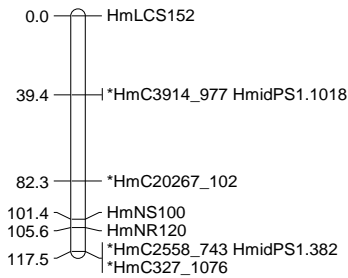
POP_LG_10a



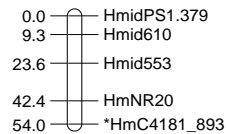
POP_LG_10b



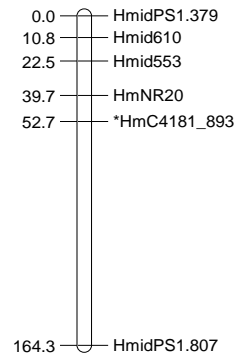
P2_LG_10a



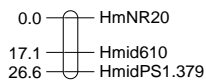
P1_LG_12



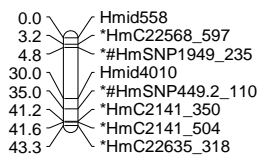
POP_LG_12



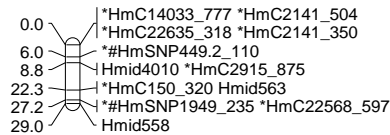
P2_LG_12



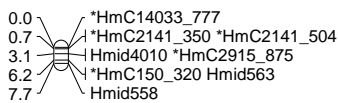
P1_LG_13



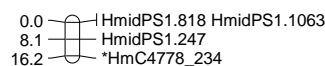
POP_LG_13



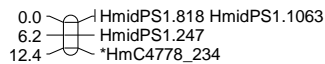
P2_LG_13



POP_LG_14



P2_LG_14



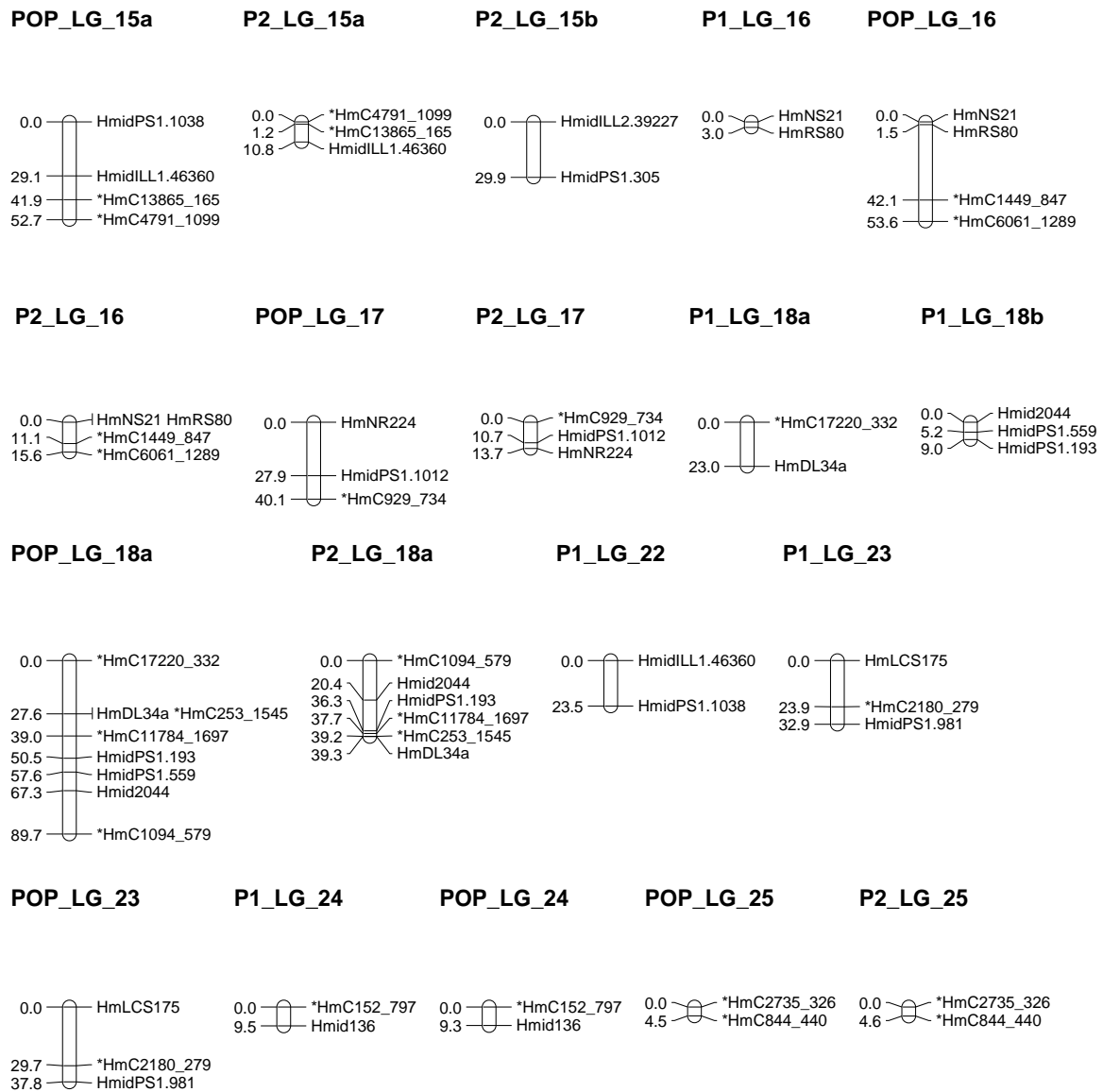


Figure 4.2: Maternal (P1), sex-average (POP) and paternal (P2) maps of family DS1.

3.2.3 Integrated linkage map

Of the 314 markers (118 of which were newly developed *in silico* SNP markers) that were informative in all the families, 186 (59.2%) could be mapped to the integrated map. The integrated map consisted of SNP markers developed in the current study as well as markers (microsatellites and SNPs) developed in previous studies (Bester *et al.* 2004, Bester *et al.* 2008; Slabbert *et al.* 2008, 2010; Hepple 2010; Rhode *et al.* 2008; Rhode 2010; Blaauw 2012; Jansen 2012; Slabbert *et al.* 2012) (Figure 4.3).

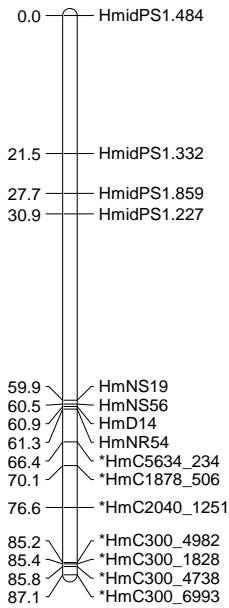
The number of markers per linkage group ranged from three to 20 and the length of the linkage groups ranged from 1.3 cM to 87.1 cM with an average marker spacing of 6.9 cM (Table 4.5). The genome length, calculated with G_e1 and G_e2 , respectively was 1326.65 cM and 1296.88 cM. The genome coverage was 79.1%.

Due to insufficient linkage and inadequate Chi-square values (Regression mapping algorithm), 128 of the markers could not be grouped or mapped to linkage groups.

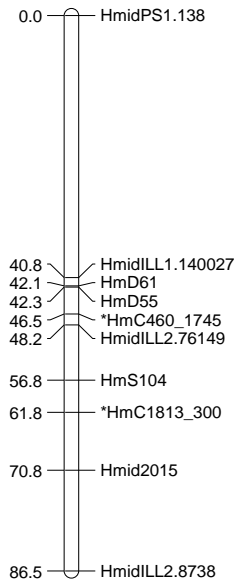
Table 4.5: Number of markers per linkage group, length of linkage groups, average marker spacing and largest interval for integrated map.

Linkage group	No. of markers	Length (cM)	Ave. spacing (cM)	Largest interval (cM)
1	15	87.1	6.2	29.0
2	10	86.5	9.6	40.8
3	5	67.5	16.9	24.9
4	10	64.7	7.2	21.4
5	19	64.3	3.6	16.1
6	12	64.0	5.8	19.0
7	9	62.5	7.8	16.2
8	20	62.0	3.3	11.9
9	8	58.7	8.4	17.7
10	10	51.0	5.7	9.8
11	6	50.9	10.2	27.8
12	6	46.4	9.3	17.1
13	13	43.2	3.6	10.3
14	7	42.6	7.1	10.8
15	7	33.6	5.6	10.2
16	10	29.3	3.3	10.6
17	3	15.2	7.6	11.9
18a	4	1.3	0.4	0.6
18b	6	64.0	12.8	25.5
18c	6	36.8	7.4	20.3
18d	3	5.9	3.0	3.6
Total	189	1037.5	144.6	355.5
Average	9.0	49.4	6.9	16.9

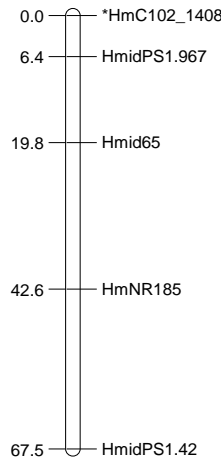
INT_LG_1



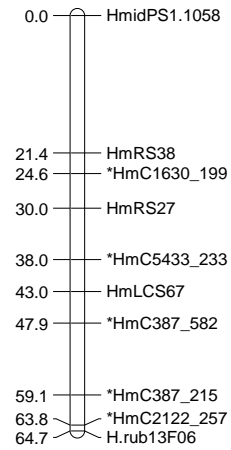
INT_LG_2



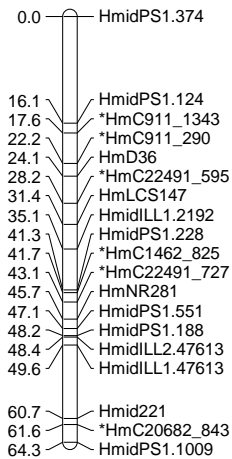
INT_LG_3



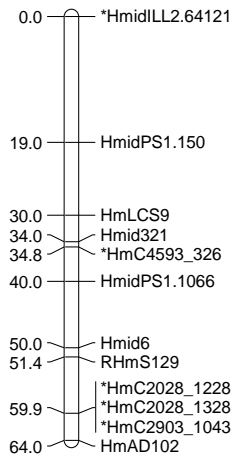
INT_LG_4



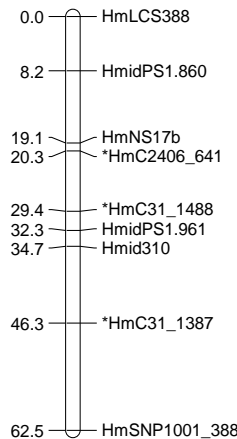
INT_LG_5



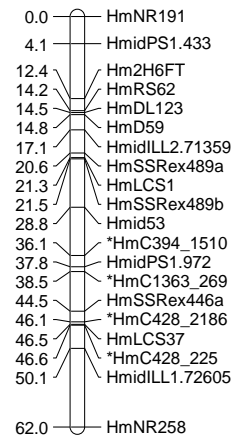
INT_LG_6



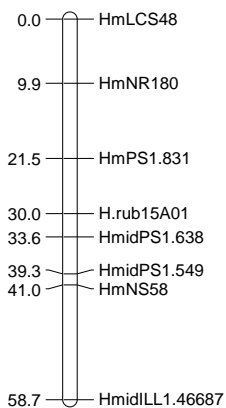
INT_LG_7



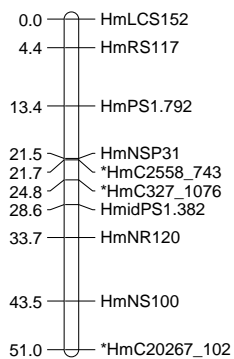
INT_LG_8



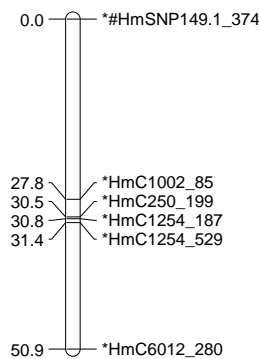
INT_LG_9



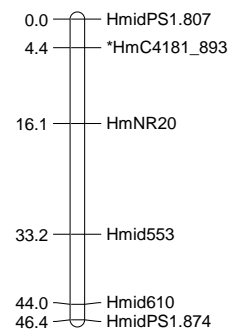
INT_LG_10



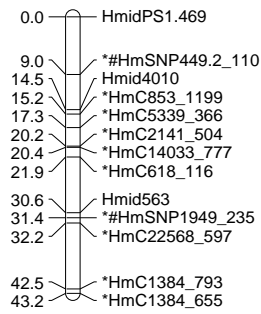
INT_LG_11



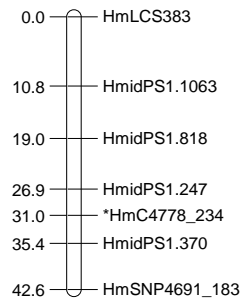
INT_LG_12



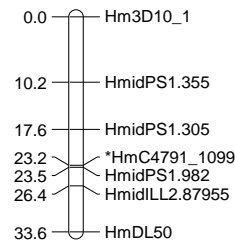
INT_LG_13



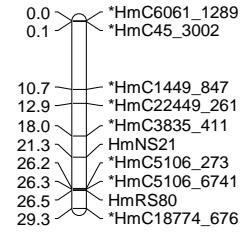
INT_LG_14



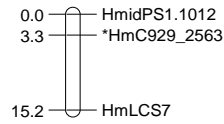
INT_LG_15



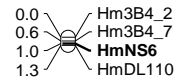
INT_LG_16



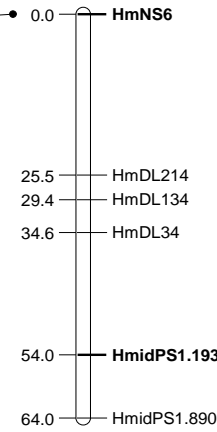
INT_LG_17



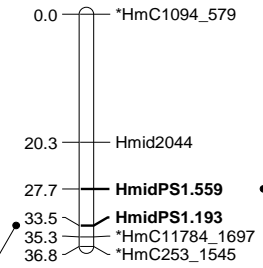
INT_LG_18a



INT_LG_18b



INT_LG_18c



INT_LG_18d



Figure 4.3: Integrated map [* indicates SNP markers developed in current study, # indicates positive SNP controls developed by Blaauw (2012)].

3.2.4 Linkage group one (LG_1) comparison

Comparisons were not made across all linkage groups due to the large number of families used. LG_1 is used as an illustration of conserved marker order between a sex-average map and the integrated map. This linkage group is also used to indicate that a larger number of markers mapped to FamDS1 (comprised of newly and previously developed markers) compared to FamD (only contained newly developed SNP markers), resulting in a difference in marker spacing and observed length between FamD and FamDS1. Furthermore, compared to the integrated map less markers mapped to the linkage groups of the two respective sex-average family maps; a trend observed for all the linkage groups in all mapping families. LG_1 was chosen for the comparison of abovementioned due to the observed length of the linkage group (longest linkage group).

In comparison with the integrated map, LG_1 of FamD consisted of only four markers, whilst LG_1 of FamDS1 contained 11 markers. Overlapping loci are highlighted in different colours and connected with a line. Overlapping loci (anchor loci) are in the same order on the three maps.

Table 4.6: LG_1: Number of markers per map, lengths of LG_1, average marker spacing and largest interval.

(LG_1)	No. of markers	Length (cM)	Ave. spacing (cM)	Largest interval (cM)
FamD	4	82.9	27.6	43.3
Integrated map	15	87.1	6.2	29.0
FamDS1	11	25.4	2.5	8.0

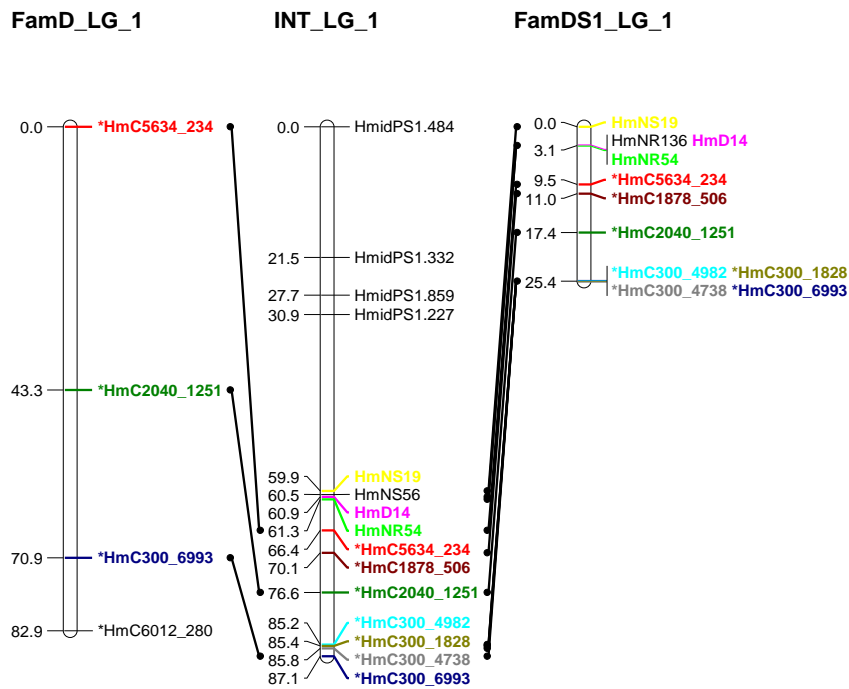


Figure 4.4: LG_1: A comparison of marker order and marker density of sex-average maps from family D (SNPs only) and DS1 (SNPs and microsatellites) with the integrated map.

4. Discussion

This study focused on saturating the first generation linkage map for *Haliotis midae* with 118 newly developed SNP markers. Previous linkage maps constructed by Hepple (2010) and Jansen (2012) were mostly saturated with microsatellite markers and contained only a few SNPs. Due to the error rate of microsatellite markers that cause various problems for linkage mapping including map inflation and marker order ambiguity (Ball *et al.* 2010), a shift has been recognised from microsatellite markers to SNPs in many species for linkage map construction. The abundance of SNPs as well as their high-throughput discovery and genotyping have also contributed to their popularity as molecular markers for linkage mapping studies.

