Linkage Disequilibrium in the South African Abalone, *Haliotis midae*

by

Ruth Dale Kuys

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

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

Co-Supervisor: Rouvay Roodt-Wilding, Ph.D.

Department of Genetics

December 2015

Declaration

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

December 2015

Abstract

Linkage disequilibrium (LD) is defined as the non-random association of alleles at two or more loci within a population. It is sensitive to a variety of locus-specific- and demographic factors, and can thus provide much insight into the micro-evolutionary factors that have shaped species of interest. It can also be exploited to identify the genomic regions determining complex traits of interest, which can then be applied as performance evaluation markers in marker-assisted selection (MAS). The South African abalone, *Haliotis midae*, supports a rapidly developing aquaculture production industry, in which genetic improvement potential is high. This species also represents an opportunistic model for studying the effects of early domestication in a shellfish species. The aim of this study was therefore to quantify and characterise levels of genome-wide LD within the South African abalone, and to demonstrate its utility within population genetic investigations and the characterisation of complex traits. Estimates of LD between 112 mapped microsatellite markers within wild and cultured *H. midae* revealed that levels of LD in abalone are high relative to other aquaculture species. This was attributed primarily to small effective population sizes produced by a combination of natural- and anthropogenic factors. The decay of LD with genetic distance was evident in both cultured cohorts, but almost absent in wild cohorts, likely reflecting the differences in size, age and sampling of wild populations relative to cultured. Putative evidence for the effects of recombination, selection, and epistasis were also evident in distinctive locus-specific patterns of LD on some of the linkage groups, many of which could represent the effects of domestication. The effects of selection associated with the domestication event were further investigated using a candidate locus LD mapping approach to determine the proportion of candidate loci under selection associated with artificial selection for faster growth rate in cultured abalone. Two loci (15%) were found to be significantly associated with differences in size of individual animals, both of which could be linked with genes potentially involved in growth and development. These markers could therefore find application in MAS programmes for abalone. Several promising candidates for natural selection were also identified based on similarity with known genes. As the latter represented the majority, natural selection, rather than artificial selection, appears to be predominant during the early stages of domestication in abalone. While some conclusions within the current study were speculative, both the direct and indirect applications of LD were clearly demonstrated. Linkage disequilibrium data can provide a unique perspective on many of the commonly used population genetic estimates, and is therefore of great value in

population genetic investigations. Furthermore, these results also highlighted the effectiveness of the candidate locus approach in species with both limited molecular resources and extensive LD.

Opsomming

Koppelingsonewewig (KO) word gedefinieer as die nie-lukrake assosiasie van allele by twee of meer lokusse binne 'n populasie. Koppelingsonewewig is sensitief vir 'n verskeidenheid van lokus-spesifieke- en demografiese faktore, en kan dus insiggewend wees m.b.t. mikro-evolusionêre faktore wat spesies van belang beïnvloed het. Dit kan ook benut word om die genoom-gebiede onderligend tot komplekse eienskappe te bespeur; wat dan aangewend kan word vir prestasie-evaluering m.b.v. merkerbemiddelde seleksie (MBS). Die Suid-Afrikaanse perlemoen, *Haliotis midae*, ondersteun 'n vinnig ontwikkelende akwakultuur produksie bedryf, waarin genetiese verbeteringspotensiaal hoog is. Hierdie spesie verteenwoordig ook 'n opportunistiese model vir die bestudering van die gevolge van vroeë domestiseering in 'n skulpvis spesie. Die doel van hierdie studie was dus om vlakke van genoom-wye KO binne die Suid-Afrikaanse perlemoen te kwantifiseer en te karakteriseer, en om die toepassing hiervan binne populasiegenetiese ondersoeke en die karakterisering van komplekse eienskappe te demonstreer. Ramings van KO tussen 112 gekarteerde mikrosatelliet-merkers binne wilde en gekultiveerde *H. midae* het aan die lig gebring dat die vlakke van KO in perlemoen hoog was, in vergelyking met ander akwakultuur spesies. Dit word hoofsaaklik toegeskryf aan klein effektiewe populasiegroottes wat deur 'n kombinasie van natuurlike- en antropogeniese faktore teweeg gebring word. Die verval van KO met genetiese afstand was duidelik waarneembaar in gekultiveerde kohorte, maar amper afwesig in die wilde kohorte, waarskynlik a.g.v. verskille in populasiegrootte, ouderdom, en streekproef-neemings metodieke van die verskeie populasies. Vermeende bewyse vir die gevolge van rekombinasie, seleksie en epistase kon ook gesien word a.g.v. lokus-spesifieke patrone van KO op sommige van die koppelingsgroepe, moontlik ‘n gevolg van domestisering. Die gevolge van seleksie wat verband hou met die domestiseringsgebeurtenis is verder ondersoek m.b.v 'n kandidaat-lokus KO karteringsbenadering om die verhouding van kandidaat lokusse wat geassosieer is met kunsmatige seleksie (vir vinniger groeikoers in perlemoen) te bepaal. Twee lokusse (15%) was beduidend geassosieer met verskille in grootte tussen individuele diere. Beide van die lokusse was gekoppel met gene wat potensieel betrokke is by groei en ontwikkeling. Hierdie merkers kan dus moontlik aangewend word in MBS programme vir perlemoen. Verskeie belowende kandidaat lokusse vir natuurlike seleksie is ook geïdentifiseer gebaseer op ooreenkoms met bekende gene. Gegewe dat die laasgenoemde die meerderheid van die merkers verteenwoordig, kan daar afgelei word dat natuurlike seleksie, eerder as kunsmatige seleksie, oorheersend