Using a total of 314 informative molecular markers, a linkage map for *H. midae* was constructed which included 178 microsatellite markers and 136 SNPs (including the 118 markers developed during the current study). Of these 314 DNA markers, 186 mapped to the integrated map which was constructed by combining the sex-average

maps from each family and were arranged into 18 linkage groups (not counting "a", "b", "c" and "d" separately, but as one linkage group where applicable). It has been established that the haploid chromosomal number for *H. midae* is 18 (Van der Merwe & Roodt-Wilding 2008), therefore the anticipated number of linkage groups per mapping family is also 18.

For map order construction, regression mapping as well as maximum likelihood (ML) mapping algorithms were applied. The ML mapping algorithm proved to be adequate (if not better) for number of markers mapped as well as determining marker order. Only in limited instances where a small number of markers mapped or problems were experienced with regards to marker order, was the regression mapping algorithm considered. When constructing the integrated maps only the regression mapping algorithm was used as the necessary algorithm for creating integrated maps with the maximum likelihood algorithm has not yet been developed. Although ML mapping is computationally more demanding and slower than regression mapping, it is more accurate and in the presence of missing data it is also more robust as it uses nearby markers to estimate the missing genotypes by taking possible recombinations into consideration (De Keyser *et al.* 2010).

Although the integrated map contained 18 linkage groups, the separate family maps rarely displayed this specific number and the number of linkage groups ranged from nine to 22. FamD, H, I and J only had 11, nine, ten and nine linkage groups, respectively (not counting "a" and "b" separately, but as one linkage group where applicable) with only 63.3%, 50.0%, 63.0% and 53.2% of the informative markers mapping to the sex-average maps. Due to SNP markers being bi-allelic these markers have to be heterozygous in both or at least one of the linkage mapping family parents otherwise they contain no valuable information for determining linkage between markers. Therefore, although a relatively large number of SNPs was developed for use in this linkage mapping study (142 successful genotypes), a great deal (24) had to be excluded from further analysis due to either both parents being homozygous for different alleles or the SNP marker being monomorphic in all the families used. Because the linkage maps constructed for FamD, H, I and J only contained SNP markers, the map density was very low making it difficult to determine linkage between markers and mapping all the available markers. There

were also not enough markers to cover all 18 of the chromosomes. According to Liu *et al.* (2006), this is a regular phenomenon when constructing preliminary maps.

As mentioned before, FamDS1 and DS2 contained markers (microsatellites and SNPs) that were developed in previous studies. These two families along with the previously developed markers were included in the current study to help integrate the newly developed SNP markers on the linkage maps. The two families contained 21 (DS1) and 22 (DS2) linkage groups each (not counting "a", "b" and "c" separately, but as one linkage group where applicable) with 87.9% and 81.4% of the informative markers mapping to the sex-average maps. A possible reason for the linkage groups amounting to more than 18 for both families could be due to markers that are located too far apart from each other (50 cM or more, Miké 1977) making it impossible to obtain linkage information and subsequently link markers with one another (Pérez *et al.* 2004; Chistiakov *et al.* 2005). Consequently, the formation of two or more linkage groups as opposed to only one is the result of insufficient marker density that could be circumvented if more markers are available to map. Since more markers were included in constructing linkage maps for these two families relative to the other families, it explains the higher number of linkage groups per family as well. Also, due to the highly polymorphic nature of microsatellite markers, they contain more information per family with regards to the segregation of alleles (Hauser & Seeb 2008) than bi-allelic SNPs. The higher percentage of mapped markers for DS1 and DS2 can therefore be explained as well by the fact that a significantly larger number of SNPs are needed in order to reach the same level of information content supplied by microsatellites (Schaid *et al.* 2004).

Some linkage groups proved to be problematic and had to be designated "a", "b", "c" or "d". This occurred either when the separate family maps were constructed or when the sex-average maps were combined in order to create the integrated map, and markers selected as anchor loci were used to link the linkage groups together (Table 4.7). One such linkage group that proved to be challenging was INT_LG_18. Four different integrated maps containing different markers could be drawn for this linkage group, but with every map a different marker existed that could connect the four maps (HmNS6: INT_LG_18a & INT_LG_18b; HmidPS1.193: INT_LG_18b & INT_LG_18c; HmidPS1.559: INT_LG_18c & INT_LG_18d) (Figure 4.3). A possible

reason for LG_18 resulting in different groups in the families could be due to no anchor loci existing for this linkage group. Therefore no marker exist that can help "integrate" map "a" "b" "c" and "d" to form only one map. It is important to identify anchor loci in order to integrate additional markers on a linkage map. Markers developed from EST sequences are extremely useful in identifying anchor loci as these markers are derived from more conserved genomic regions and are expected to be more transferable to other mapping populations (Brown *et al.* 2001; Studer *et al.* 2010).

Table 4.7: Anchor loci informative in four or more families.

Linkage group	Anchor loci
1	HmD14, HmNR54, HmNS19, *HmC2040_1251, *HmC300_6993
2	HmidILL1.140027, HmD55, HmidILL2.8738
3	Hmid65
4	HmidPS1.1058, *HmC5433_233, *HmC387_582, *HmC387_215
5	HmidPS1.374, HmidPS1.228, HmidPS1.551, HmidILL1.47613, Hmid221
6	HmRS129
7	HmLCS388, HmidPS1.860, Hmid310
8	*HmC1363_269, *HmC428_2186, HmLCS37, *HmC428_225
9	HmidPS1.638, HmidPS1.549
10	HmNR120, HmNS100
11	*HmC1254_187, *HmC1254_529
12	HmNR20, Hmid553, Hmid610, HmidPS1.874
13	HmSNP449.2_110, Hmid4010, *HmC2141_504, Hmid563, HmSNP1949_235
14	HmidPS1.1063, HmidPS1.818, HmidPS1.247
15	-
16	HmNS21, HmRS80
17	-
18	-

* Indicates *in silico* SNP markers developed in current study.

When comparing FamD, FamDS1 and the integrated map, the marker order of LG_1 is maintained although not all markers are present on the different maps. Considering that it is to be expected that less markers will map to the "SNP-only" map of FamD (bi-allelic SNPs provide less segregation information), it demonstrates that SNPs are also reliable markers to use for linkage map construction and determining marker order on a chromosome. Taking the marker spacing into

account, the average spacing of markers found on LG_1 of FamD (27.6 cM) is not ideal and this problem can only be addressed by including more markers. The small number of markers that mapped to LG_1 of FamD is also responsible for the extremely long observed length (compared to LG_1 of INT map and FamDS1 and taking into consideration number of markers mapped) as it has been shown that a genetic map may be inflated by low marker density (Yu & Guo 2003). As genetic maps then shorten with increased marker density, developing more markers can also address this problem. It is evident that the inclusion of microsatellites and SNPs are more effective for linkage map construction than only SNPs (Figure 4.4). It was found that a linkage map constructed with only SNPs or microsatellite markers contained less markers than when both types of markers were included (highly polymorphic microsatellite loci provide more information regarding segregation).

The integrated map contained an average marker spacing of 6.9 cM that is sufficient for QTL detection (Massault *et al.* 2008 estimated 10 cM to be adequate). However, this spacing is not uniform across all the markers and the large intervals between some loci on the map poses a problem for QTL mapping. It is easier to map a QTL in an interval of defined genetic distance due to recombination events than being at a minimum. Thus, the accuracy of identifying a QTL is dependent on the number of recombination events occurring between markers (which are less the smaller the genetic distance is between them) (Doerge 2002). For the individual families, the average marker spacing for the sex-average maps ranged between 11.7 cM and 34.1 cM. The family maps however did not contain as many markers per linkage group as the integrated map, and some linkage groups even contained only two markers; leaving a large gap between markers. To address this problem, more markers need to be developed in order to saturate the integrated map of *H. midae* for QTL studies.

When markers were prepared for linkage map construction, all informative markers were tested for segregation distortion by making use of a Chi-square test. Distorted markers displayed a p-value of less than 0.05 but were still included in linkage analysis as distorted markers can sometimes aid in QTL mapping (Zhan & Xu 2011). After linkage maps were constructed, only some of the distorted loci included could be mapped as the other did not group to any linkage groups. Segregation distortion

refers to an occurrence where the observed genotypic frequencies differ considerably from the expected Mendelian frequencies (Sandler *et al.* 1959). According to Charlesworth and Charlesworth (1998) it is expected to see distorted markers cluster together on chromosomes as it can possibly indicate viability selection, but in the current study that was not the case. The only instance where two distorted markers mapped close together was in family DS1 at LG_2. Both microsatellite markers (HmSSRgd842 and HmidILL2.8738) were distorted in more than one family (FamDS1 and FamDS2), but marker HmSSRgd842 only mapped to LG_2 in one of the families (FamDS1). HmidILL2.8738 did however also map to the integrated map, so it is important to note that the map distance between Hmid2015 and HmidILL2.8738 may be inaccurate. All families except for FamI displayed distorted SNP markers, and it is important to keep those markers in mind when marker order on a map is inspected. A SNP marker that was distorted in FamD, but not in FamJ was HmC387_215 (LG_4). This marker did not map to the same position in FamD as it did in FamJ and the integrated map, so this should be taken into account when examining marker distances on that particular linkage group. It is also possible that markers exhibiting segregation distortion may actually point towards genotyping or scoring errors. Ball *et al.* (2010) calculated that an error rate of 5% may cause up to 50% map inflation while influencing map lengths and marker orders. In this study, GenTrain and GenCall cutoff values were employed during genotyping in order to minimise errors associated with genotyping (see chapter 3).

Differences in recombination rates between sexes have been observed for a number of fish species with the female maps frequently showing larger map distances than the male maps (Wang *et al.* 2011). Although this phenomenon is not yet completely understood, various factors including transcriptional activity of certain genes during meiosis, the presence of sequences recognised by sex-specific enzymes and differences observed between sexes in the time spent in meiotic prophase has been proposed to influence this observation (Wang *et al.* 2004; Baranski *et al.* 2006). While the molecular mechanisms responsible for different recombination rates between sexes is not yet fully comprehended, it has been suggested that many factors including pericentromeric suppression, GC content, LINE and SINE elements, CpG islands, polyA/polyT content, simple repeats and other sequence features could influence these rates (Xia *et al.* 2010). For FamD, FamH, FamI and

FamJ this however was not the situation. The maternal map lengths ranged from 26.6 cM to 106.2 cM (Table 4.3; Addendums 4,6,8) and the paternal map lengths ranged from 76.4 cM to 124.5 cM (Table 4.3; Addendums 4,6,8), with the male maps of FamD, FamI and FamJ being longer than their respective female maps. The genome lengths for the maternal and paternal maps confirmed this observation with only FamH having a larger female than male map. A possible reason for this observation could be due to only SNPs being used for map construction of these families. Very few of the available SNP markers for each family could be mapped to the maternal and paternal maps, making these preliminary maps unreliable for determining sex-specific recombination rates. The sex-specific maps of families DS1 and DS2 did however conform to this trend. The female map of family DS1 was 816.1 cM (Table 4.4)(genome length, G_e1 : 1431.2 cM and G_e2 : 1360.2 cM) compared to the male map that was 637.3 cM (Table 4.4)(genome length, G_e1 : 961.4 cM and G_e2 : 900.5 cM), and the female map of family DS2 was 1824.2 cM (Addendum 10)(genome length, G_e1 : 2402.6 cM and G_e2 : 1763.6 cM) compared to 1075.0 cM (Addendum 10)(genome length, G_e1 : 2682.6 cM and G_e2 : 1777.9 cM) of the male map. This could be attributed to the fact that SNPs and microsatellites were used to construct the maps of these two families, yielding more information regarding the segregation of markers and ultimately mapping more markers per linkage group (higher marker density results in more accurately estimated genome lengths, Yu & Guo 2003). This corresponds to map lengths observed in other fish (*Salmo salar*, *S. trutta*) and other halitid species (*Haliotis discus hannai*, *H. diversicolor*, *H. rubra*) (Table 4.8) where the female maps were also longer than the male maps (Baranski *et al.* 2006; Liu *et al.* 2006; Lien *et al.* 2011).

Compared to linkage maps constructed for other halitid species [*H. rubra* (Baranski *et al.* 2006), *H. diversicolor* (Shi *et al.* 2010; Zhan *et al.* 2011), *H. discus hannai* (Liu *et al.* 2006)] the linkage map of *H. midae* is the only one containing SNP markers. The first linkage map (Badenhorst 2008) constructed for *H. midae* consisted only of AFLPs, but subsequent maps also contained microsatellite- and SNP markers (Hepple 2010; Jansen 2012). The current map is the most recent map constructed for perlemon, and also consists of microsatellites and SNPs.

Table 4.8: Examples of linkage maps constructed for fish and shellfish species.

Species	Map length (cM)*	Mapped markers	Linkage groups*	Reference
<i>Salmo salar</i>	2402.3 / 1746.2	5650	29	Lien <i>et al.</i> 2011
<i>Lates calcarifer</i>	2411.5	822	24	Wang <i>et al.</i> 2011
<i>Haliotis diversicolor</i>	758.3 / 676.2	175	16	Zhan <i>et al.</i> 2011
<i>Gadus morhua</i>	1421.92	924	23	Hubert <i>et al.</i> 2010
<i>Haliotis diversicolor</i>	2152.8 / 2032.7	90 / 94	17 / 18	Shi <i>et al.</i> 2010
<i>Ctenopharyngodon idella</i>	1176.1	279	24	Xia <i>et al.</i> 2010
<i>Haliotis rubra</i>	766 / 621	98 / 102	20 / 17	Baranski <i>et al.</i> 2006
<i>Salmo trutta</i>	912 / 346	288	37	Gharbi <i>et al.</i> 2006
<i>Haliotis discus hannai</i>	1774 / 1366	119 / 94	22 / 19	Liu <i>et al.</i> 2006

* Female map length / linkage group indicated first where two map lengths / linkage groups are shown.

It however includes more SNPs than the previous map as well as more mapping families and this can explain the considerably higher genome coverage of the current integrated map (79.1%) as opposed to the 65% which was obtained in the map previously constructed by Jansen (2012). The average marker spacing also improved from 9.3 cM (Jansen 2012) to 6.9 cM. When comparing the map of *H. midae* to an integrated map constructed for *H. diversicolor* (Zhan *et al.* 2011), the genome coverage was approximately the same (79.1% *H. midae*, 80.7% *H. diversicolor*), while the average marker spacing for *H. diversicolor* was better (4.6 cM). This could possibly be due to the map for *H. diversicolor* saturated with microsatellites (more informative than SNPs) thus leading to more markers mapped per chromosome and therefore containing improved average marker spacing.

In conclusion, SNPs were found to be informative markers and are recommended for linkage map construction. However, due to their bi-allelic nature, a large number of SNP markers should be included in linkage mapping studies. Due to microsatellites being more informative than SNPs and SNPs being easier to develop than microsatellites it is advantageous to use SNPs and microsatellites in the construction of a linkage map as the two markers integrate well and results in a higher number of mapped markers when used in conjunction than when used separately. A way to ensure that less SNP markers are excluded from linkage analysis as a result of being non-informative is to initially genotype the marker in the parents to determine

informativeness in the mapping families (SNPs have to be heterozygous in both or at least one of the parents). That way more informative markers can be included in linkage map construction without having to discard monomorphic markers. For QTL detection it is also more valuable to employ fewer larger families than a greater number of families with less offspring (Massault *et al.* 2008). It is evident that the development of more markers will aid in the difficulties associated with linkage map construction of a non-model species. Not only will it aid in identifying anchor loci and provide higher coverage of the genome, but it will also lead to a decrease in the average marker spacing which will be beneficial for future QTL mapping and marker-assisted selection.

Chapter Five

Conclusions and Future Considerations

1. Introduction

During this study the success of developing *in silico* SNPs from next-generation sequencing data was investigated by applying specific criteria and assigning quality scores for identifying and genotyping putative SNPs. Identification of SNPs was performed using CLC Genomics Workbench and genotyping was conducted with the Illumina GoldenGate genotyping assay with the VeraCode technology on the BeadXpress platform. Successfully genotyped SNPs were subjected to segregation analysis using the mapping software JoinMap for the construction of a more dense linkage map for *Haliotis midae*. As previous maps constructed by Badenhorst (2008), Hepple (2010) and Jansen (2012) contained either AFLP markers, microsatellite markers and only few SNPs, it was necessary to develop more SNP markers to use in conjunction with previously developed SNPs and microsatellites in an attempt to saturate the genetic linkage map of *H. midae* for future QTL mapping and ultimately marker-assisted selection.

2. Marker development and validation

Single nucleotide polymorphisms are abundant and widespread in genomes; making these markers ideal for use in genetic applications. Due to the bi-allelic nature of these polymorphisms, they are simple to score and can also easily be subjected to high-throughput genotyping (Garvin *et al.* 2010).

With the introduction of next-generation sequencing technologies, developing SNPs has become more feasible due to the decrease in time and costs associated with marker development. Next-generation sequencing technologies include pyrosequencing, sequencing-by-synthesis, sequencing-by-ligation and single-molecule sequencing. Various platforms exist for these sequencing technologies, but the two most commonly used in species with no or little genomic information are Roche 454 FLX Titanium and Illumina Genome Analyser II (pyrosequencing and sequence-by-synthesis, respectively). In the current study, the Illumina Genome Analyser II sequence-by-synthesis method was employed which resulted in a large amount of sequenced data. Drawbacks such as high error rate, short read length

and shallow coverage are however associated with non-model organisms and the lack of a reference genome complicate the assembly of generated reads, as was also observed with the assembled transcriptome of *H. midae*. Nevertheless, as current technologies continue to improve (as they undoubtedly will) error rates will decrease and longer read lengths will be produced (average read lengths produced in current study: 260.62 bp). Associated costs will also decrease as technology advances. Accompanied by improved data analysis algorithms and computing capacity, higher quality assemblies of sequenced data will be feasible (Ekblom & Galindo 2010).

In order to obtain deep assemblies of redundant contigs (necessary for identifying SNPs), a genome reduction step is required for organisms without a reference genome and transcriptome sequencing has been shown to be one of the most frequently used reduction methods (Seeb *et al.* 2011a). In the current study and other similar studies (for example Wang *et al.* 2008; Lepoittevin *et al.* 2010; Milano *et al.* 2011), it has been shown that ESTs provide a rich resource for identifying SNPs in non-model species. An advantage associated with *in silico* SNP identification from transcribed sequences is that no additional bench work is required, only bioinformatic analysis; thus contributing to the time- and cost effectiveness of developing these markers. As the availability of sequenced transcriptomic data continues to increase, the identification of gene-linked SNPs will also increase as will the utility of these markers in applications such as parentage assignment, population genetic studies, comparative studies, QTL mapping and MAS.

In order to be able to use *in silico* identified SNPs in downstream applications, a validation step is required. To recognise the polymorphic state of the developed markers in different individuals, various genotyping platforms exist that need to be considered in terms of cost, accuracy, equipment, difficulty of assay, throughput and multiplexing capacity. During this study the GoldenGate genotyping assay proved successful for this validation step. This platform is relatively flexible regarding the number of loci genotyped (48 - 384), does not require a large amount of preparation and demonstrated a high genotyping success rate in the current study (76.34%) as well as previous studies where it was used (69%, Wang *et al.* 2008; 66.9%, Lepoittevin *et al.* 2010; 85.4%, Blaauw 2012; 87%, Grattapaglia *et al.* 2011).

When a larger number of SNPs need to be genotyped (>384), Illumina also provides the GoldenGate genotyping assay with the BeadArray Reader. This platform allows for plex levels of up to 3072 SNP loci; the highest multiplex levels offered by Illumina's GoldenGate genotyping assay. For even higher plex levels, the Infinium iSelectHD can be employed. This platform can genotype from 3072 to up to one million markers per sample. This technology however makes use of a BeadChip which only exists for a limited number of species (Illumina 2012).

Overall it can be concluded that making use of NGS data (transcriptomic data in particular) is sufficient for identifying SNP markers and that using a medium-throughput genotyping platform for validation of *in silico* SNPs (in a species with little genetic information) proved to be highly successful. The aim was therefore to identify SNPs that could be used for saturating the linkage map of *H. midae*. From the 186 newly developed SNP markers, 128 were polymorphic and could be used for linkage map construction.

3. Linkage mapping

As confirmed by Van der Merwe and Roodt-Wilding (2008) the haploid chromosome number of *H. midae* is 18. Therefore, when the linkage map was constructed 18 linkage groups were expected. However, as the number of markers available for mapping purposes in *H. midae* (as in many aquaculture species) is limited, this was not easily accomplished. Either fewer or more linkage groups than expected were obtained; but this is to be expected when constructing preliminary linkage maps for non-model species (Liu *et al.* 2006). The linkage map constructed in the current study is the fourth map constructed for perlemoen. Previous maps were constructed based on AFLP markers, microsatellite markers and a limited number of SNPs. During map construction in the current study 118 newly developed SNP markers were available for use [10 markers had to be excluded due to both parents being different homozygotes (e.g. AA x GG); therefore making it impossible to determine which parent passed on which allele].

Six families were used for map construction in the current study of which four were new families and two were families used in previous studies (Blaauw 2012; Jansen 2012). A total of 314 markers (SNPs and microsatellites) were available for use of which 186 mapped to the integrated map (combined family maps). The number of linkage groups obtained for the different families ranged from nine to 22, with the two families previously used containing the highest number of linkage groups. This was attributed to the larger number of available markers (microsatellites and SNPs) for these two families and less available markers for the four new families (SNPs only).

It was also concluded that maps based on only SNP or microsatellite markers were not as dense as maps constructed using both types of markers. Although for linkage mapping focus is increasingly shifting towards SNPs as a result of their abundance, microsatellites are still very valuable for linkage map construction due to their high levels of polymorphism. A significantly larger number of SNPs are needed to obtain the same amount of information content supplied by only a few microsatellites. Therefore it is beneficial for linkage map construction to use more than one type of marker that has different advantages (low error rate of SNPs; high polymorphism levels of microsatellites) so as to be able to construct the highest quality map possible (Ball *et al.* 2010).

Genome coverage of 79.1% was obtained in the current study. This is 14.1% higher than the genome coverage found in the previously constructed linkage map by Jansen (2012). This was made possible by the newly developed SNP markers in the current study; contributing to a higher total number of markers available for mapping. Also, due to more available markers, the average marker spacing, which is a crucial factor in QTL mapping (10 cM, Massault *et al.* 2008), was decreased to 6.9 cM from the 9.3 cM obtained by Jansen (2012). Developing more markers (SNPs and microsatellites) will decrease the average marker distance even further as well as lead to higher genome coverage. Also, it will be easier to identify anchor loci which will facilitate with integrating newly developed markers on linkage maps as well as assist with the merging of linkage groups where currently an insufficient number of markers are available.

It is evident that the development of additional informative markers has numerous benefits. The aim of this section was to use newly developed SNP markers in

conjunction with previously developed SNP and microsatellite markers in the construction and saturation of the linkage map of *H. midae*. As results obtained show, this was successfully achieved with the use of JoinMap. This software proved to be very user-friendly and allowed for easy integration of the separate family maps to construct an integrated linkage map of *H. midae*. It also offers two mapping algorithms (regression mapping and ML mapping) which are very valuable especially when marker order or map length seems to be unreliable. It was possible to compare the maps constructed using the different mapping algorithms and verify marker order where necessary.

4. Conclusions and future considerations

As a result of illegal, unreported and unregulated fishing of perlemoen, wild populations have almost been entirely depleted. This led to the artificial cultivation of this natural resource in order to supply the world demand. In order to assist breeding practices as well as ensure the sustainability of these commercial populations, genetic management is needed. Genetic markers are employed for this purpose and are useful in amongst others genetic diversity studies, parentage assignment and linkage mapping that will enable the identification of QTL for selective breeding purposes.

The development of SNPs from EST sequences is gaining popularity as these polymorphisms represent gene-linked markers which are very valuable for QTL analysis. During this study it was shown that developing more markers and using more mapping families (as well as larger mapping families) greatly benefits the process of constructing linkage maps. A possible way to prevent the inclusion of non-informative SNP markers in the genotyping assay would be to initially genotype potential SNPs in the parents of the mapping families in order to determine possible homo- or heterozygote state of the offspring. It should be noted however that the smallest number of samples that can be genotyped with the GoldenGate assay is 96 samples (one plate), thus it would be beneficial to determine the SNP genotypes of the parents first by sequencing or a similar method that would be cost- and time-efficient for a smaller number of individuals.

During this study some difficulties were encountered. Various samples were not successfully genotyped due to low DNA concentrations. Obtaining enough individuals per family also proved to be problematic as families contained non-family members (other families' offspring). To address the problem associated with low DNA concentration, alternative DNA extraction methods that have a higher DNA yield could be investigated in the future. This may include the use of commercial extraction kits. It is however quite problematic to keep families separate within the commercial farm setup due to logistical issues and the nocturnal movements of the animals which do not confine them to specific baskets. It would however be beneficial in future studies to include a larger number of offspring so that in the event of non-family members being present in a certain family, enough family members would remain to be available for linkage analyses. Another difficulty that presented itself was the very stringent criteria that were set to identify putative SNPs. Although a large number of SNPs were present in the transcriptome data (11 934 SNPs, chapter two), the number decreased significantly when the MAF was set to 10% and the coverage to 80. This is however crucial for identifying true SNPs and lowering these values will only lead to a higher number of false positives. The 60 bp flanking regions containing no other polymorphisms however led to the exclusion of many additional putative SNPs. As this flanking region criterion is a necessary prerequisite for genotyping with the GoldenGate assay (primer binding may be affected), it is however also important to adhere to this criterion. If large numbers of SNP markers are to be developed in future, other medium- to high-throughput genotyping methods that do not require as large flanking regions [such as MALDI-TOF (see chapter one)] could prove to be more successful.

The higher number of markers used for linkage map construction in the current study has aided the identification of anchor loci that will greatly benefit the construction of future linkage maps for *Haliotis midae*. The denser linkage map constructed in the current study provides a step closer to QTL mapping in *H. midae*; an essential tool for identifying markers associated with economically important traits that can be used in future marker-assisted selection for this valuable South African aquaculture species.

References

- Aitken, N., Smith, S., Schwarz, C., Morin, P.A., 2004. Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach. *Molecular Ecology* 13, 1423-1431.
- 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.
- Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L., Lander, E.S., 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407, 513-516.
- Andreassen, R., Lunner, S., Høyheim, B., 2010. Targeted SNP discovery in Atlantic salmon (*Salmo salar*) genes using a 3'UTR-primed SNP detection approach. *BMC Genomics* 11, 706.
- Arnheim, N. and Calabrese, P., 2009. Understanding what determines the frequency and pattern of human germline mutations. *Nature Reviews Genetics* 10, 478-488.
- Artamonova, V., 2007. Genetic markers in population studies of Atlantic salmon *Salmo salar* L.: Analysis of DNA sequences. *Russian Journal of Genetics* 43, 341-353.
- 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.
- Aslam, M.L., Bastiaansen, J.W.M., Crooijmans, R.P.M.A., Vereijken, A., Megens, H.J., Groenen, M.A.M., 2010. A SNP based linkage map of the turkey genome reveals multiple intrachromosomal rearrangements between the Turkey and Chicken genomes. *BMC Genomics* 11, 647.

- Badenhorst, D., 2008. Development of AFLP markers for *Haliotis midae* for linkage mapping. [Unpublished Master of Science thesis] Stellenbosch University, South Africa.
- Ball, A.D., Stapley, J., Dawson, D.A., Birkhead, T.R., Burke, T., Slate, J., 2010. A comparison of SNPs and microsatellites as linkage mapping markers: Lessons from the zebra finch (*Taeniopygia guttata*). BMC Genomics 11, 218.
- Baranski, M., Loughnan, S., Austin, C. M., Robinson, N., 2006. A microsatellite linkage map of the blacklip abalone, *Haliotis rubra*. Animal Genetics 37, 563-570.
- 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.
- Beaumont, A.R., Boudry, P., Hoare, K., 2010. Biotechnology and genetics in fisheries and aquaculture. Second Edition. Wiley-Blackwell. John Wiley and Sons, Inc. pp 121.
- Beckmann, J.S., Estivill, X., Antonarakis, S.E., 2007. Copy number variants and genetic traits: Closer to the resolution of phenotypic to genotypic variability. Nature Reviews Genetics 8, 639-646.
- 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.
- Beuzen, N.D., Stear, M.J., Chang, K.C., 2000. Molecular markers and their use in animal breeding. The Veterinary Journal 160, 42-52.
- Blaauw, S., 2012. SNP screening and validation in *Haliotis midae*. [Unpublished Master of Science thesis] Stellenbosch University, South Africa.

- Bowman, S., Hubert, S., Higgins, B., Stone, C., Kimball, J., Borza, T., Bussey, J., Simpson, G., Kozera, C., Curtis, B., Hall, J., Hori, T., Feng, C., Rise, M., Booman, M., Gamperl, A., Trippel, E., Symonds, J., Johnson, S., Rise, M.L., 2011. An integrated approach to gene discovery and marker development in Atlantic cod (*Gadus morhua*). *Marine Biotechnology* 13, 242-255.
- Brown, G.R., Kadel, E.E., Bassoni, D.L., Kiehne, K.L., Temesgen, B., van Buijtenen, J.P., Sewell, M.M., Marshall, K.A., Neale, D.B., 2001. Anchored reference loci in loblolly pine (*Pinus taeda* L.) for integrating pine genomics. *Genetics* 159, 799-809.
- Bunje, P., 2010. The Mollusca: Sea slugs, squid, snails, and scallops. [Online]. Available: <http://www.ucmp.berkeley.edu/taxa/inverts/mollusca/mollusca.php> [Accessed 3 September 2012].
- Campino, S., Auburn, S., Kivinen, K., Zongo, I., Ouedraogo, J.B., Mangano, V., Djimde, A., Doumbo, O.K., Kiara, S.M., Nzila, A., Borrmann, S., Marsh, K., Michon, P., Mueller, I., Siba, P., Jiang, H., Su, X.Z., Amaratunga, C., Socheat, D., Fairhurst, R.M., Imwong, M., Anderson, T., Nosten, F., White, N.J., Gwilliam, R., Deloukas, P., MacInnis, B., Newbold, C.I., Rockett, K., Clark, T.G., Kwiatkowski, D.P., 2011. Population genetic analysis of *Plasmodium falciparum* parasites using a customized Illumina GoldenGate genotyping assay. *PLoS One* 6, e20251.
- Castaño Sánchez, C., Smith, T.P.L., Wiedmann, R.T., Vallejo, R.L., Salem, M., Yao, J., Rexroad, C.E., 2009. Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics* 10, 559.
- Castaño-Sánchez, C., Fuji, K., Ozaki, A., Hasegawa, O., Sakamoto, T., Morishima, K., Nakayama, I., Fujiwara, A., Masaoka, T., Okamoto, H., Hayashida, K., Tagami, M., Kawai, J., Hayashizaki, Y., Okamoto, N., 2010. A second generation genetic linkage map of Japanese flounder (*Paralichthys olivaceus*). *BMC Genomics* 11, 554.

- Chakravarti, A., Lasher, L.K., Reefer, J.E., 1991. A maximum likelihood method for estimating genome length using genetic linkage data. *Genetics* 128, 175-182.
- Chancerel, E., Lepoittevin, C., Le Provost, G., Lin, Y., Jaramillo-Correa, J.P., Eckert, A.J., Wegrzyn, J.L., Zelenika, D., Boland, A., Frigerio, J., Chaumail, P., Garnier-Géré, P., Boury, C., Grivet, D., González-Martínez, S.C., Rouzé, P., Van de Peer, Y., Neale, D.B., Cervera, M.T., Kremer, A., Plomion, C., 2011. Development and implementation of a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine. *BMC Genomics* 12, 368.
- Charlesworth, B. and Charlesworth, D., 1998. Some evolutionary consequences of deleterious mutations. *Genetica* 102, 3-19.
- Chistiakov, D.A., Hellemans, B.C., Haley, S., Law, A.S., Tsigenopoulos, C.S., Katoulas, G., Bertotto, D., Libertini, A., Volckaert, F.A.M., 2005. A microsatellite linkage map of the European sea bass *Dicentrarchus labrax* L. *Genetics* 170, 1821-1826.
- Chistiakov, D.A., Hellemans, B., Volckaert, F.A.M., 2006. Microsatellites and their genomic distribution, evolution, function and applications: A review with special reference to fish genetics. *Aquaculture* 255, 1-29.
- Coates, B.S., Sumerford, D.V., Miller, N.J., Kim, K.S., Sappington, T.W., Siegfried, B.D., Lewis, L.C., 2009. Comparative performance of single nucleotide polymorphism and microsatellite markers for population genetic analysis. *Journal of Heredity* 100, 556-564.
- Collard, B.C.Y., Jahufer, M.Z.Z., Brouwer, J.B., Pang, E.C.K., 2005. An introduction to markers, quantitative trait loci (QTLs) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* 142, 169-196.
- Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., Robles, M., 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674-3676.

- Cook, P.A. and Gordon, H.R., 2010. World abalone supply, markets, and pricing. *Journal of Shellfish Research* 29, 569-571.
- Dames, S., Durtschi, J., Geiersbach, K., Stephens, J., Voelkerding, K.V., 2010. Comparison of the illumina genome analyzer and Roche 454 GS FLX for resequencing of hypertrophic cardiomyopathy-associated genes. *Journal of Biomolecular Techniques* 21, 73-80.
- Degnan, S.M., Imron, Geiger, D.L., Degnan, B.M., 2006. Evolution in temperate and tropical seas: Disparate patterns in southern hemisphere abalone (Mollusca: Vetigastropoda: Haliotidae). *Molecular Phylogenetics and Evolution* 41, 249-256.
- De Keyser, E., Shu, Q.Y., Van Bockstaele, E., De Riek, J., 2010. Multipoint-likelihood maximization mapping on 4 segregating populations to achieve an integrated framework map for QTL analysis in pot azalea (*Rhododendron simsii* hybrids). *BMC Molecular Biology* 11, 1.
- De la Cruz, F.L. and Gallardo-Escárate, C., 2011. Intraspecies and interspecies hybrids in *Haliotis*: natural and experimental evidence and its impact on abalone aquaculture. *Reviews in Aquaculture* 3, 74-99.
- Department of Agriculture, Forestry and Fisheries, 2011. *Aquaculture Annual Report 2011*.
- Doerge, R.W., 2002. Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* 3, 43-52.
- Doyle, J. and Doyle, J.L., 1987. Genomic plant DNA preparation from fresh tissue – CTAB method. *Phytochemical Bulletin* 19, 11.
- Echt, C., Knapp, S., Liu, B.H., 1992. Genome mapping with non-inbred crosses using GMendel 2.0. *Maize Genetics Cooperation Newsletter* 66, 27-29.
- Eklom, R. and 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 single nucleotide polymorphisms in weathervane scallop. *Molecular Ecology Notes* 5, 406-409.
- Elliott, N.G., 2000. Genetic improvement programmes in abalone: What is the future? *Aquaculture Research* 31, 51-59.
- Estes, J.A., Lindberg, D.R., Wray, C., 2005. Evolution of large body size in abalone (*Haliotis*): Patterns and implications. *Paleobiology* 31, 591-606.
- Fallu, R., 1991. Abalone farming. Oxford: Fishing News Books. A division of Blackwell Scientific Publications Ltd, United Kingdom, pp 1-120.
- 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.
- Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R., Bateman, A., 2010. The pfam protein families database. *Nucleic Acids Research* 38, D211-D222.
- Fishman, L., Kelly, A.J., Morgan, E., Willis, J.H., 2001. A genetic map in the *Mimulus guttatus* species complex reveals transmission ratio distortion due to heterospecific interactions. *Genetics* 159, 1701-1716.
- 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.
- 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., 1999. A total evidence cladistic analysis of the Haliotidae (Gastropoda: Vetigastropoda). [Unpublished Doctor of Philosophy Thesis] University of Southern California, California.
- Geiger, D.L., 2000. Distribution and biogeography of the Haliotidae (Gastropoda: Vestigastropoda) world-wide. *Bolletino Malacologia* 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.
- Gharbi, K., Gautier, A., Danzmann, R.G., Gharbi, S., Sakamoto, T., Hoyheim, B., Taggart, J.B., Cairney, M., Powell, R., Krieg, F., Okamoto, N., Ferguson, M.M., Holm, L.E., Guyomard, R., 2006. A linkage map for brown trout (*Salmo trutta*): Chromosome homeologies and comparative genome organization with other salmonid fish. *Genetics* 172, 2405-2419.
- Gheyas, A.A., Houston, R.D., Mota-Velasco, J.C., Guy, D.R., Tinch, A.E., Haley, C.S., Woolliams, J.A., 2010. Segregation of infectious pancreatic necrosis resistance QTL in the early life cycle of Atlantic salmon (*Salmo salar*). *Animal Genetics* 41, 531-536.
- Gilbey, J., Verspoor, E., McLay, A., Houlihan, D., 2004. A microsatellite linkage map for Atlantic salmon (*Salmo salar*). *Animal Genetics* 35, 98-105.
- Gjedrem, T., Robinson, N., Rye, M., 2012. The importance of selective breeding in aquaculture to meet future demands for animal protein: A review. *Aquaculture* 350-353, 117-129.
- Glover, K.A., Hansen, M.M., Lien, S., Als, T.D., Høyheim, B., Skaala, Ø., 2010. A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genetics* 11, 2.
- Gordon, H.R. and Cook, P.A., 2001. World abalone supply, markets and pricing: historical, current and future. *Journal of Shellfish Research* 20, 567-570.