is in die vroeë stadia van domestisering in perlemoen. Terwyl sommige gevolgtrekkings binne die huidige studie spekulatief was, is beide die direkte en indirekte toepassings van KO duidelik gedemonstreer. Koppelingsonewewig-data kan 'n unieke perspektief gee op baie van die algemeen gebruikte populasie genetiese skattings, en is dus van groot waarde in populasie genetiese ondersoeke. Verder demonstreer hierdie resultate ook die doeltreffendheid van die kandidaat lokus benadering in spesies met beide beperkte molekulêre hulpbronne en uitgebreide KO.

Acknowledgements

I would like to thank the following institutions for their contributions to this study (in alphabetical order): Atlantic Sea Farm (Pty) Ltd, the Central Analytical Facility, I&J Danger Point Abalone Farm (Pty) Ltd, the National Research Foundation, Stellenbosch University, and Wild Coast Abalone (Pty) Ltd. I would also like to thank the following people for their academic guidance and encouragement: my supervisor, Dr Clint Rhode, and co-supervisor, Prof. Rouvay Roodt-Wilding (I hope I do all your faith in me justice), our lab manager, Jessica Vervalle (you are the rock on which we all stand), and fellow members of the Molecular Breeding and Biodiversity research group, especially Charné Rossouw, Shaun Lesch and William Versfeld (you shared the drama with me and made me laugh at myself). Lastly, I would like to acknowledge the non-academic support of friends and family, who all shouted the loudest when the finish line was close, but seemed so far away. Every one of you was there when I came to the end of my tether, and you made me stronger than I ever was on my own. Thank you for having faith in me when I had none in myself, and for all the prayers you offered on my behalf.

Table of Contents

Chapter 2: Genome-wide Linkage Disequilibrium in Abalone

Chapter 3: Association Analysis of Candidate Loci under Selection with Size in the South African Abalone

List of Figures

List of Tables

List of Abbreviations

%	Percentage
>	Greater than
<	Less than
≥	Greater than or equal to
~	Approximately
∞	Infinity
±	Plus-minus
5'	Five prime
3'	Three prime
2n	Diploid chromosome number
ABC	Adenosine triphosphate-binding cassette
A_e	Effective number of alleles
AMOVA	Analysis of molecular variance
A_n	Number of alleles
ASF	Atlantic Sea Farm cohort
ATP	Adenosine triphosphate
b_j	Coefficient of linkage disequilibrium decay
BLAST	Basic local alignment search tool
Bsl	Baseline value
c	Recombination rate
°C	Degrees Celsius
CI/s	Confidence interval/s
cM	centiMorgan
CTAB	Cetyltrimethylammonium bromide
CV	Coefficient of variance
D'	Multi-allelic extension of Lewontin's standardised linkage disequilibrium coefficient
D'_{ij}	Lewontin's standardised linkage disequilibrium coefficient
D_{ABC}	Linkage disequilibrium coefficient for higher-order disequilibria
DAFF	Department of Agriculture, Forestry and Fisheries
Df	Degrees of freedom
D_{ij}	Linkage disequilibrium coefficient

$D'_{IS}{}^2$	Ohta's D-statistic (variance of the correlation of genes of the two loci of one gamete in a subpopulation relative to that of the total population)
$D'_{ST}{}^2$	Ohta's D-statistic (variance of the disequilibrium of the total population)
$D_{IS}{}^2$	Ohta's D-statistic (variance of within-subpopulation disequilibrium)
$D_{IT}{}^2$	Ohta's D-statistic (total variance of disequilibrium)
DNA	Deoxyribonucleic acid
$D_{ST}{}^2$	Ohta's D-statistic (variance of the correlation of genes of the two loci of different gametes of one subpopulation relative to that of the total population)
e.g.	*exempli gratia* (for example)
EST	Expressed sequence tag
et al.	*et alii* (and others)
E-value	Expect value
EW	Ewens-Watterson
F_1	First generation
F_2	Second generation
FAO	Food and Agriculture Organisation
FBC	Family-bias corrected cohort
F_{IS}	Wright's fixation index (individual relative to the sub-population)
F_{IT}	Wright's fixation index (individual relative to the total population)
F_{SC}	Derivative of Wright's fixation index adapted for hierarchical AMOVA (sub-population relative to the group of populations)
F_{ST}	Wright's fixation index (subpopulation relative to the total population)
g	Grams
G''_{ST}	Hedrick's standardised G_{ST}, corrected for bias when number of populations is small
GAS	Gene-assisted selection
GSL	Glucosinolate
GWAS	Genome-wide association study
HIV-1	Human immunodeficiency virus type 1
H_o	Observed heterozygosity
HW	Hardy-Weinberg
I	Shannon's information index
I&J	Irvin and Johnson
I_A	Index of association
i.e.	*id est* (that is to say)