- Grattapaglia, D., Silva-Junior, O.B., Kirst, M., de Lima, B.M., Faria, D.A., Pappas, G.J., 2011. High-throughput SNP genotyping in the highly heterozygous genome of *Eucalyptus*: Assay success, polymorphism and transferability across species. *BMC Plant Biology* 11, 65.
- Guillot, G. and Foll, M., 2009. Correcting for ascertainment bias in the inference of population structure. *Bioinformatics* 25, 552-554.
- Gut, I.G., Lathrop, M.G., 2004. Duplicating SNPs. *Nature Genetics* 36, 789-790.
- Guyomard, R., Mauger, S., Tabet-Canale, K., Martineau, S., Genet, C., Krieg, F., Quillet, E., 2006. A Type I and Type II microsatellite linkage map of Rainbow trout (*Oncorhynchus mykiss*) with presumptive coverage of all chromosome arms. *BMC Genomics* 7, 302.
- Hauck, M. and Kroese, M., 2006. Fisheries compliance in South Africa: A decade of challenges and reform 1994–2004. *Marine Policy* 30, 74-83.
- Hauck, M. and Sweijd, N.A., 1999. A case study of poaching in South Africa and its impact on fisheries management. *Journal of Marine Science* 56, 1024-1032.
- Hauser, L. and Seeb J.E., 2008. Advances in molecular technology and their impact on fisheries genetics. *Fish and Fisheries* 9, 473-486.
- Hayes, B., Baranski, M., Goddard, M.E., Robinson, N., 2007a. Optimisation of marker assisted selection for abalone breeding programs. *Aquaculture* 265, 61-69.
- Hayes, B., Laerdahl, J.K., Lien, S., Moen, T., Berg, P., Hindar, K., Davidson, W.S., Koop, B.F., Adzhubei, A., Høyheim, B., 2007b. An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. *Aquaculture* 265, 82-90.
- 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 opportunities and challenges. *Molecular Ecology Resources* 11, 123-136.

Helyar, S.J., Limborg, M.T., Bekkevold, D., Babbucci, M., van Houdt, J., Maes, G.E., Bargelloni, L., Nielsen, R.O., Taylor, M.I., Ogden, R., Cariani, A., Carvalho, G.R., FishPopTrace Consortium, Panitz, F., 2012. SNP discovery using next generation transcriptomic sequencing in Atlantic herring (*Clupea harengus*). *PLoS One* 7, e42089.

Hepple, J., 2010. An integrated linkage map of Perlemoen (*Haliotis midae*). [Unpublished Master of Science thesis] Stellenbosch University, South Africa.

Ho, M.R., Tsai, K.W., Chen, C., Lin, W., 2010. dbDNV: a resource of duplicated gene nucleotide variants in human genome. *Nucleic Acids Research* 39, D920-D925.

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

Hyten, D., Song, Q., Choi, I., Yoon, M., Specht, J., Matukumalli, L., Nelson, R., Shoemaker, R., Young, N., Cregan, P., 2008. High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. *Theoretical and Applied Genetics* 116, 945-952.

Illumina, 2008. GoldenGate® Genotyping with VeraCode™ Technology: Custom 48-plex and 384-plex Assays. [Online]. Available: [www.illumina.com: http://www.illumina.com/Documents/products/technotes/technote_veracode_goldengate_genotyping.pdf](http://www.illumina.com/Documents/products/technotes/technote_veracode_goldengate_genotyping.pdf) [Accessed 5 September 2012].

Illumina, 2012. Illumina custom genotyping options. [Online]. Available: [www.illumina.com: http://www.illumina.com/Documents/products/datasheets/datasheet_custom_gt.pdf](http://www.illumina.com/Documents/products/datasheets/datasheet_custom_gt.pdf) [Accessed 17 September 2012].

Jansen, J., De Jong, A.G., Van Ooijen, J.W., 2001. Constructing dense genetic linkage maps. *Theoretical and Applied Genetics* 102, 1113-1122.

- Jansen, S., 2012. Linkage mapping in *Haliotis midae* using gene-linked markers. [Unpublished Master of Science thesis] Stellenbosch University, South Africa.
- Kang, J., Appleyard, S.A., Elliott, N.G., Jee, Y., Lee, J.B., Kang, S.W., Baek, M.K., Han, Y.S., Choi, T., Lee, Y.S., 2011. Development of genetic markers in abalone through construction of a SNP database. *Animal Genetics* 42, 309-315.
- Khan, M.A., Han, Y., Zhao, Y.F., Korban, S.S., 2012. A high-throughput apple SNP genotyping platform using the GoldenGate™ assay. *Gene* 494, 196-201.
- Kucuktas, H., Wang, S., Li, P., He, C., Xu, P., Sha, Z., Liu, H., Jiang, Y., Baoprasertkul, P., Somridhivej, B., Wang, Y., Abernathy, J., Guo, X., Liu, L., Muir, W., Liu, Z., 2009. Construction of genetic linkage maps and comparative genome analysis of catfish using gene-associated markers. *Genetics* 181, 1649-1660.
- Landau, M., 1992. Introduction to aquaculture. Stockton State College. John Wiley and Sons Inc, United States, pp 181-186.
- Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E., Newberg, L., 1987. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1, 174-181.
- 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.
- Lee, Y.H. and Vacquier, V.D., 1995. Evolution and systematics in Haliotidae (Mollusca, Gastropoda): Inference from DNA sequences of sperm lysin. *Marine Biology* 124, 267-278.
- Lee, Y.H., Ota, T., Vacquier, V.D., 1995. Positive selection is a general phenomenon in the evolution of abalone sperm. *Molecular Biology and Evolution* 12, 231-238.

- Lepoittevin, C., Frigerio, J., Garnier-Géré, P., Salin, F., Cervera, M., 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, e11034.
- Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., Quackenbush, J., 2000. An optimized protocol for analysis of EST sequences. *Nucleic Acids Research* 28, 3657-3665.
- Lien, S., Gidskehaug, L., Moen, T., Hayes, B.J., Berg, P.R., Davidson, W.S., Omholt, S.W., Kent, M.P., 2011. A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC Genomics* 12, 615.
- Lindberg, D.R., 1992. Evolution, distribution and systematics of Haliotidae. In: *Abalone of the World: Biology, Fisheries and Culture*. (Ed. by Shepherd, S.A., Tegner, M.J. & Guzman del Proo, S.A.). Fishing News Books, Great Britain. pp 3-18.
- 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, X., Liu, X., Guo, X., Gao, Q., Zhao, H., Zhang, G., 2006. A preliminary genetic linkage map of the pacific abalone *Haliotis discus hannai* Ino. *Marine Biotechnology* 8, 386-397.
- Liu, X., Liu, X., Zhang, G., 2007. Identification of quantitative trait loci for growth-related traits in the Pacific abalone *Haliotis discus hannai* Ino. *Aquaculture Research* 38, 789-797.
- Liu, Z., 2007. Microsatellite markers and assessment of marker utility. In: *Aquaculture Genome Technologies*. Blackwell Publishing Ltd, United Kingdom, pp 43-58.

- Liu, Z.J. and Cordes, J.F., 2004. DNA marker technologies and their applications in aquaculture genetics. *Aquaculture* 238, 1-37.
- Luo, C., Tsementzi, D., Kyripides, N., Read, T., Konstantinidis, K.T., 2012. Direct comparisons of Illumina vs Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* 7, 30087.
- Massault, C., Bovenhuis, H., Haley, C., de Koning, D., 2008. QTL mapping designs for aquaculture. *Aquaculture* 285, 23-29.
- McAndrew, B. and Napier, J., 2011. Application of genetics and genomics to aquaculture development: Current and future directions. *The Journal of Agricultural Science* 149, 143-152.
- Miké, V., 1977. Theories of quasi-linkage and "affinity": Some implications for population structure. *Proceedings of the National Academy of Sciences, USA* 74, 3513-3517.
- Milano, I., Babbucci, M., Panitz, F., Ogden, R., Nielsen, R.O., Taylor, M.I., Helyar, S.J., Carvalho, G.R., Espineira, M., Atanassova, M., Tinti, F., Maes, G.E., Patarnello, T., FishPopTrace Consortium, Bargelloni, L., 2011. Novel tools for conservation genomics: Comparing two high-throughput approaches for SNP discovery in the transcriptome of the European hake. *PLoS One* 6, e28008.
- Nichols, K.M., Young, W.P., Danzmann, R.G., Robison, B.D., Rexroad, C., Noakes, M., Phillips, R.B., Bentzen, P., Spies, I., Knudsen, K., Allendorf, F.W., Cunningham, B.M., Brunelli, J., Zhang, H., Ristow, S., Drew, R., Brown, K.H., Wheeler, P.A., Thorgaard, G. H., 2003. A consolidated linkage map for rainbow trout (*Oncorhynchus mykiss*). *Animal Genetics* 34, 102-115.
- Nielsen, E.E., Cariani, A., Aoidh, E.M., Maes, G.E., Milano, I., Ogden, R., Taylor, M., Hemmer-Hansen, J., Babbucci, M., Bargelloni, L., Bekkevold, D., Diopere, E., Grenfell, L., Helyar, S., Limborg, M.T., Martinsohn, J.T., McEwing, R., Panitz, F., Patarnello, T., Tinti, F., Van Houdt, J.K.J., Volckaert, F.A.M., Waples, R.S., Carvalho, G.R., 2012. Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nature Communications* 3, 851.

- Pérez, F., Erazo, C., Zhinaula, M., Volckaert, F., Calderón, J., 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.
- Postlethwait, J.H., Johnson, S.L., Midson, C.N., Talbot, W.S., Gates, M., Ballinger, E.W., Africa, D., Andrews, R., Carl, T., Eisen, J.S., 1994. A genetic linkage map for the zebrafish. *Science* 264, 699-703.
- 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 polymorphisms 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.
- 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.
- 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.

- Renaut, S., Nolte, A.W., Bernatchez, L., 2010. Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). *Molecular Ecology* 19, 115-131.
- Rexroad, C.E., Palti, Y., Gahr, S.A., Vallejo, R.L., 2008. A second generation genetic map for rainbow trout (*Oncorhynchus mykiss*). *BMC Genetics* 9, 74.
- Rhode, C., 2010. Development of gene-linked molecular markers in South African abalone (*Haliotis midae*) using an *in silico* mining approach. [Unpublished Master of Science thesis] Stellenbosch University, South Africa.
- Rhode, C., Slabbert, R., Roodt-Wilding, R., 2008. Microsatellite flanking regions: a SNP mine in South African abalone (*Haliotis midae*). *Animal Genetics* 39, 329.
- Roberts, S.B., Hauser, L., Seeb, L.W., Seeb, J.E., 2012. Development of genomic resources for pacific herring through targeted transcriptome pyrosequencing. *PLoS One* 7, e30908.
- Robinson, N., Li, X., Hayes, B., 2010. Testing options for the commercialization of abalone selective breeding using bioeconomic simulation modelling. *Aquaculture Research* 41, 268-288.
- Ronaghi, M., 2001. Pyrosequencing sheds light on DNA sequencing. *Genome Research* 11, 3-11.
- Roodt-Wilding, R., 2007. Abalone ranching: A review on genetic considerations. *Aquaculture Research* 38, 1229-1241.
- Roodt-Wilding, R. and Brink, D., 2011. Selection and sea snails: the South African story. *Philosophical Transactions in Genetics* 1, 1-41.
- Roodt-Wilding, R. and Slabbert, R., 2006. Molecular markers to assist the South African abalone industry. *South African Journal of Science* 102, 99-102.
- Ryynänen, H.J. and Primmer, C.R., 2006. Single nucleotide polymorphism (SNP) discovery in duplicated genomes: intron-primed exon-crossing (IPEC) as a

strategy for avoiding amplification of duplicated loci in Atlantic salmon (*Salmo salar*) and other salmonid fishes. *BMC Genomics* 7, 192.

Sakamoto, T., Danzmann, R.G., Gharbi, K., Howard, P., Ozaki, A., Khoo, S.K., Woram, R.A., Okamoto, N., Ferguson, M.M., Holm, L.E., Guyomard, R., Hoyheim, B., 2000. A microsatellite linkage map of rainbow trout (*Oncorhynchus mykiss*) characterized by large sex-specific differences in recombination rates. *Genetics* 155, 1331-1345.

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

Sandler, L., Hiraizumi, Y., Sandler, I., 1959. Meiotic drive in natural populations of *Drosophila melanogaster*. I. The cytogenetic basis of segregation-distortion. *Genetics* 44, 233-250.

Schaid, D.J., Guenther, J.C., Christensen, G.B., Hebring, S., Rosenow, C., Hilker, C.A., McDonnell, S.K., Cunningham, J.M., Slager, S.L., Blute, M.L., Thibodeau, S.N., 2004. Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility loci. *The American Journal of Human Genetics* 75, 948-965.

Schlötterer, C., 2004. The evolution of molecular markers – just a matter of fashion? *Nature Reviews Genetics* 5, 63-68.

Sechi, T., Coltman, D.W., Kijas, J.W., 2010. Evaluation of 16 loci to examine the cross-species utility of single nucleotide polymorphism arrays. *Animal Genetics* 41, 199-202.

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

Seeb, J.E., Pascal, C.E., Grau, E.D., Seeb, L.W., Templin, W.D., Harkins, T., Roberts, S.B., 2011b. Transcriptome sequencing and high-resolution melt

analysis advance single nucleotide polymorphism discovery in duplicated salmonids. *Molecular Ecology Resources* 11, 335-348.

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

Shen, R., Fan, J., 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.

Shi, Y., Guo, X., Zhifeng, G. U., Wang, A., Wang, Y., 2010. Preliminary genetic linkage map of the abalone *Haliotis diversicolor* Reeve. *Chinese Journal of Oceanology and Limnology* 28, 549-557.

Singhal, D., Gupta, P., Sharma, P., Kashyap, N., Anand, S., Sharma, H., 2011. *In-silico* single nucleotide polymorphisms (SNP) mining of *Sorghum bicolor* genome. *African Journal of Biotechnology* 10, 580-583.

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 Society* 39, 429-435.

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

Slabbert, R., Hepple, J.A., Rhode, C., Bester-van der Merwe, A.E., Roodt-Wilding, R., 2012. New microsatellite markers for the abalone *H. midae* developed by 454 pyrosequencing and *in silico* analyses. *Genetics and Molecular Research* 11, 2769-2779.

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.

- Sonesson, A.K., 2007. Within-family marker-assisted selection for aquaculture species. *Genetics Selection Evolution* 7, 301-317.
- Souche, E.L., Hellemans, B., Van Houdt, J.K.J., Canario, A., Klages, S., Reinhardt, R., Volckaert, F.A.M., 2007. Mining for single nucleotide polymorphisms in expressed sequence tags. *Journal of Integrative Bioinformatics* 4, 73.
- Stam, P., 1993. *Construction of integrated genetic linkage maps by means of a new computer package: JoinMap*. *The Plant Journal* 3, 739-744.
- Studer, B., Kolliker, R., Muylle, H., Asp, T., Frei, U., Roldan-Ruiz, I., Barre, P., Tomaszewski, C., Meally, H., Barth, S., Skot, L., Armstead, I.P., Dolstra, O., Lubberstedt, T., 2010. EST-derived SSR markers used as anchor loci for the construction of a consensus linkage map in ryegrass (*Lolium* spp.). *BMC Plant Biology* 10, 177.
- Sturtevant, A.H., 1913. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology* 74, 43-59.
- Suiter, K.A., Wendel, J.F., Case, J.S., 1983. LINKAGE-1: a pascal computer program for the detection and analysis of genetic linkage. *Journal of Heredity* 74, 203-204.
- Sunnucks, P., 2000. Efficient genetic markers for population biology. *Trends in Ecology and Evolution* 15, 199-203.
- Tarr, R.J.Q., 1989. Abalone. In: *Oceans of life off southern Africa*. Payne, A.I.L., and R. Crawford, J. M. (eds). Vlaeberg Publishers, Cape Town pp 62-69.
- 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.

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

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.

Tsuchihashi, Z. and Dracopoli, N.C., 2002. Progress in high throughput SNP genotyping methods. *The Pharmacogenomics Journal* 2, 103-110.

Twyman, R.M., 2005. Single Nucleotide Polymorphism (SNP) genotyping techniques - An overview. University of York. Marcel Dekker Inc, United Kingdom, pp 1202-1207.

Useche, F.J., Gao, G., HanaFey, M., Rafalski, A., 2001. High-throughput identification, database storage and analysis of SNPs in EST sequences. *Genome Informatics* 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 der Merwe, M. and Roodt-Wilding, R., 2008. Chromosome number of the South African abalone *Haliotis midae*. *African Journal of Marine Science* 30, 195-198.

Van Ooijen, J.W., 2006. JoinMap® v4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma, B. V. Wageningen, The Netherlands.

Van Ooijen, J.W., 2011. Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genetics Research* 93, 343-349.

Van Wormhoudt, A., Le Bras, Y., Huchette, S., 2009. *Haliotis marmorata* from Senegal; A sister species of *Haliotis tuberculata*: Morphological and molecular evidence. *Biochemical Systematics and Ecology* 37, 747-755.

- 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.
- Vervalle, J., Hepple, J., Jansen, S., Du Plessis, J., Wang, P., Rhode, C., Roodt-Wilding, R., In press. Integrated linkage map of *Haliotis midae* Linnaeus based on microsatellites and SNPs. *Journal of Shellfish Research*.
- Vignal, A., Milan, D., SanCristobal, M., Eggan, A., 2002. A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution* 34, 275-305.
- Waldbieser, G.C., Bosworth, B.G., Nonneman, D.J., Wolters, W.R., 2001. A microsatellite-based genetic linkage map for channel catfish, *Ictalurus punctatus*. *Genetics* 158, 727-734.
- Wang, C.M., Bai, Z.Y., He, X.P., Lin, G., Xia, J.H., Sun, F., Lo, L.C., Feng, F., Zhu, Z.Y., Yue, G.H., 2011. A high-resolution linkage map for comparative genome analysis and QTL fine mapping in Asian seabass, *Lates calcarifer*. *BMC Genomics* 12, 174.
- Wang, S., Bao, W., Pan, J., Zhang, L., Yao, B., Zhan, A., Bi, K., Zhang, Q., 2004. AFLP linkage map of an intraspecific cross in *Chlamys farreri*. *Journal of Shellfish Research* 23, 491-499.
- Wang, S., Sha, Z., Sonstegard, T.S., Liu, H., Xu, P., Somridhivej, B., Peatman, E., Kucuktas, H., Liu, Z., 2008. Quality assessment parameters for EST-derived SNPs from catfish. *BMC Genomics* 9, 45.
- Xia, J.H., Liu, F., Zhu, Z.Y., Fu, J., Feng, J., Li, J., Yue, G.H., 2010. A consensus linkage map of the grass carp (*Ctenopharyngodon idella*) based on microsatellites and SNPs. *BMC Genomics* 11, 135.
- Yu, Z. and Guo, X., 2003. Genetic linkage map of the eastern oyster *Crassostrea virginica* Gmelin. *The Biological Bulletin* 204, 327-338.

Zhan, H. and Xu, S., 2011. Generalized linear mixed model for segregation distortion analysis. *BMC Genetics* 12, 97.

Zhan, X., Fan, F., You, W., Yu, J., Ke, C., 2011. Construction of an integrated map of *Haliotis diversicolor* using microsatellite markers. *Marine Biotechnology* 14, 79-86.

Addendum 1: BLASTX search results for 139 contigs containing 186 SNPs.

SNP name	Contig	Sequence description	E-Value	Organism		Identity	Genbank Accession number
				Scientific name	Common name		
C31_1387	Contig 31	Tyrosine 3-monooxygenase / tryptophan 5-monooxygenase activation epsilon polypeptide	1.05E-135	<i>Nasonia vitripennis</i>	Wasp	90	XP_001600604
C31_1488							
C45_3002	Contig 45	Kyphoscoliosis peptidase	0	<i>Hirudo medicinalis</i>	Leech	57	AAK49949
C48_636	Contig 48	Mannose c type 2	6.75E-87	<i>Crassostrea virginica</i>	Atlantic oyster	55	AAB34577
C48_933							
C67_2395	Contig 67	-	-	-	-	-	-
C102_1408	Contig 102	-	-	-	-	-	-
C140_2112	Contig 140	Na ⁺ K ⁺ -ATPase alpha subunit	0	<i>Paroctopus digueti</i>	Pacific pygmy octopus	92	AEH68841
C140_2421							
C150_320	Contig 150	Ubiquinol-cytochrome c reductase core protein 2	4.72E-133	<i>Branchiostoma floridae</i>	Florida lancelet	68	XP_002595411
C152_797	Contig 152	Transcription elongation factor b polypeptide 1	7.69E-69	<i>Saccoglossus kowalevskii</i>	Acorn worm	96	XP_002737213
C158_67	Contig 158	Alpha-sarcomeric-like isoform 2	0	<i>Apis florea</i>	Dwarf honeybee	84	XP_003693213
C158_238							
C184_1379	Contig 184	Chorion peroxidase	4.14E-117	<i>Acyrtosiphon pisum</i>	Pea aphid	53	XP_001946672
C229_2772	Contig 229	14-3-3 zeta	1.21E-103	<i>Heliothis virescens</i>	Tobacco budworm moth	78	ACR07788
C231_2116	Contig 231	-	-	-	-	-	-
C236_970	Contig 236	Isoform a	1.74E-97	<i>Crassostrea virginica</i>	Atlantic oyster	60	AAB34577
C250_199	Contig 250	-	-	-	-	-	-

C253_1545	Contig 253	PDZ and LIM domain protein ZASP	1.91E-29	<i>Clonorchis sinensis</i>	Chinese liver fluke	55	GAA56670
C300_1828	Contig 300	-	-	-	-	-	-
C300_4738							
C300_4982							
C300_6993							
C311_406	Contig 311	Chromosome segregation protein SMC	4.51E-34	<i>Saccoglossus kowalevskii</i>	Acorn worm	37	XP_002738323
C311_1293							
C314_1039	Contig 314	-	-	-	-	-	-
C327_1076	Contig 327	Beta-glucanase	0	<i>Haliotis discus discus</i>	Disk abalone	90	ABO26613
C347_1008	Contig 347	Ribosome biogenesis protein NSA2 homolog	3.85E-168	<i>Branchiostoma floridae</i>	Florida lancelet	96	XP_002609891
C379_2197	Contig 379	Heat shock protein 90	0	<i>Haliotis discus hannai</i>	Ezo abalone	99	ACX94847
C387_215	Contig 387	Indoleamine 2,dioxygenase- like	0	<i>Haliotis diversicolor</i>	Japanese abalone	98	Q01966
C387_582							
C394_1510	Contig 394	Heat shock protein	2.22E-37	<i>Ruditapes philippinarum</i>	Manila clam	82	ACU83231
C421_541	Contig 421	Calcium-binding protein	2.05E-06	<i>Entamoeba dispar</i> Strain SAW760	None	46	XP_001736602
C428_225	Contig 428	Myosin light chain kinase	2.76E-32	<i>Nasonia vitripennis</i>	Wasp	57	XP_003425477
C428_306							
C428_2101							
C428_2186							
C450_1390	Contig 450	Intermediate filament protein	5.99E-15	<i>Helix aspersa</i>	Garden snail	52	P22488
C460_1184	Contig 460	Transport protein Sec61 subunit alpha 2	0	<i>Acyrtosiphon pisum</i>	Pea aphid	96	NP_001119639
C460_1745							

C549_128	Contig 549	Nuclear ribonucleoprotein	1.05E-66	<i>Saccoglossus kowalevskii</i>	Acorn worm	80	XP_002741832
C570_855	Contig 570	Tumor suppressor candidate 5 homolog	1.35E-05	<i>Xenopus (Silurana) tropicalis</i>	Western clawed frog	55	NP_001005003
C570_1064							
C618_116	Contig 618	S-adenosylmethionine synthase isoform type-2	0	<i>Branchiostoma floridae</i>	Florida lancelet	91	XP_002596166
C681_815	Contig 681	Spectrin alpha	0	<i>Tribolium castaneum</i>	Red flour beetle	86	XP_973750
C844_440	Contig 844	Selenoprotein 1	2.22E-28	<i>Danio rerio</i>	Zebrafish	60	AAH86844
C844_751							
C853_1199	Contig 853	Myosin heavy chain	2.78E-142	<i>Argopecten irradians</i>	Atlantic bay scallop	87	AAF62393
C910_1175	Contig 910	X-box binding protein 1	2.91E-31	<i>Mytilus edulis</i>	Blue mussel	64	ABA43316
C911_290	Contig 911	ATP H ⁺ mitochondrial F1 alpha subunit cardiac muscle	0	<i>Taeniopygia guttata</i>	Zebra finch	92	XP_002195743
C911_461							
C911_839							
C911_1343							
C929_475	Contig 929	Hexokinase	0	<i>Crassostrea gigas</i>	Pacific oyster	81	CAJ28915
C929_734							
C929_1311							
C929_2563							
C980_261	Contig 980	-	-	-	-	-	-
C1001_964	Contig 1001	Serca (sarco-endoplasmic reticulum calcium ATPase) family member (sca-1)	6.71E-26	<i>Pinctada fucata</i>	Pearl oyster	89	ABS19817
C1002_85	Contig 1002	Protein translation factor SUI1 homolog	1.83E-43	<i>Acyrtosiphon pisum</i>	Pea aphid	77	XP_001948896
C1002_557							

C1094_579	Contig 1094	Eukaryotic translation initiation factor 5A-1	2.27E-63	<i>Oreochromis niloticus</i>	Nile tilapia	82	XP_003442933
C1177_137	Contig 1177	Dopamine beta-hydroxylase	1.12E-159	<i>Haliotis discus discus</i>	Disk abalone	61	ABO26633
C1254_187	Contig 1254	Beta 2c	0	<i>Saccostrea kegaki</i>	Spiny oyster	100	BAG55008
C1254_529							
C1363_269	Contig 1363	Na ⁽⁺⁾ H ⁽⁺⁾ exchange regulatory cofactor NHE-RF1	5.83E-48	<i>Ailuropoda melanoleuca</i>	Giant panda	59	EFB25599
C1384_655	Contig 1384	Voltage-dependent anion channel 2	0	<i>Haliotis diversicolor</i>	Japanese abalone	98	ADI56517
C1384_793							
C1449_847	Contig 1449	Kazal-type serine protease inhibitor domain-containing protein 1	1.02E-173	<i>Haliotis diversicolor</i>	Japanese abalone	99	AEE01360
C1462_825	Contig 1462	Elongation factor 2	0	<i>Caenorhabditis elegans</i>	Nematode	83	NP_492457
C1462_917							
C1462_1238							
C1520_1336	Contig 1520	Collagen alpha-1 chain	0	<i>Haliotis diversicolor</i>	Japanese abalone	94	AEW42986
C1630_199	Contig 1630	Cytosolic malate dehydrogenase	3.66E-151	<i>Lottia pelta</i>	Shield limpet	83	ACJ64673
C1726_472	Contig 1726	-	-	-	-	-	-
C1783_492	Contig 1783	Cat eye syndrome chromosome candidate 5 homolog	1.28E-116	<i>Clonorchis sinensis</i>	Chinese liver fluke	66	GAA47299
C1797_660	Contig 1797	Ornithine decarboxylase antizyme	2.48E-109	<i>Haliotis diversicolor</i>	Japanese abalone	98	ACV32415
C1797_1023							
C1797_1098							
C1813_219	Contig 1813	Protein BTG2-like	1.00E-45	<i>Crassostrea gigas</i>	Pacific oyster	81	ACH92125

C1813_300							
C1878_506	Contig 1878	PREDICTED: Heterogeneous nuclear ribonucleoprotein A1-like	4.91E-65	<i>Saccoglossus kowalevskii</i>	Acorn worm	76	XP_002741832
C2028_1228	Contig 2028	PREDICTED: Hypothetical protein	8.97E-05	<i>Hydra magnipapillata</i>	Fresh water polyp	56	XP_002154876
C2028_1328							
C2040_468	Contig 2040	Multiple banded antigen	9.87E-59	<i>Haliotis discus</i>		95	BAA75669
C2040_1251							
C2122_257	Contig 2122	Ependymin related protein-1 precursor	3.09E-116	<i>Haliotis discus discus</i>	Disk abalone	95	ABO26653
C2141_350	Contig 2141	Mitochondrial ATP synthase	1.98E-107	<i>Haliotis discus discus</i>	Disk abalone	96	ABO26657
C2141_504							
C2180_279	Contig 2180	Myosin alkali light chain 1	1.86E-80	<i>Haliotis discus discus</i>	Disk abalone	89	ABO26638
C2236_965	Contig 2236	High mobility group-T protein	3.74E-60	<i>Saccostrea kegaki</i>	Spiny oyster	70	BAG55013
C2362_847	Contig 2362	-	-	-	-	-	-
C2406_641	Contig 2406	Ribosomal protein L23a	6.18E-66	<i>Argopecten irradians</i>	Atlantic bay scallop	92	AAN05592
C2558_743	Contig 2558	PREDICTED: Neurogenic locus Notch protein-like	5.73E-23	<i>Strongylocentrotus purpuratus</i>	Purple sea urchin	42	XP_782555
C2735_326	Contig 2735	-	-	-	-	-	-
C2899_1354	Contig 2899	ATP H ⁺ mitochondrial F1 beta polypeptide	0	<i>Pinctada fucata</i>	Pearl oyster	90	ABC86835
C2903_286	Contig 2903	Ribosomal protein S17	2.77E-11	<i>Lepidochitona cinerea</i>	Grey chiton	91	ACR24968
C2903_1043							
C2915_875	Contig 2915	20 kDa Calcium-binding	1.46E-58	<i>Ruditapes</i>	Manilla clam	82	AFB83400

		protein		<i>philippinarum</i>			
C3107_715	Contig 3107	Actin-related protein 2/3 complex subunit 1A	1.07E-132	<i>Saccostrea kegaki</i>	Spiny oyster	81	BAG55010
C3220_245	Contig 3220	-	-	-	-	-	-
C3495_541	Contig 3495	Universal stress protein	2.36E-23	<i>Nematostella vectensis</i>	Starlet sea anemone	57	XP_001636809
C3676_443	Contig 3676	Actin 2	0	<i>Haliotis iris</i>	Blackfoot paua	99	AAX19286
C3835_411	Contig 3835	Sorbitol dehydrogenase	2.67E-114	<i>Gallus gallus</i>	Red junglefowl	68	XP_413719
C3914_977	Contig 3914	-	-	-	-	-	-
C4144_389	Contig 4144	Cathepsin D	0	<i>Pinctada maxima</i>	Pearl oyster	80	AEI58896
C4147_533	Contig 4147	EF-hand family protein	9.79E-08	<i>Littorina littorea</i>	Common periwinkle	43	AAM20842
C4181_822	Contig 4181	PREDICTED: Protein	9.59E-32	<i>Nematostella vectensis</i>	Starlet sea anemone	66	XP_001625221
C4181_893							
C4223_662	Contig 4223	Profilin	7.79E-56	<i>Haliotis diversicolor</i>	Japanese abalone	78	ABY87349
C4463_182	Contig 4463	-	-	-	-	-	-
C4593_326	Contig 4593	-	-	-	-	-	-
C4778_234	Contig 4778	Beta-Ig-H3/ fasciclin	7.47E-150	<i>Haliotis discus discus</i>	Disk abalone	81	ADJ21804
C4778_642							
C4791_1099	Contig 4791	Carboxypeptidase A1 precursor	2.53E-65	<i>Daphnia pulex</i>	Waterflea	52	EFX83250
C5054_124	Contig 5054	Cartilage matrix protein	1.70E-51	<i>Amphimedon queenslandica</i>	Demosponge	50	XP_003391549
C5054_1800							
C5106_273	Contig 5106	-	-	-	-	-	-

C5106_6741							
C5339_366	Contig 5339	Myosin heavy chain	1.79E-122	<i>Placopecten magellanicus</i>	Atlantic sea scallop	85	AAB03660
C5433_233	Contig 5433	Cytochrome c oxidase subunit mitochondrial	1.28E-04	<i>Danio rerio</i>	Zebrafish	55	XP_685495
C5634_234	Contig 5634	Unknown	4.30E-19	<i>Chrysomela tremula</i>	Poplar leaf beetle	44	ACP18834
C5741_659	Contig 5741	-	-	-	-	-	-
C6012_280	Contig 6012	ATP Synthase lipid-binding mitochondrial precursor	7.07E-55	<i>Haliotis diversicolor</i>	Japanese abalone	98	ABY87376
C6012_652							
C6061_1289	Contig 6061	Phosphoglycerate mutase	6.18E-115	<i>Clonorchis sinensis</i>	Chinese liver fluke	79	ABZ82035
C6631_237	Contig 6631	-	-	-	-	-	-
C7947_662	Contig 7947	NADH dehydrogenase subunit 5	0	<i>Haliotis rubra</i>	Blacklip abalone	96	YP_026073
C7947_1013							
C7947_1867							
C8539_132	Contig 8539	Calmodulin	4.93E-58	<i>Schistosoma mansoni</i>	Blood fluke	98	XP_002574095
C9238_1342	Contig 9238	-	-	-	-	-	-
C9471_299	Contig 9471	60S Ribosomal protein L8	0	<i>Haliotis discus discus</i>	Disk abalone	99	ABO26687
C9511_498	Contig 9511	-	-	-	-	-	-
C10524_242	Contig 10524	Protein transport protein sec61 subunit gamma-like	2.54E-27	<i>Ciona intestinalis</i>	Sea squirt	97	NP_001027676
C11784_1697	Contig 11784	-	-	-	-	-	-
C12119_299	Contig 12119	-	-	-	-	-	-
C13865_165	Contig 13865	Cathepsin S	1.30E-17	<i>Pinctada fucata</i>	Pearl oyster	70	ADC52431