KW	Kruskal-Wallis
LD	Linkage disequilibrium
LG	Linkage group
MAS	Marker-assisted selection
Max	Maximum
Min	Minimum
mm	Millimeters
n	Sample size
N_e	Effective population size
PCoA	Principal coordinate analysis
PCR	Polymerase chain reaction
(Pty) Ltd	Proprietary limited
P-value	Probability value
QTL/s	Quantitative trait locus/loci
®	Registered trademark
r	Relatedness
R^2	Squared correlation coefficient
r_{ij}^2 or r^2	Squared correlation coefficient between alleles at two loci
RNA	Ribonucleic acid
RP	Riet Point cohort
RSA	Republic of South Africa
SAL	Saldana Bay cohort
SD	Standard deviation
SNP	Single nucleotide polymorphism
SSD	Sum of squared differences
TAC	Total allowable catch
™	Trademark
uH_e	Unbiased expected heterozygosity
UTR	Untranslated region
v	Version
WCA	Wild Coast Abalone cohort
χ^2	Chi-squared statistic
$\chi^{2\prime}$	Standardised χ^2 statistic as a measure of linkage disequilibrium

This thesis is dedicated to the extraordinary “luck” that led a humble man to grow pea plants and change the world.

Chapter 1
Literature Review and Introduction

"Linkage disequilibrium (LD) is one of those unfortunate terms that does not reveal its meaning." (Slatkin 2008)

1.1) Linkage Disequilibrium: A Brief History and Explanation

1.1.1) Mendel and the modern evolutionary synthesis

Biological populations are diverse; a characteristic which has long been the focus of a number of different fields within the biological sciences. Population genetics aims to investigate the extent to which genetic diversity is responsible for creating and maintaining biological diversity. Within this broad aim, particular emphasis is placed on elucidating the various molecular genetic mechanisms that determine traits, as well as the manner and extent to which environmental factors direct changes in genetic diversity. However, at the core of such studies remains a thorough understanding of the fundamental concepts governing particulate inheritance, *i.e.* how genetic material is physically arranged and inherited.

Even before it was known that DNA is the genetic material and that genes are arranged on structurally distinct chromosomes, Mendel (1866) determined that traits are inherited in a specific and predictable manner by conducting controlled breeding experiments in pea plants. Using the phenotypic data from the various crosses he conducted, he formulated the expected genotypic frequencies for one and two locus combinations, on which all modern population genetics is based, and which he used to construct his four postulates:

i) *Unit factors occur in pairs:* Genetic characteristics are determined by "unit factors", or genes, that exist in pairs within each individual.

ii) *Different forms of the same factor are either dominant or recessive*: When two unlike factors, or alleles, for a single trait are present within a single individual, one is expressed (dominant), while the other is not (recessive).

iii) *Pairs of alleles segregate randomly:* During gamete formation, each pair of alleles segregates randomly into different gametes, so that each gamete only carries one copy or the other.

iv) *Different pairs of alleles assort independently:* When considering more than one trait, each pair of alleles will segregate independently of each other pair of alleles during this process.

After its rediscovery in 1900 (Correns 1900; De Vries 1900; von Tschermak 1900), Mendel's work not only revolutionised our understanding of the inheritance of biological traits, but also paved the way for a number of equally fundamental breakthroughs, allowing the amalgamation of theories of heredity and evolution. However, although numerous later studies confirmed many of Mendel's findings, many others also reported inconsistencies with Mendelian principles (Miko 2008; Hill 2009; Smýkal 2014). For example, according to Mendel's postulate of independent assortment, each copy of a gene should be assigned to a particular gamete in a random manner relative to every other gene. However, Bateson *et al*. (1905) noticed in their work with pea plants, that not all of their crosses produced results that were consistent with this principle. In particular, certain combinations of alleles appeared far more frequently than predicted by Mendelian genetics. As a result, the authors suspected that certain allelic variants had to be linked in some manner, although it was not until after Morgan's (1910, 1911) work on the chromosome theory, and subsequent discovery of genetic recombination, that the phenomenon of genetic linkage was fully elucidated. Because he observed independent assortment in his experiments, Mendel's view of genetic material was that of independently inherited "unit factors"; however, the way in which genetic material is actually "packaged" makes this scenario often untenable. Chromosomes containing genes, and not genes themselves, are the units of transmission during gamete formation, resulting in genes on the same chromosome tending to assort together, rather than randomly. However, what prevents these genes from being permanently linked, and potentially explains Mendel's observations (Blixt 1975; Smýkal 2014), is the process of genetic recombination, which occurs during the early stages of gamete formation (meiosis), before chromosomes begin segregating. During this process, the arms of homologous chromosomes make contact at random points, called chiasmata. Crossing over then occurs, leading to an exchange of short segments of DNA, and resulting in "chimeras" of both ancestral chromosomes (recombinants) (Figure 1.1a-c). Naturally, if crossing over does not take place, the chromosomes remain unchanged (non-recombinants), which results in a mixture of recombinant and non-recombinant chromosomes that segregate randomly into gametes (Figure 1.1d-e).