C14033_777	Contig 14033	Myosin heavy chain	2.23E-149	<i>Mytilus galloprovincialis</i>	Blue mussel	80	CAB64664
C15455_325	Contig 15455	Nucleoside diphosphate kinase	1.71E-102	<i>Haliotis discus discus</i>	Disk abalone	93	ABO26651
C16314_519	Contig 16314	Guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1	0	<i>Platynereis dumerilii</i>	Dumeril's clam worm	90	ACQ83470
C17220_332	Contig 17220	Glyceraldehyde-3-phosphate dehydrogenase	1.41E-101	<i>Daphnia magna</i>	Waterflea	84	CAB99475
C17943_1521	Contig 17943	PREDICTED: Zonadhesin-like	2.82E-19	<i>Saccoglossus kowalevskii</i>	Acorn worm	44	XP_002738323
C17963_254	Contig 17963	-	-	-	-	-	-
C17963_499							
C18184_390	Contig 18184	FK506-binding protein	2.20E-57	<i>Haliotis discus discus</i>	Disk abalone	89	ABO26659
C18679_62	Contig 18679	Kielin / chordin-like protein	1.25E-41	<i>Haliotis discus</i>		91	BAA75668
C18774_676	Contig 18774	-	-	-	-	-	-
C18774_877							
C18782_153	Contig 18782	-	-	-	-	-	-
C19500_354	Contig 19500	Acyl-binding protein	1.46E-25	<i>Carassius auratus</i>	Goldfish	83	BAJ83550
C19952_651	Contig 19952	-	-	-	-	-	-
C20003_92	Contig 20003	Hypothetical protein IscW_ISCW002118	3.21E-09	<i>Ixodes scapularis</i>	Deer tick	62	XP_002409469
C20142_203	Contig 20142	60S Ribosomal protein L10a	2.80E-105	<i>Branchiostoma floridae</i>	Florida lancelet	90	XP_002607839
C20174_2657	Contig 20174	Universal minicircle sequence binding protein	1.94E-26	<i>Latrodectus hesperus</i>	Western widow	64	ADV40117

C20267_102	Contig 20267	Polcalcin Bra r	2.49E-05	<i>Paramecium tetraurelia</i> Strain d4-2	None	51	XP_001449509
C20267_179							
C20427_267	Contig 20427	Myosin alkali light chain 1	2.54E-109	<i>Haliotis discus discus</i>	Disk abalone	100	ABO26638
C20580_186	Contig 20580	-	-	-	-	-	-
C20648_3041	Contig 20648	Filamin-C isoform 3	0	<i>Daphnia pulex</i>	Waterflea	72	EFX86436
C20682_843	Contig 20682	PREDICTED: protein	4.25E-33	<i>Nematostella vectensis</i>	Starlet sea anemone	66	XP_001625221
C20776_135	Contig 20776	Alginate lyase	2.45E-135	<i>Haliotis discus hannai</i>	Ezo abalone	94	BAC87758
C21068_302	Contig 21068	60S Ribosomal protein L7	4.78E-121	<i>Crassostrea gigas</i>	Pacific oyster	87	CAD89885
C21134_331	Contig 21134	60S Ribosomal protein L22	9.24E-35	<i>Lepidochitona cinerea</i>	Grey chiton	91	ACR24950
C21581_136	Contig 21581	Ribosomal protein L19	3.09E-93	<i>Argopecten irradians</i>	Atlantic bay scallop	94	AAN05588
C21593_75	Contig 21593	Ribosomal protein L15	8.19E-31	<i>Mus musculus</i>	House mouse	100	CAX15910
C21645_185	Contig 21645	Alpha tubulin	4.12E-43	<i>Saccoglossus kowalevskii</i>	Acorn worm	98	XP_002738641
C21673_148	Contig 21673	Ribosomal protein L28	6.61E-56	<i>Haliotis asinina</i>	Donkey's ear abalone	98	AAX11340
C21706_435	Contig 21706	NADH dehydrogenase	3.68E-18	<i>Drosophila grimshawi</i>	Fruit fly	83	XP_001990826
C21880_565	Contig 21880	Zinc-binding dehydrogenase family	1.14E-66	<i>Caenorhabditis elegans</i>	Nematode	61	NP_502269
C21880_1003							
C21880_1111							
C21908_64	Contig 21908	-	-	-	-	-	-
C22317_403	Contig 22317	Proline-rich extensin-like family protein	1.81E-17	<i>Drosophila willistoni</i>	Fruit fly	47	XP_002071748
C22340_159	Contig 22340	Anoxia-induced grl-like protein	1.09E-10	<i>Haliotis discus discus</i>	Disk abalone	49	ACZ15980

C22347_319	Contig 22347	-	-	-	-	-	-
C22431_278	Contig 22431	EF-hand domain protein	6.27E-07	<i>Biomphalaria glabrata</i>	Bloodfluke planorb	54	AAV91525
C22449_261	Contig 22449	Protein disulfide isomerase	0	<i>Haliotis discus discus</i>	Disk abalone	96	ABO26667
C22491_341	Contig 22491	Translation elongation factor 2	0	<i>Lycosa singoriensis</i>	Wolf spider	88	ABX75376
C22491_595							
C22491_727							
C22521_249	Contig 22521	Peptidyl-prolyl cis-trans isomerase CYP20-2	9.26E-40	<i>Oryza sativa</i> Japonica group	Asian rice	84	NP_001054392
C22537_1941	Contig 22537	Cytochrome c oxidase subunit 1	0	<i>Haliotis discus hannai</i>	Ezo abalone	97	ACB73222
C22568_597	Contig 22568	Ribosomal protein L3	2.80E-161	<i>Haliotis diversicolor</i>	Japanese abalone	100	AEW42982
C22574_507	Contig 22574	-	-	-	-	-	-
C22635_318	Contig 22635	Myosin heavy chain	2.06E-10	<i>Lycosa singoriensis</i>	Wolf spider	91	ABX75479

Addendum 2: Parent panel consisting of seven microsatellite loci.

Microsatellite	Primer sequence (5'-3')	Size (bp)	Genbank Accession number	Repeat tract
PS1.870 ^a	F: ACAACAACACACAGCACA R: GTGCCAAAACATATTTCAAAC	142	GU256718	(CACACG) _n (AC) _n
PS1.818 ^a	F: AATGTAGGGTTGCTTCAAATG R: GAGTGTGTGGGTGTCTCTTTC	244	GU256711	(ATGG) _n (TGGA) _n (AC) _n
PS1.305 ^a	F: CTCGAGTTTCAACCATTGAGT R: GGGTGGGTGTTACGAGTG	215	GU256679	(GCAC) _n
NR106 ^b	F: TCCTTGGCCAGAATAACC R: TATATGGTCTGCATCGCTG	395	DQ825709	(TG) _n
NR120 ^b	F: TTGAGCATGAGTCGTTGAGC R: ACCTGCTCTTTAGCTCAGATGG	502	EF121745	(TGAG) _n
NR20 ^b	F: CTACAACAAACGCCGATG R: TGCAGTAATAGGGGTACCAG	384	EF063097	(TCC) _n (TAC) _n
NS19 ^b	F: ACAACAACAAAGGTGGTCAA R: CAATGAATAGCTATGGGTCTG	380	EF033330	(AAGACCC) _n

(^a) Slabbert *et al.* (2012); (^b) Slabbert *et al.* (2008)

Addendum 3: Genotyping results of 192 SNPs.

Isolation method	SNP name	Variant	SNP position
<i>In silico</i> SNPs	C31_1387	A/G	1387
	C31_1488	A/G	1488
	C45_3002	A/C	3002
	C48_636*	T/C	636
	C48_933*	A/G	933
	C67_2395	A/T	2395
	C102_1408	A/T	1408
	C140_2112*	T/G	2112
	C140_2421	A/G	2421
	C150_320	A/G	320
	C152_797	A/G	797
	C158_67*	T/C	67
	C158_238	A/G	238
	C184_1379	A/C	1379
	C229_2772	A/G	2772
	C231_2116*	T/C	2116
	C236_970*	T/G	970
	C250_199	A/G	199
	C253_1545	A/C	1545
	C300_1828	A/G	1828
	C300_4738	A/T	4738
	C300_4982	A/G	4982
	C300_6993	A/G	6993
	C311_406*	A/T	406
	C311_1293	A/C	1293
	C314_1039	A/G	1039
	C327_1076	C/G	1076
	C347_1008	A/T	1008
	C379_2197*	A/T	2197
	C387_215	A/G	215

C387_582	A/G	582
C394_1510	A/G	1510
C421_541	A/G	541
C428_225	A/G	225
C428_306	A/G	306
C428_2101	A/G	2101
C428_2186	A/G	2186
C450_1390**	A/G	1390
C460_1184	A/G	1184
C460_1745	A/C	1745
C549_128*	A/T	128
C570_855	A/T	855
C570_1064	A/C	1064
C618_116	A/G	116
C681_815**	A/G	815
C844_440	A/G	440
C844_751**	A/G	751
C853_1199	A/G	1199
C910_1175	A/G	1175
C911_290	A/T	290
C911_461*	T/C	461
C911_839	A/G	839
C911_1343	A/T	1343
C929_475	A/C	475
C929_734	A/G	734
C929_1311**	A/C	1311
C929_2563	A/G	2563
C980_261	A/G	261
C1001_964**	A/T	964
C1002_85	A/T	85
C1002_557*	A/T	557
C1094_579	A/G	579

C1177_137	A/C	137
C1254_187	A/G	187
C1254_529	A/G	529
C1363_269	A/G	269
C1384_655	A/G	655
C1384_793	A/T	793
C1449_847	A/C	847
C1462_825	A/G	825
C1462_917	A/C	917
C1462_1238*	A/C	1238
C1520_1336	A/C	1336
C1630_199	A/G	199
C1726_472*	T/C	472
C1783_492	A/C	492
C1797_660**	A/G	660
C1797_1023*	A/G	1023
C1797_1098*	A/G	1098
C1813_219**	A/C	219
C1813_300	A/C	300
C1878_506	A/G	506
C2028_1228	A/T	1228
C2028_1328	A/G	1328
C2040_468*	T/C	468
C2040_1251	A/T	1251
C2122_257	A/G	257
C2141_350	A/T	350
C2141_504	A/C	504
C2180_279	A/G	279
C2236_965	A/C	965
C2362_847**	A/T	847
C2406_641	A/G	641
C2558_743	A/G	743

C2735_326	A/C	326
C2899_1354	A/C	1354
C2903_286*	A/G	286
C2903_1043	A/T	1043
C2915_875	A/G	875
C3107_715**	A/G	715
C3220_245	A/T	245
C3495_541	A/G	541
C3676_443	A/T	443
C3835_411	A/G	411
C3914_977	A/C	977
C4144_389*	A/C	389
C4147_533**	C/G	533
C4181_822	A/T	822
C4181_893	A/C	893
C4223_662	A/T	662
C4463_182	A/T	182
C4593_326	A/C	326
C4778_234	A/G	234
C4778_642*	T/C	642
C4791_1099	A/C	1099
C5054_124	A/G	124
C5054_1800*	A/T	1800
C5106_273	A/G	273
C5106_6741	A/G	6741
C5339_366	A/G	366
C5433_233	A/G	233
C5634_234	A/G	234
C5741_659	A/C	659
C6012_280	A/G	280
C6012_652**	C/G	652
C6061_1289	A/C	1289

C6631_237*	A/G	237
C7947_662**	A/G	662
C7947_1013	A/G	1013
C7947_1867	A/G	1867
C8539_132*	T/C	132
C9238_1342*	T/G	1342
C9471_299*	A/G	299
C9511_498*	A/T	498
C10524_242*	A/G	242
C11784_1697	A/G	1697
C12119_299	A/G	299
C13865_165	A/G	165
C14033_777	A/G	777
C15455_325*	T/C	325
C16314_519*	C/G	519
C17220_332	A/G	332
C17943_1521*	T/G	1521
C17963_254*	T/C	254
C17963_499*	T/G	499
C18184_390	A/G	390
C18679_62	A/G	62
C18774_676	A/G	676
C18774_877*	A/G	877
C18782_153**	A/G	153
C19500_354*	T/C	354
C19952_651	A/G	651
C20003_92*	A/G	92
C20142_203*	T/G	203
C20174_2657	A/G	2657
C20267_102	A/G	102
C20267_179*	A/G	179
C20427_267	A/G	267

	C20580_186**	A/G	186
	C20648_3041	A/G	3041
	C20682_843	A/T	843
	C20776_135*	T/C	135
	C21068_302	A/G	302
	C21134_331*	T/C	331
	C21581_136*	T/C	136
	C21593_75*	T/G	75
	C21645_185*	T/G	185
	C21673_148*	A/G	148
	C21706_435*	T/C	435
	C21880_565	A/G	565
	C21880_1003	C/G	1003
	C21880_1111	A/G	1111
	C21908_64	A/C	64
	C22317_403	A/T	403
	C22340_159*	T/C	159
	C22347_319	A/G	319
	C22431_278	A/G	278
	C22449_261	A/G	261
	C22491_341	A/G	341
	C22491_595	A/G	595
	C22491_727	A/G	727
	C22521_249	A/T	249
	C22537_1941	A/G	1941
	C22568_597	A/G	597
	C22574_507	A/C	507
	C22635_318	A/G	318
	SNP name	Variant	SNP position
Positive controls	3B4_7	A/T	492
	SNP146.2_132	A/G	132
	SNP149.1_374	C/G	374

	SNP1834_464**	A/G	464
	SNP1949_235	A/C	235
	SNP449.2_110	A/G	110

* SNPs that failed to cluster correctly

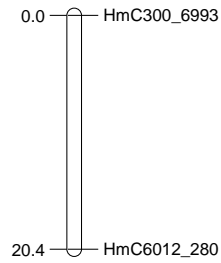
** Monomorphic SNPs

Addendum 4: Number of markers per linkage group, length of linkage groups, average marker spacing and largest interval for maternal (P1), sex-average (POP) and paternal (P2) maps of family H.

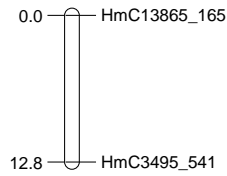
Linkage group	No. of markers			Length (cM)			Ave. spacing (cM)			Largest interval (cM)		
	P1	POP	P2	P1	POP	P2	P1	POP	P2	P1	POP	P2
1	-	2	-	-	20.4	-	-	20.4	-	-	20.4	-
2	2	2	2	12.8	25.5	12.8	12.8	25.5	12.8	12.8	25.5	12.8
4	-	2	2	-	0.0	0.0	-	0.0	0.0	-	0.0	0.0
5	-	2	2	-	26.0	26.0	-	26.0	26.0	-	26.0	26.0
8	5	5	-	38.1	38.9	-	9.5	9.7	-	14.3	14.4	-
11	2	2	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	4	7	5	28.1	49.8	37.6	9.4	8.3	9.4	13.8	19.5	23.4
16	2	2	-	12.5	12.5	-	12.5	12.5	-	12.5	12.5	-
17	2	2	-	14.7	14.7	-	14.7	14.7	-	14.7	14.7	-
Total	17.0	26.0	13.0	106.2	187.8	76.4	58.9	117.1	48.2	68.1	133.0	62.2
Average	2.8	2.9	2.6	17.7	20.9	15.3	9.8	13.0	9.6	11.4	14.8	12.4

Addendum 5: Maternal (P1), sex-average (POP) and paternal (P2) maps of family H.