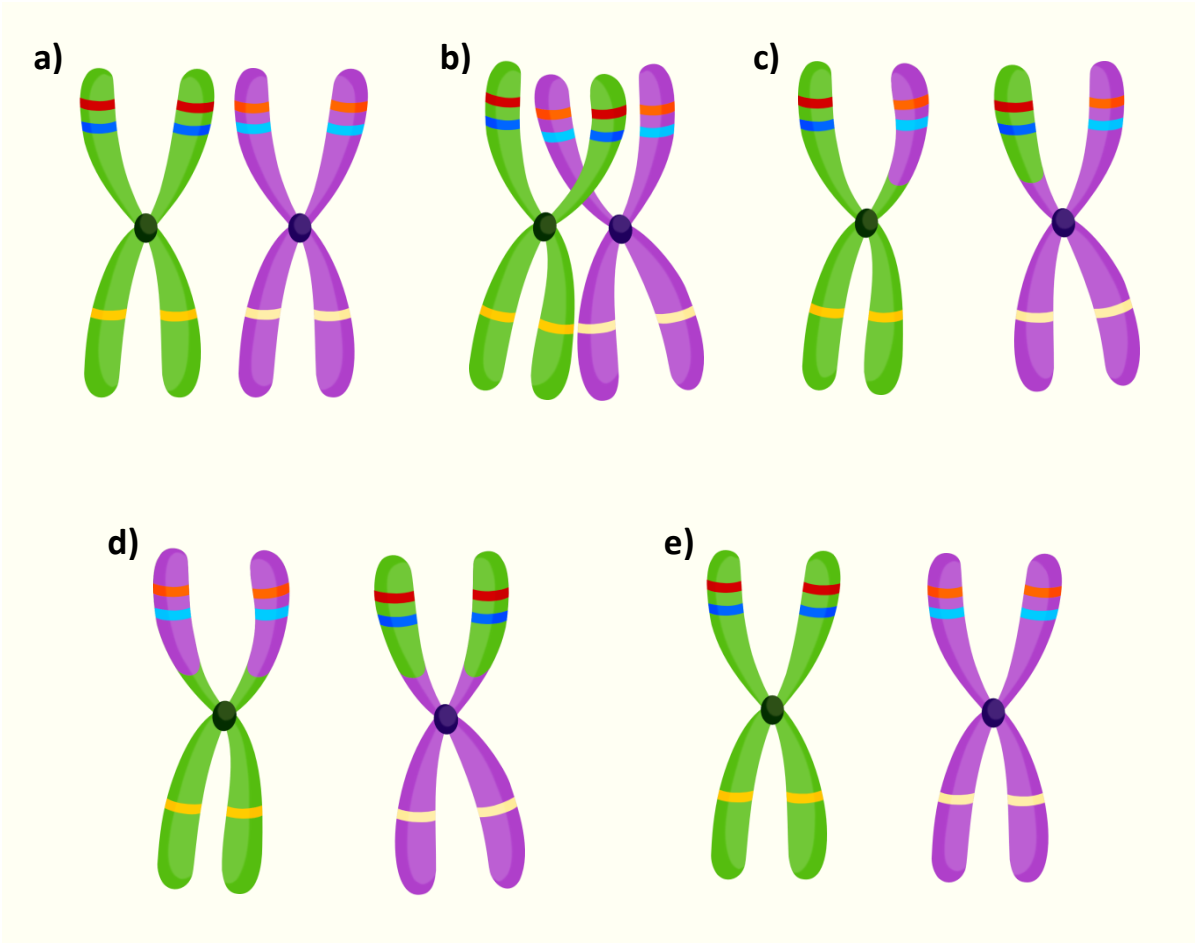


Figure 1.1: A pair of homologous chromosomes during the process of genetic recombination: **a)** Homologous chromosomes (heterozygous at three loci) pair up during prophase I of meiosis ; **b)** Arms of chromosomes (non-sister chromatids) over-lap to form cross overs; **c)** Non-sister chromatids exchange segments of DNA; **d)** The resulting recombinant, and **e)** non-recombinant chromosomes that segredate into gametes.

Therefore, although Mendel was unaware of recombination, in effect, it was assumed that recombination would occur between every gene on every chromosome, *i.e.* a complete "reshuffling" of the genetic material. However, the physical properties of chromosomes are such that only a limited number of cross overs can occur at any one time, making such extensive recombination virtually impossible (Sturtevant 1913; McPeek & Speed 1995; Hassold *et al*. 2000). It is therefore expected that large segments of each recombinant chromosome remain unchanged following a single recombination event, causing any genes, or loci, therein to remain "linked" in the following generation (Figure 1.1d) (Culleton *et al*. 2005; Ben-Ari *et al*. 2006; López *et al*. 2010). As the location of cross overs is to an extent determined randomly, the probability of one forming between any two loci is increased the further they are apart, thus causing linkage to be most prevalent between loci that are located closer together along the chromosome (Morgan 1911). This phenomenon, referred to as genetic linkage, causes the assortment of affected loci to no longer be random, thus violating Mendel's postulate, and altering how linked loci are

inherited relative to unlinked loci. In accordance with the probability of sequential independent events, the gametic frequencies of loci that assort independently should be equivalent to the products of their respective allele frequencies. However, if assortment is no longer random, it is possible to observe certain combinations of alleles more or less often than would be expected under the null model of independent assortment, which is what Bateson *et al*. (1905) observed in their own research.

1.1.2) Allelic associations within populations

Thus far, non-random associations between alleles had only been investigated within the highly simplified context of carefully constructed and isolated pedigrees. However, although the effect of linkage is most easily observed within this arrangement, it can also be observed within natural, outbred populations, which typically consist of numerous multigenerational, interbreeding families, in which multiple different alleles are segregating at each locus. Although the vast number of recombination events occurring within such an environment generally serves to abolish most associations between alleles at loci on the same chromosome, a state referred to as linkage equilibrium, significant associations can still be maintained between very closely linked loci, termed linkage disequilibrium (LD). However, despite its name, genetic linkage only represents one of a number of factors that are capable of causing LD. Although it remains most fundamentally the result of alleles at different loci failing to assort randomly relative to each other, the exact reason for deviation from the null model of independent assortment within the context of a population can vary greatly. As such, LD is more accurately expressed as a probabilistic relationship between the alleles at two or more loci, rather than as the result of any one cause. Population genetics theory commonly defines it as the non-random association of alleles at two or more loci within a population (Slatkin 2008), or as the ability of an allele from one locus to predict the allelic state of another (Meadows *et al*. 2008).

1.1.3) Linkage disequilibrium within the context of micro-evolutionary processes

Although much of the linkage disequilibrium observed within populations can be attributed to the effects of genetic linkage, the level of LD present within a population is, in principle, sensitive to any factor that can influence the independence of alleles at different loci. These include a wide variety of population-specific biological phenomena that are broadly subdivided into locus-specific and demographic factors, depending on their origin and sphere of influence. The effects of locus-specific factors (*i.e.* recombination rate, selection,

mutation, and epistasis) are characteristically limited to a small subset of loci, which are often located on the same chromosome (syntenic loci), but can also be located on different chromosomes (non-syntenic loci) (Slatkin 1994; Frisse *et al*. 2001; Gregersen *et al*. 2006; McVean 2007; Slate & Pemberton 2007; Slatkin 2008; Baird 2015). As recombination is predominantly responsible for preventing and/or eliminating non-random associations between syntenic loci, the overall frequency of recombination events within a particular region, or local recombination rate, represents a primary influencing factor in determining whether LD between these loci exists or not, and if so, how readily it is dissolved in subsequent generations. Rather than being consistent throughout, studies in an increasing number of species (*e.g.* humans, apes, mice, birds, fish, and plants) have demonstrated that recombination rates across the genome are distinctly heterogeneous, with certain areas experiencing significantly higher rates of recombination ("hot-spots"), while others experience significantly lower rates ("cold-spots") (McVean *et al*. 2004; Drouaud *et al*. 2006; Slate & Pemberton 2007; Auton *et al*. 2010; Li & Merilä 2010a; Smagulova *et al*. 2011; Hohenlohe *et al*. 2012; Singhal *et al*. 2015). For example, Backstrom *et al*. (2010) observed significantly higher levels of recombination towards the ends of chromosomes (telomeres) when comparing the relationship between recombination rate and distance to chromosome end in the zebra finch (*Taeniopygia guttata*), chicken (*Gallus gallus*), mouse (*Mus musculus*), and human (*Homo sapiens*). The presence of recombination hot-spots has also been positively correlated with a number of sequence features, *e.g.* GC content (Fullerton *et al*. 2001; Groenen *et al*. 2009; Giraut *et al*. 2011; Auton *et al*. 2013; Singhal *et al*. 2015) and transcription start sites (Pan *et al*. 2011; Choi *et al*. 2013; Singhal *et al*. 2015), as well as the presence of the DNA-binding protein, PRDM9, which binds to specific sequence motifs during meiotic prophase and eventually leads to recombination at those sites (Baudat 2010; Berg *et al*. 2010; Myers *et al*. 2010). As a result, levels of LD can be highly variable across the genome, with non-random associations between alleles within recombination hot-spots tending to be weak and not extend over very many loci, while the opposite is true for recombination cold-spots (Reich *et al*. 2001; Kim *et al*. 2007; Slate & Pemberton 2007; Li & Merilä 2010b). Within this context, LD over short distances therefore tends to decay in a step-wise manner, rather than as a linear function of distance (*i.e.* the strength of associations decreases sharply after a certain distance, rather than a continuous reduction with distance), with 'blocks' of LD in lower recombination regions being separated by recombination hot-spots (Daly *et al*. 2001; Goldstein 2001; Slate & Pemberton 2007).