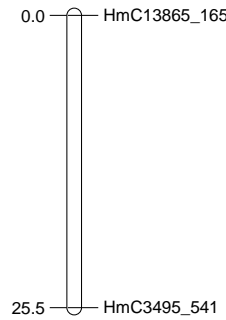
POP_LG_1



P1_LG_2



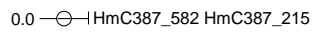
POP_LG_2



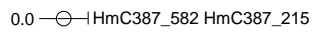
P2_LG_2



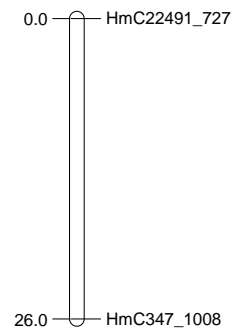
POP_LG_4



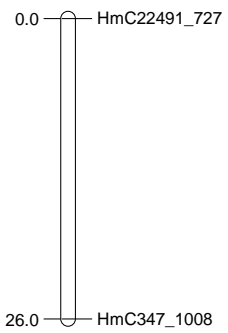
P2_LG_4



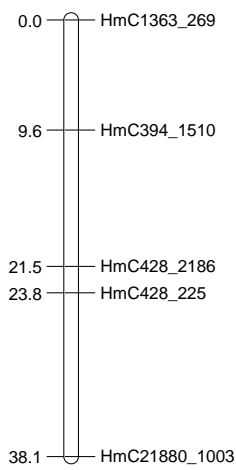
POP_LG_5



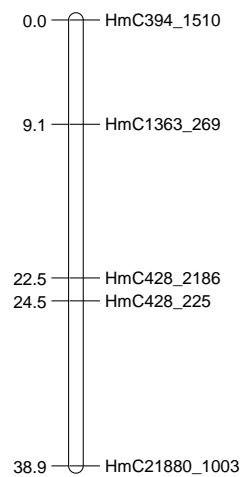
P2_LG_5



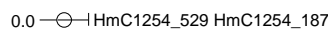
P1_LG_8



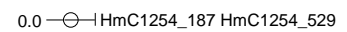
POP_LG_8



P1_LG_11



POP_LG_11



P2_LG_11

0.0 —○—| HmC1254_529 HmC1254_187

P1_LG_13

0.0 —○—| HmC2915_875
 13.8 —| HmC853_1199
 23.2 —| HmC2141_504
 28.1 —○—| HmC22635_318

POP_LG_13

0.0 —○—| HmC2915_875
 19.5 —| HmC853_1199 HmSNP449.2_110
 26.3 —| HmC2141_350 HmC2141_504
 34.8 —| HmC22635_318
 49.8 —○—| HmSNP1949_235

P2_LG_13

0.0 —○—| HmSNP1949_235
 23.4 —| HmC2141_504
 23.7 —| HmC2141_350
 37.3 —| HmSNP449.2_110
 37.6 —| HmC853_1199

P1_LG_16

0.0 —○—| HmC22449_261
 12.5 —○—| HmC5106_273

POP_LG_16

0.0 —○—| HmC22449_261
 12.5 —○—| HmC5106_273

P1_LG_17

0.0 —○—| HmC5741_659
 14.7 —○—| HmC929_475

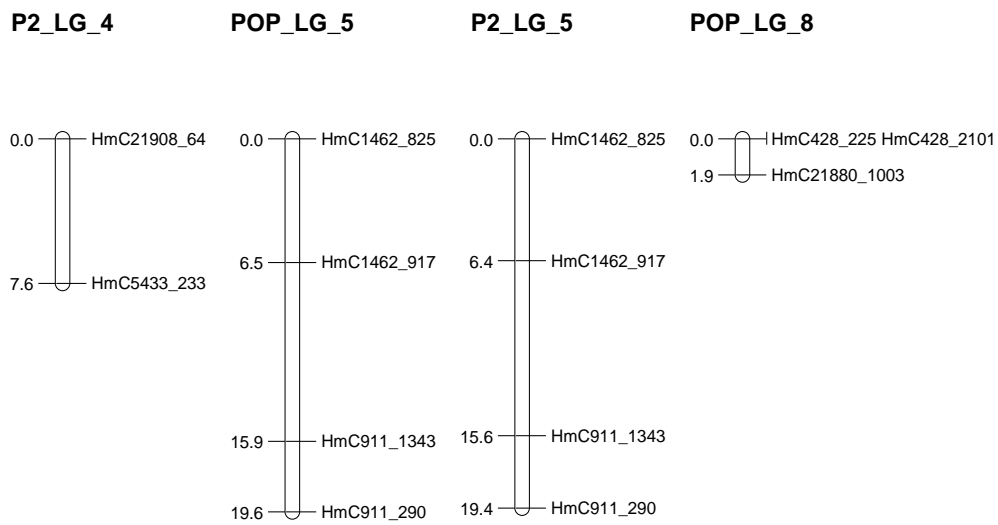
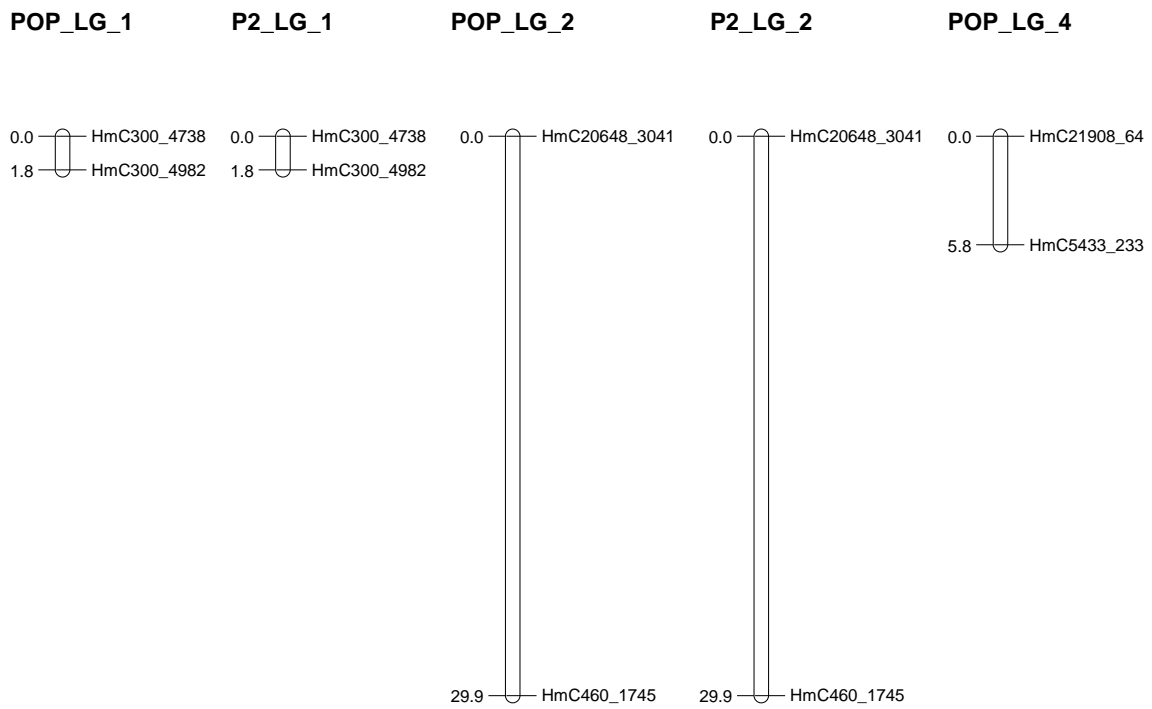
POP_LG_17

0.0 —○—| HmC5741_659
 14.7 —○—| HmC929_475

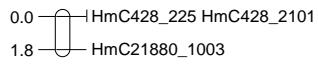
Addendum 6: Number of markers per linkage group, length of linkage groups, average marker spacing and largest interval for maternal (P1), sex-average (POP) and paternal (P2) maps of family I.

Linkage group	No. of markers			Length (cM)			Ave. spacing (cM)			Largest interval (cM)		
	P1	POP	P2	P1	POP	P2	P1	POP	P2	P1	POP	P2
1	-	2	2	-	1.8	1.8	-	1.8	1.8	-	1.8	1.8
2	-	2	2	-	29.9	29.9	-	29.9	29.9	-	29.9	29.9
4	-	2	2	-	5.8	7.6	-	5.8	7.6	-	5.8	7.6
5	-	4	4	-	19.6	19.4	-	6.5	6.5	-	9.4	9.2
8	-	3	3	-	1.9	1.8	-	1.0	0.9	-	1.9	1.8
11	3	3	-	25.1	25.0	-	12.6	12.5	-	25.1	25.0	-
13	3	6	4	6.2	30.3	26.9	3.1	6.1	9.0	6.2	14.1	14.1
15	2	2	-	9.7	9.7	-	9.7	9.7	-	9.7	9.7	-
16	-	3	3	-	27.0	26.5	-	13.5	13.3	-	18.8	18.0
19	2	2	-	30.1	30.1	-	30.1	30.1	-	30.1	30.1	-
Total	10.0	29.0	20.0	71.1	181.1	113.9	55.5	116.8	68.9	71.1	146.5	82.4
Average	2.5	2.9	2.9	17.8	18.1	16.3	13.9	11.7	9.8	17.8	14.7	11.8

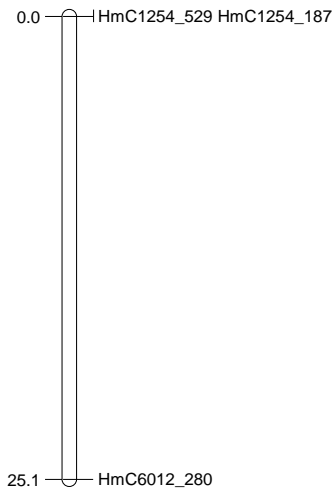
Addendum 7: Maternal (P1), sex-average (POP) and paternal (P2) maps of family I.



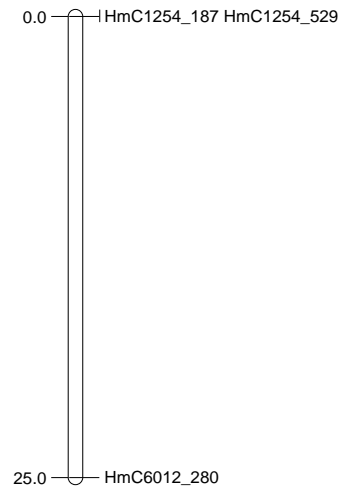
P2_LG_8



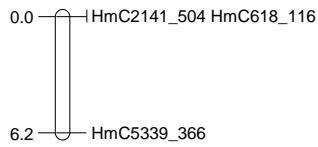
P1_LG_11



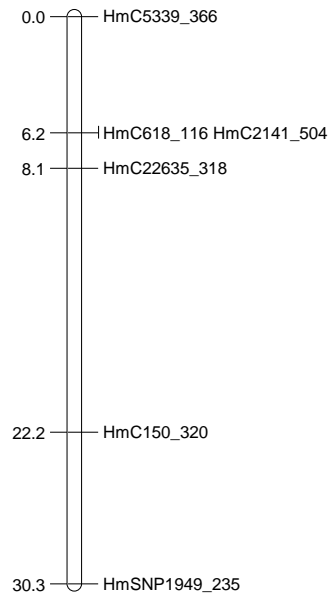
POP_LG_11



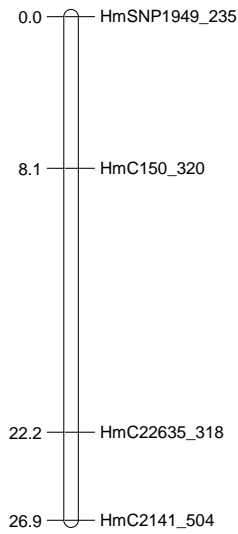
P1_LG_13



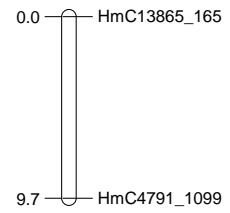
POP_LG_13



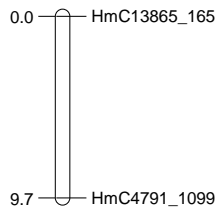
P2_LG_13



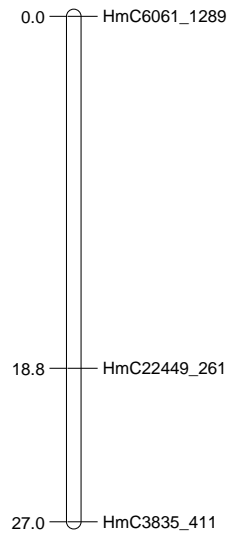
P1_LG_15



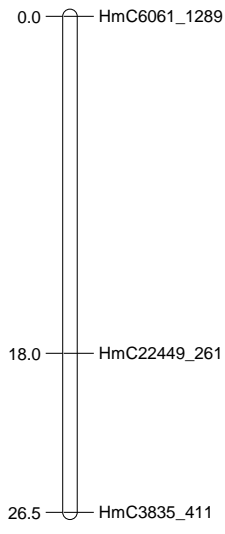
POP_LG_15



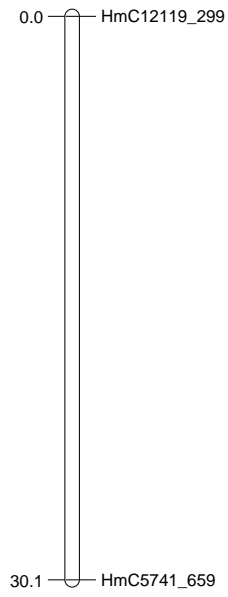
POP_LG_16



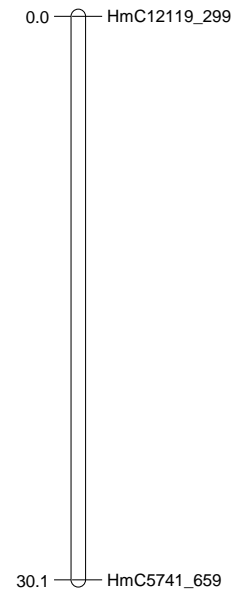
P2_LG_16



P1_LG_19



POP_LG_19

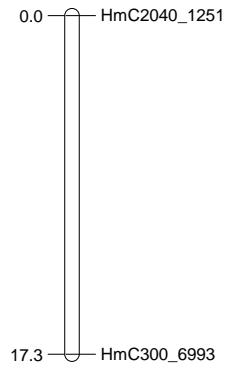


Addendum 8: Number of markers per linkage group, length of linkage groups, average marker spacing and largest interval for maternal (P1), sex-average (POP) and paternal (P2) maps of family J.

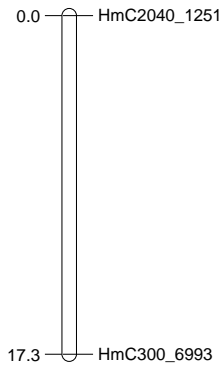
Linkage group	No. of markers			Length (cM)			Ave. spacing (cM)			Largest interval (cM)		
	P1	POP	P2	P1	POP	P2	P1	POP	P2	P1	POP	P2
1	-	2	2	-	17.3	17.3	-	17.3	17.3	-	17.3	17.3
2	-	2	2	-	16.9	16.9	-	16.9	16.9	-	16.9	16.9
4	3	3	-	12.7	10.7	-	6.4	5.4	-	10.8	8.8	-
5	2	2	-	8.8	4.6	-	8.8	4.6	-	8.8	4.6	-
8	5	6	4	22.9	12.2	0.2	5.7	2.4	0.1	14.3	7.0	0.2
13	2	4	5	16.3	7.9	9.4	16.3	2.6	2.4	16.3	4.0	4.6
16	-	2	2	-	28.9	28.9	-	28.9	28.9	-	28.9	28.9
20	-	2	2	-	18.7	18.0	-	18.7	18.0	-	18.7	18.0
21	-	2	2	-	28.1	28.1	-	28.1	28.1	-	28.1	28.1
Total	12.0	25.0	19.0	60.7	145.3	118.8	37.2	124.9	111.6	50.2	134.3	114.0
Average	3.0	2.8	2.7	15.2	16.1	17.0	9.3	13.9	15.9	12.6	14.9	16.3

Addendum 9: Maternal (P1), sex-average (POP) and paternal (P2) maps of family J.

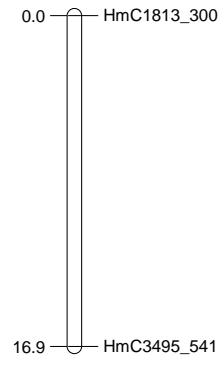
POP_LG_1



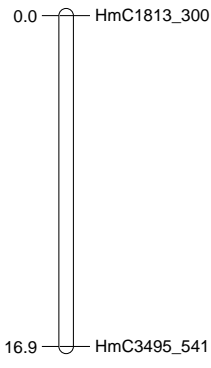
P2_LG_1



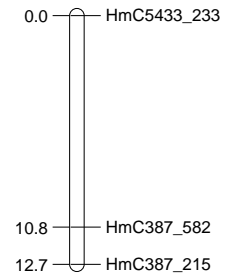
POP_LG_2



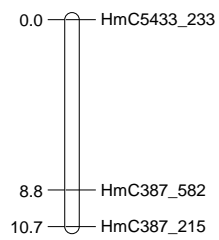
P2_LG_2



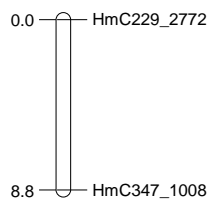
P1_LG_4



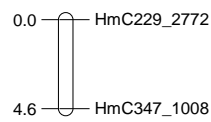
POP_LG_4



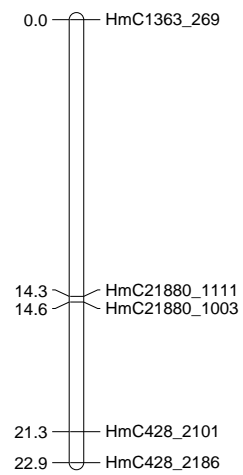
P1_LG_5



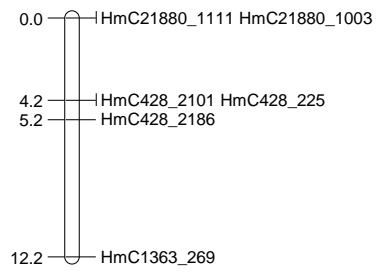
POP_LG_5



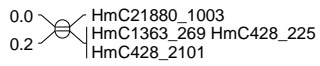
P1_LG_8



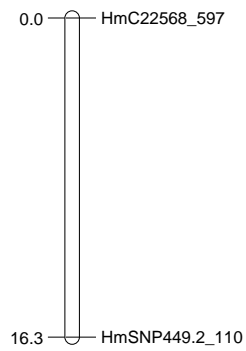
POP_LG_8



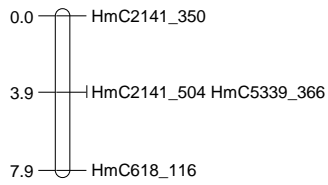
P2_LG_8



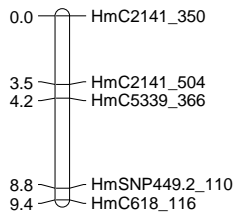
P1_LG_13



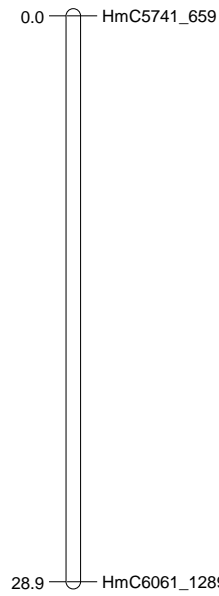
POP_LG_13



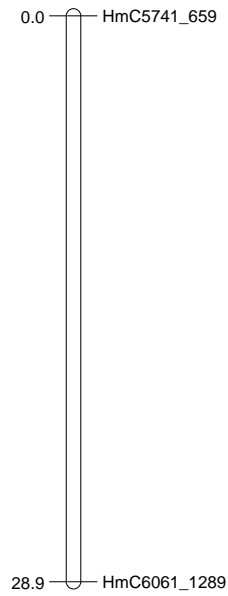
P2_LG_13



POP_LG_16



P2_LG_16

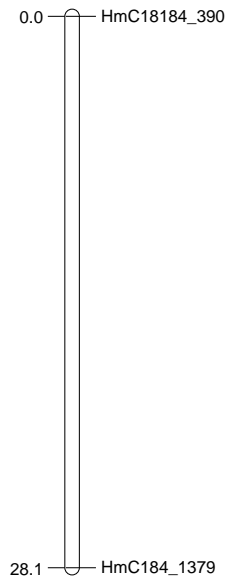
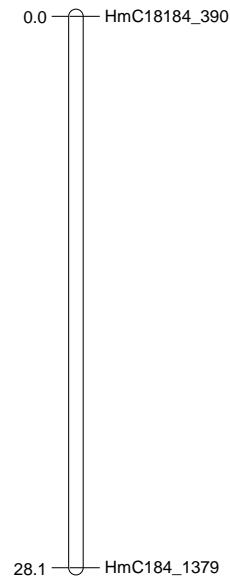
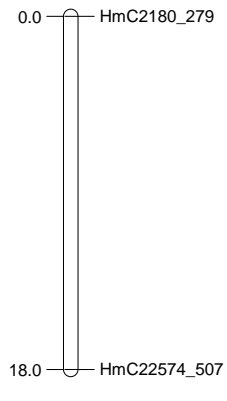
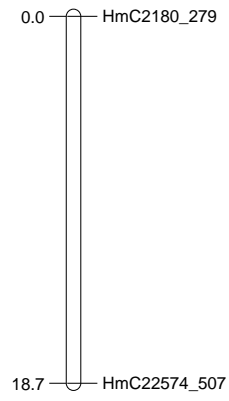