However, in addition to being a factor of local recombination rates, the distance-dependent co-segregation of loci can also be strongly influenced by locus-specific evolutionary factors, such as selection and mutation. For example, strong positive selection for an advantageous allele during a selective sweep would serve to rapidly increase its frequency within the population. However, because of their proximity, any closely linked loci would tend to co-segregate with the advantageous allele, referred to as genetic "hitch-hiking" (Maynard Smith & Haigh 1974; McVean 2007), resulting in a characteristic block of surrounding LD (syntenic LD). In contrast, selection can also create LD between loci that are located distinctly further away from each other, or even on different chromosomes (non-syntenic LD). In the event that two or more loci are under selective constraint concurrently, the advantageous allelic combination/s would tend to be over-represented within the population, thus creating non-random associations between those alleles (Chan *et al*. 2010a, 2010b; Rhode *et al*. 2013; Stapper *et al*. 2015). In such cases, LD can be maintained between any number of loci, regardless of chromosomal location, as the non-random association causing LD is due to a functional relationship between loci (*e.g.* epistatic interactions), rather than a spatial one. Interestingly, non-syntenic LD has also been observed as being the result of assortative mating. Stapper *et al*. (2015) investigated the possibility of non-random associations between the genes for egg (*EBR1*) and sperm (*Bindin*) recognition proteins within the sea urchin (*Strongylocentrotus purpuratus*), which are known to determine fertilisation success. Although the genes were determined to be located on separate chromosomes, significant LD was observed between them, suggesting that assortative mating preferentially selects for particular combinations of *Bindin* and *EBR1* genotypes, thus creating non-random associations between compatible alleles.

In the case of a neutral mutation, a somewhat similar pattern of syntenic LD to that of selection might be observed, although, such signatures can be distinguished from those surrounding a locus under selection based on their respective allele frequencies within the population (Kimura 1984; Bomba *et al*. 2015). Although the new allele would start out in perfect LD with all other alleles on the chromosome at the time, variants with an elevated level of surrounding LD due to selection would tend to be at a much higher frequency within the population than a neutral variant that had only just arisen due to mutation. Furthermore, as the length of time required for such a variant to reach higher frequencies *via* genetic drift alone would also allow for numerous recombination opportunities between the new variant and surrounding loci, the initially high levels of LD surrounding it would

likely have dissipated by that point. As with genetic distance, there is therefore an inverse relationship between the persistence of syntenic LD and the amount of time that has passed since the LD was created, which serves as a proxy for the number of recombination opportunities between linked loci within that time (Ardlie *et al*. 2002*)*.

In contrast with locus-specific factors, the effects of demographic factors (*i.e.* effective population size, genetic drift, population subdivision, mating system, migration and admixture) on LD is generally observed across the genome, as these processes are not targeted at specific loci (Terwilliger *et al*. 1998; Charlesworth & Wright 2001; Frisse *et al*. 2001; Weiss & Clark 2002; Wakeley & Lessard 2003; Tenesa *et al*. 2007; Uimari & Tapio 2011; Baird 2015). Changes of this nature, often preceded by a sudden increase or decrease in population genetic diversity, are also usually characterised by a sharp overall increase in LD, which then dissipates over time and as a function of local recombination rates. For example, following a significant decrease in effective population size, as might occur during a population bottleneck or founder event, the resulting decrease in haplotype diversity, as well as the sampling effect of random genetic drift, would serve to significantly increase genome-wide levels of LD (Reich *et al*. 2001; Flint-Garcia *et al*. 2003; Gaut & Long 2003; Slatkin 2008; Goddard & Hayes 2009). As a decrease in haplotype diversity would greatly increase the likelihood that any two parents within the population will be homozygous at a given set of loci, such an event would significantly decrease the effective rate of recombination, *i.e.* recombinants are indistinguishable from non-recombinants, thus allowing LD to persist regardless of recombination occurring between linked loci. Therefore, an inverse relationship exists between effective population size and LD. As such, a similar effect on LD can be generated by the utilisation of mating systems that result in a lower effective population size (Weir & Hill 1980; Balloux *et al*. 2003; Flint-Garcia 2003; Gaut & Long 2003). LD within the genomes of out-crossing species (*i.e.* those that reproduce sexually) therefore tends to decay far more rapidly than that within selfing species (*e.g. Arabidopsis thaliana*), or those that reproduce clonally (*e.g. Candida albicans*), as these mating systems are generally associated with reduced effective population sizes (Nordborg 2000; Horn *et al*. 2014; Ozkilinc *et al*. 2015).

A sudden increase in haplotype diversity caused by admixture or migration between previously isolated populations would also result in extremely high initial levels of LD across the genome. Population structure, whether because of geographic isolation or adaptation to differing environments, typically results in the divergence of allele

frequencies between subpopulations, often expressed as an increase in homozygosity within the respective subpopulations [*i.e.* the Wahlund effect (Wahlund 1928)]. In the most extreme cases, certain alleles may be lost or go to fixation in one subpopulation, but not the other. However, in the event that gene flow is re-established between subpopulations *via* migration and interbreeding of individuals (admixture), the amalgamation of divergent haplotypes within the following generation would create a sudden spike in genome-wide LD, as the alleles from each ancestral chromosome would remain in perfect LD with one another until further recombination events are able to break down the ancestral haplotype blocks. Therefore, the elevated levels would subsequently decline over time as they are eroded by recombination (Mueller 2004; Slate & Pemberton 2007).

1.2) Quantifying Linkage Disequilibrium

1.2.1) Primary measures

As a reflection of its complexity, there are currently a number of different methods for quantifying pairwise LD. The original measure of LD between two biallelic loci, which measures the difference between the observed frequencies of recombinant and non-recombinant gametes, is the LD coefficient (or linkage disequilibrium parameter), D_{ij}:

$$D_{ij} = p(A_iB_j)p(A_kB_l) - p(A_iB_l)p(A_kB_j) \quad (1.1)$$

(Lewontin & Kojima 1960), where $p(A_iB_j)$ and (A_kB_l) are the observed frequencies of the non-recombinant gametes, A_iB_j and A_kB_l, respectively, and $p(A_iB_l)$ and $p(A_kB_j)$ are the observed frequencies of the recombinant gametes, A_iB_l and A_kB_j, respectively. Alternatively, this measure can also be expressed in terms of the deviation of observed gamete frequencies from what is expected under independent assortment:

$$D_{ij} = p(A_iB_j) - p(A_i)p(B_j) \quad (1.2)$$

where $p(A_i)$ is the frequency of allele i at locus A, $p(B_j)$ is the frequency of allele j at locus B, and $p(A_iB_j)$ is the frequency of the A_iB_j gamete (Slatkin 2008). However, while this measure is successful in describing the LD between individual pairs of alleles, the LD coefficient was found not to be the most effective statistic to use when comparing the LD between different pairs of loci, as its maximum value is dependent on the allele frequencies within the population, making comparisons between different pairs of loci meaningless (Harmegnies *et al.* 2006; Slatkin 2008). Similarly, comparing the same pair of

loci over different populations would also be problematic, as allele frequencies can be markedly different between populations, due to, for example, adaptive selection to differing environments or the effects of genetic drift (Ardlie *et al*. 2002). To address this, a standardised measure of D_{ij}, D'_{ij}, was developed by interpreting D_{ij} relative to its maximum theoretical value given the allele frequencies:

$$D'_{ij} = \frac{D_{ij}}{D_{ij}^{max}} \tag{1.3}$$

(Lewontin 1964), where D_{ij}^{max} is the smaller of $p(A_i)(1-(pB_j))$ and $(1-p(A_i))p(B_j)$, and the value of D'_{ij} ranges from 0 to 1. However, even this statistic proved problematic; D'_{ij} is sensitive to both allele frequency and sample size, where rare alleles and small sample sizes tend to result in inflated values of D'_{ij} (Slate & Pemberton 2007; Meadows *et al*. 2008). As an alternative measure, Hill and Robertson (1968) suggested employing the squared correlation coefficient, r_{ij}^2:

$$r_{ij}^2 = \frac{D_{ij}^2}{p(A_i)(1-p(A_i))p(B_j)(1-p(B_j))} \tag{1.4}$$

which quantifies the information one locus provides about another, and is less sensitive to small sample sizes and rare alleles than D'_{ij} (Morton *et al*. 2001; Zhao *et al*. 2005), although not unaffected. As such, r_{ij}^2 is currently the preferred measure for investigating LD between biallelic markers (Zhao *et al*. 2005).

However, as these initial measures were only suitable for comparing pairs of loci with a maximum of two alleles each (biallelic loci), they could not be used for calculating LD between pairs of markers with more than two alleles (multi-allelic loci), as LD can differ between individual pairs of alleles at the same two loci. As such, a combined measure of LD across all alleles at each pair of loci was required. Hedrick (1987) sought to address the issue by formulating a multi-allelic extension of Lewontin's D'_{ij}, termed D':

$$D' = \sum_{i=1}^{k}\sum_{j=1}^{m} p(A_i)p(B_j)\left|D'_{ij}\right| \tag{1.5}$$

(Hedrick 1987), where k is the number of alleles at locus A, m is the number of alleles at locus B, and $\left|D'_{ij}\right|$ is the absolute value of D'_{ij} for each pairwise comparison. However,

while this measure has been widely accepted within the field of population genetics, numerous studies have reported that D', as with D'_{ij}, is readily inflated in the presence of rare alleles, or when working with smaller sample sizes (Ardlie *et al*. 2002; McRae *et al*. 2002; Flint-Garcia *et al*. 2003; Pe'er *et al*. 2006). Because the likelihood of encountering all possible allelic combinations is decreased within smaller samples, particularly when one or more of the alleles is uncommon within the population, D' may indicate high levels of LD between loci even when they are not non-randomly associated, which calls into question its ultimate utility in providing an accurate estimate of LD (Heifetz *et al*. 2005).