POP_LG_20

P2_LG_20

POP_LG_21

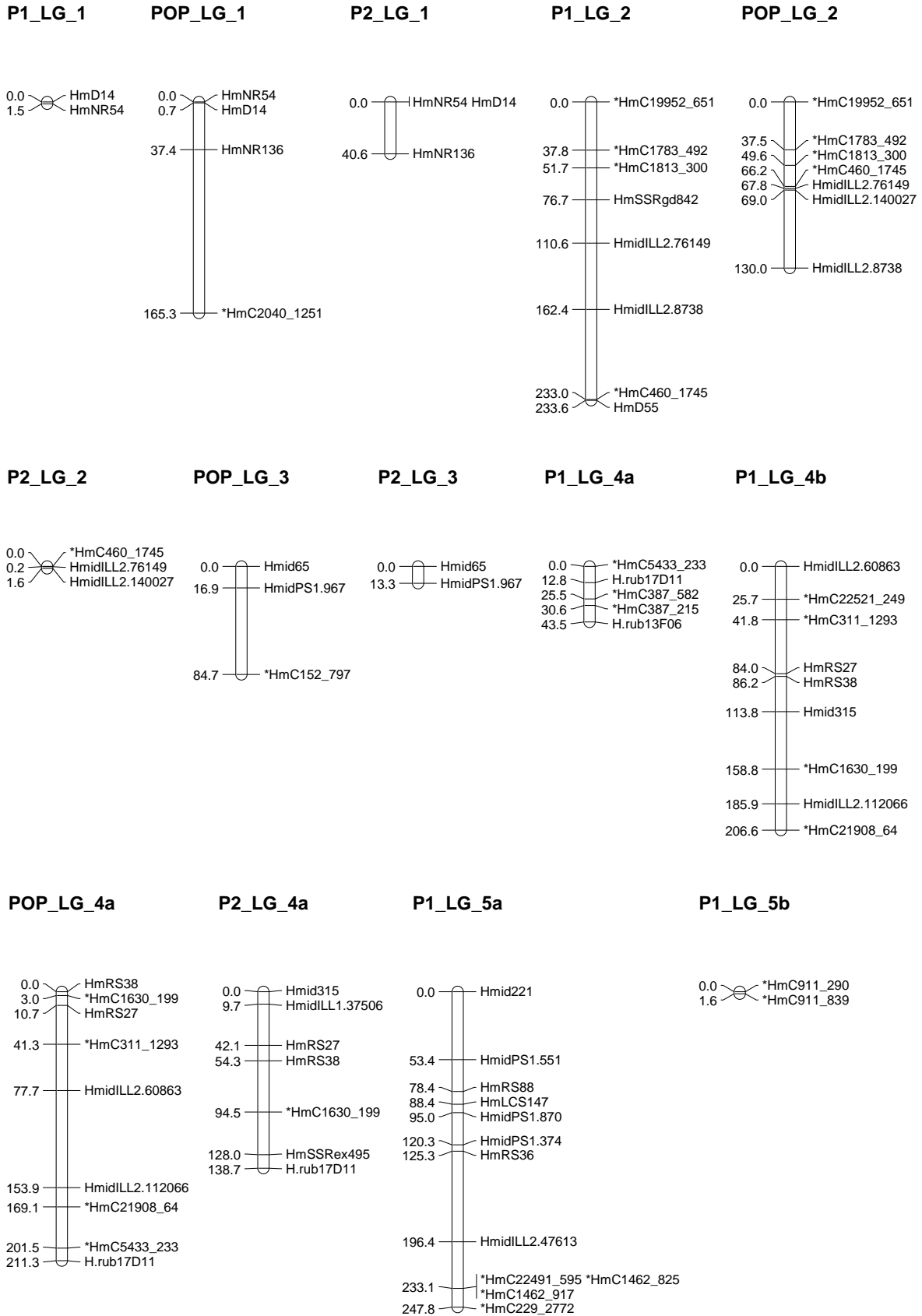
P2_LG_21



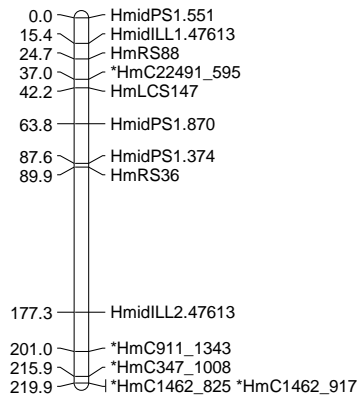
Addendum 10: Number of markers per linkage group, length of linkage groups, average marker spacing and largest interval for maternal (P1), sex-average (POP) and paternal (P2) maps of family DS2.

Linkage group	No. of markers			Length (cM)			Ave. spacing (cM)			Largest interval (cM)		
	P1	POP	P2	P1	POP	P2	P1	POP	P2	P1	POP	P2
1	2	4	2	1.5	165.3	40.6	1.5	55.1	40.6	1.5	127.9	40.6
2	8	7	3	233.6	130.0	1.6	33.4	21.7	0.8	70.6	61.0	1.4
3	-	3	2	-	84.7	13.3	-	42.4	13.3	-	67.8	13.3
4a	5	9	7	43.5	211.3	138.7	10.9	26.4	23.1	13.0	76.2	40.2
4b	9	-	-	206.6	-	-	25.8	-	-	45.0	-	-
5a	12	13	7	247.8	219.9	103.0	22.5	18.3	17.2	71.1	87.4	36.7
5b	2	-	-	1.6	-	-	1.6	-	-	1.6	-	-
6	-	2	-	-	87.7	-	-	87.7	-	-	87.7	-
7	3	7	-	6.7	257.5	-	3.4	42.9	-	6.6	108.9	-
8a	11	14	5	324.0	327.8	125.8	32.4	25.2	31.5	104.6	80.8	60.6
8b	-	-	7	-	-	101.1	-	-	16.9	-	-	33.4
9a	2	2	2	7.2	7.1	15.5	7.2	7.1	15.5	7.2	7.1	15.5
9b	2	2	-	9.2	13.3	-	9.2	13.3	-	9.2	13.3	-
10	7	10	5	168.9	345.6	50.0	28.2	38.4	12.5	82.8	116.0	23.9
11	5	7	-	42.2	175.4	-	10.6	29.2	-	38.7	90.8	-
12	4	5	3	18.9	178.9	53.2	6.3	44.7	26.6	10.1	145.3	35.6
13	10	11	8	198.7	246.0	178.2	22.1	24.6	25.5	140.7	95.0	126.3
14	6	5	3	177.0	118.8	34.7	35.4	29.7	17.4	57.3	55.3	25.5
15	-	4	4	-	76.4	75.6	-	25.5	25.2	-	67.0	64.4
16	8	9	4	107.7	153.8	65.7	15.4	19.2	21.9	63.4	96.8	34.5
17	3	5	2	11.2	79.7	9.8	5.6	19.9	9.8	6.7	68.7	9.8
18	-	2	-	-	93.0	-	-	93.0	-	-	93.0	-
26	2	-	-	1.6	-	-	1.6	-	-	1.6	-	-
27	2	4	3	14.7	150.7	43.0	14.7	50.2	21.5	14.7	103.0	25.9
28	2	2	-	1.6	1.5	-	1.6	1.5	-	1.6	1.5	-
29	-	-	2	-	-	25.2	-	-	25.2	-	-	25.2
Total	105.0	127.0	69.0	1824.2	3124.4	1075.0	289.2	716.1	344.3	748.0	1650.5	612.8
Average	5.3	6.0	4.1	91.2	148.8	63.2	14.5	34.1	20.3	37.4	78.6	36.0

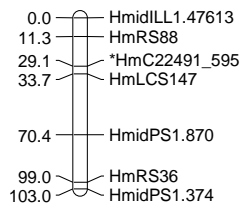
Addendum 11: Maternal (P1), sex-average (POP) and paternal (P2) maps of family DS2.



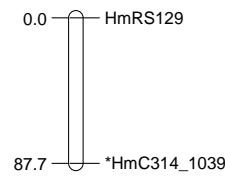
POP_LG_5a



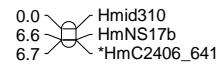
P2_LG_5a



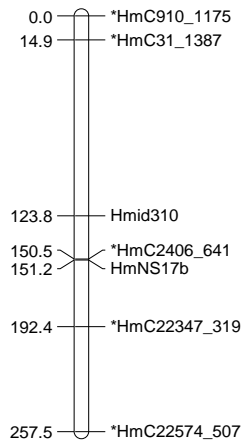
POP_LG_6



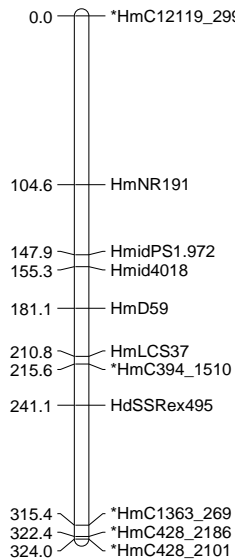
P1_LG_7



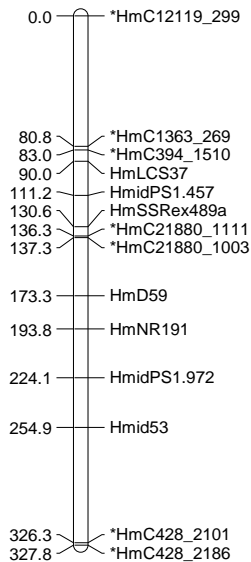
POP_LG_7



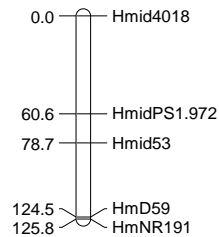
P1_LG_8a



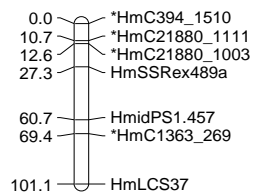
POP_LG_8a



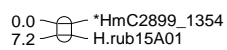
P2_LG_8a



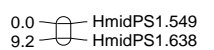
P2_LG_8b



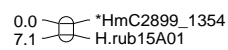
P1_LG_9a



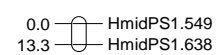
P1_LG_9b



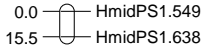
POP_LG_9a



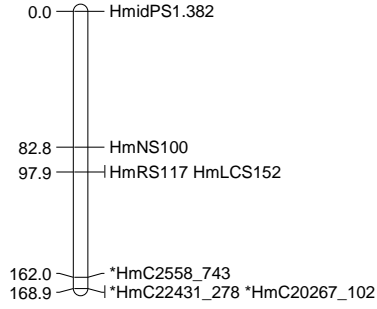
POP_LG_9b



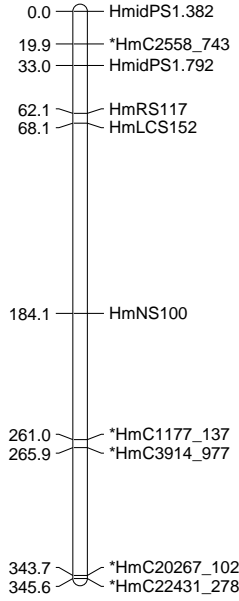
P2_LG_9a



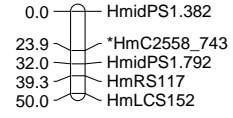
P1_LG_10



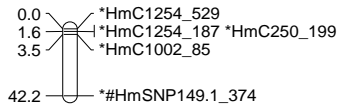
POP_LG_10



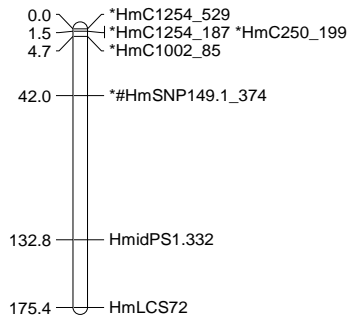
P2_LG_10



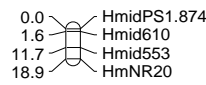
P1_LG_11



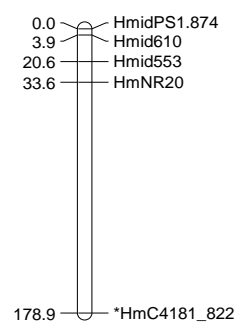
POP_LG_11



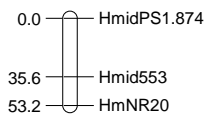
P1_LG_12



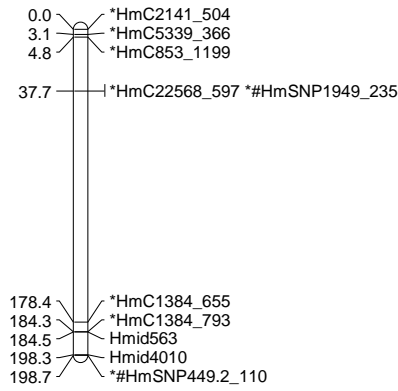
POP_LG_12



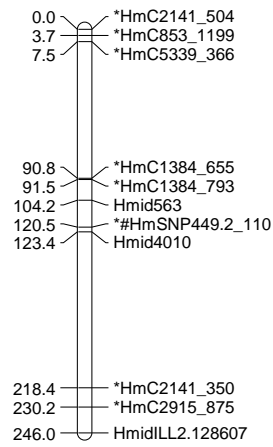
P2_LG_12



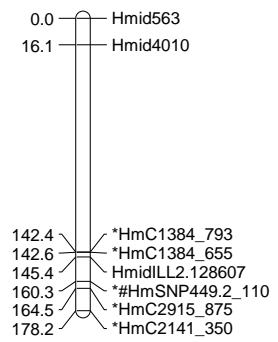
P1_LG_13



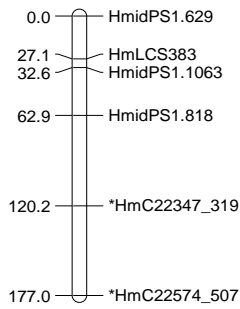
POP_LG_13



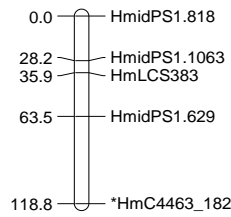
P2_LG_13



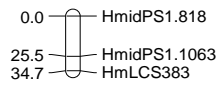
P1_LG_14



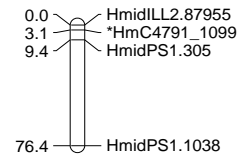
POP_LG_14



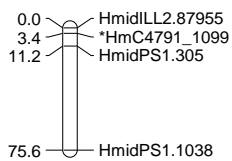
P2_LG_14



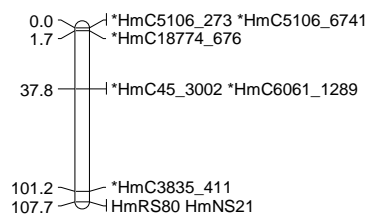
POP_LG_15



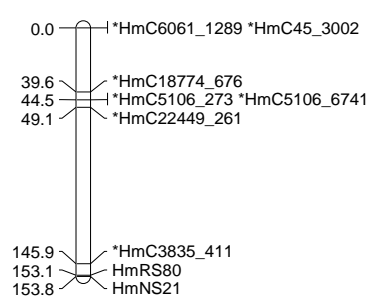
P2_LG_15



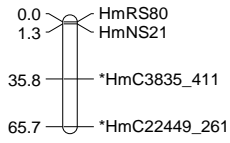
P1_LG_16



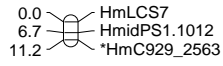
POP_LG_16



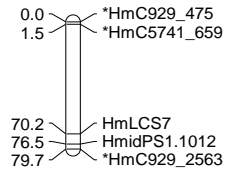
P2_LG_16



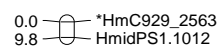
P1_LG_17



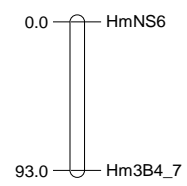
POP_LG_17



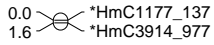
P2_LG_17



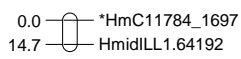
POP_LG_18a



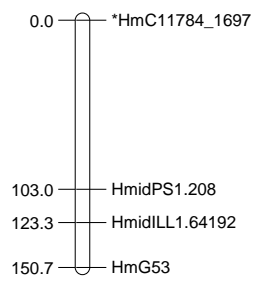
P1_LG_26



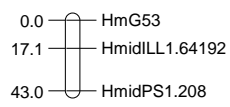
P1_LG_27



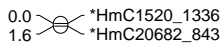
POP_LG_27



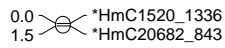
P2_LG_27



P1_LG_28



POP_LG_28



P2_LG_29