In contrast with those described above, an alternative multi-allelic measure of LD is one of a number that are based on the chi-square statistic:

$$\chi^2 = 2N \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{D_{ij}^2}{p(A_i)p(B_j)} \tag{1.6}$$

where N is the sample size and $2N$ is the number of gametes within the sample. This statistic tests for independence between alleles at two loci, and was put forth by both Hill (1975) and Hedrick (1987) as a potential LD measure. The metric presented here, $\chi^{2\prime}$, is a standardisation of the χ^2 statistic:

$$\chi^{2\prime} = \frac{\chi^2}{2N(l-1)} \tag{1.7}$$

(Yamazaki 1977), where l is the number of alleles (k or m) at the locus with the smallest number of alleles. This statistic, which is equivalent to the square of Cramér's *V* (W_n; Cramér 1946), is generally regarded as the multi-allelic extension of r_{ij}^2, and is normalised to lie between zero and one (Thomson & Single 2014). As with D'_{ij}, the denominator, $2N(l-1)$, provides a maximum value for χ^2 given the allele frequencies, by which it is standardised. While this value is actually considered to be a significant over-estimate of the maximum value of $\chi^{2\prime}$ (Kalantari *et al*. 1993), an attempt by Zhao *et al*. (2005) to provide a sharper upper bound for $\chi^{2\prime}$ found that their revised estimate, $\chi_{tr}^{2\prime}$, was actually a poorer predictor of useable marker-QTL LD than the original $\chi^{2\prime}$, which they surmised was due to the imperfect dependence of QTL alleles on marker alleles.

Despite several usable multi-allelic LD measures having been proposed, the most prominent of which are described above, an overall satisfactory measure has not yet been

settled upon (Slatkin 2008). Zhao *et al*. (2005) compared the efficacy of a number of these measures at estimating LD, and concluded that $\chi^{2\prime}$ was the most accurate measure of LD for multi-allelic markers, regardless of population size and number of alleles, and despite concerns that it underestimates LD. However, irrespective of its widely acknowledged drawbacks, D' remains one of the most widely used measures of LD (Ardlie *et al*. 2002; Flint-Garcia *et al*. 2003; Zhao *et al*. 2005). This is most likely due to the desire to retain comparability between the large number of past studies that used D' and more contemporary studies. As such, D' will most likely remain a relevant measure of LD, perhaps indefinitely, to be used in conjunction with more reliable estimates, such as $\chi^{2\prime}$ (Meadows *et al*. 2008).

1.2.2) Related measures

In general, the primary measures of pairwise LD are only able to report on the magnitude of LD observed between two or more loci. In isolation, these parameter estimates provide little indication of the persistence of LD over genetic distance (*i.e.* the relationship between LD and recombination), nor are they informative concerning the reason for the observed LD, or overall patterns of LD across multiple loci (Mueller 2004). The inclusion of related measures in LD analyses, defined here as those that make use of the basic LD measures to further investigate these additional properties, can therefore provide a much more informative picture of both the importance and the nature of observed LD.

One such measure is the rate of LD decay. As has been discussed, LD between loci on the same chromosome is primarily a function of genetic distance. This is because the probability of recombination occurring between adjacent loci increases with increasing distance, which therefore has the opposite effect on the likelihood of these loci co-segregating. However, such a relationship cannot persist indefinitely, and it is expected that after a certain critical distance, LD will reach an equilibrium point (often not zero) and stop decaying as a function of distance, *i.e.* when loci are so far apart that they assort independently as if on different chromosomes (Corbin *et al*. 2010). This critical distance is largely determined by the point at which LD ceases to be significant, referred to as baseline or background LD (Corbin *et al*. 2010; Maccaferri *et al*. 2010). As LD present between non-syntenic markers is understood to be a product of either chance alone or factors other than genetic distance (*e.g.* population substructure and admixture), assuming they are functionally independent, an effective proxy for baseline levels of LD is the level

observed between non-syntenic markers. A commonly used method for estimating baseline LD is therefore to calculate the average amount of LD present between a subset of functionally independent non-syntenic markers (Heifetz *et al*. 2005; Moen *et al*. 2008; Corbin *et al*. 2010) (Figure 1.2). If syntenic LD is then plotted against genetic distance, the distance value that corresponds with the baseline level of LD then represents the critical distance, or the point at which the LD observed is no longer significant. As this value effectively characterises the rate of LD decay, or how far LD extends on average, over a particular region, this value is of great importance in a number of applications of LD data.

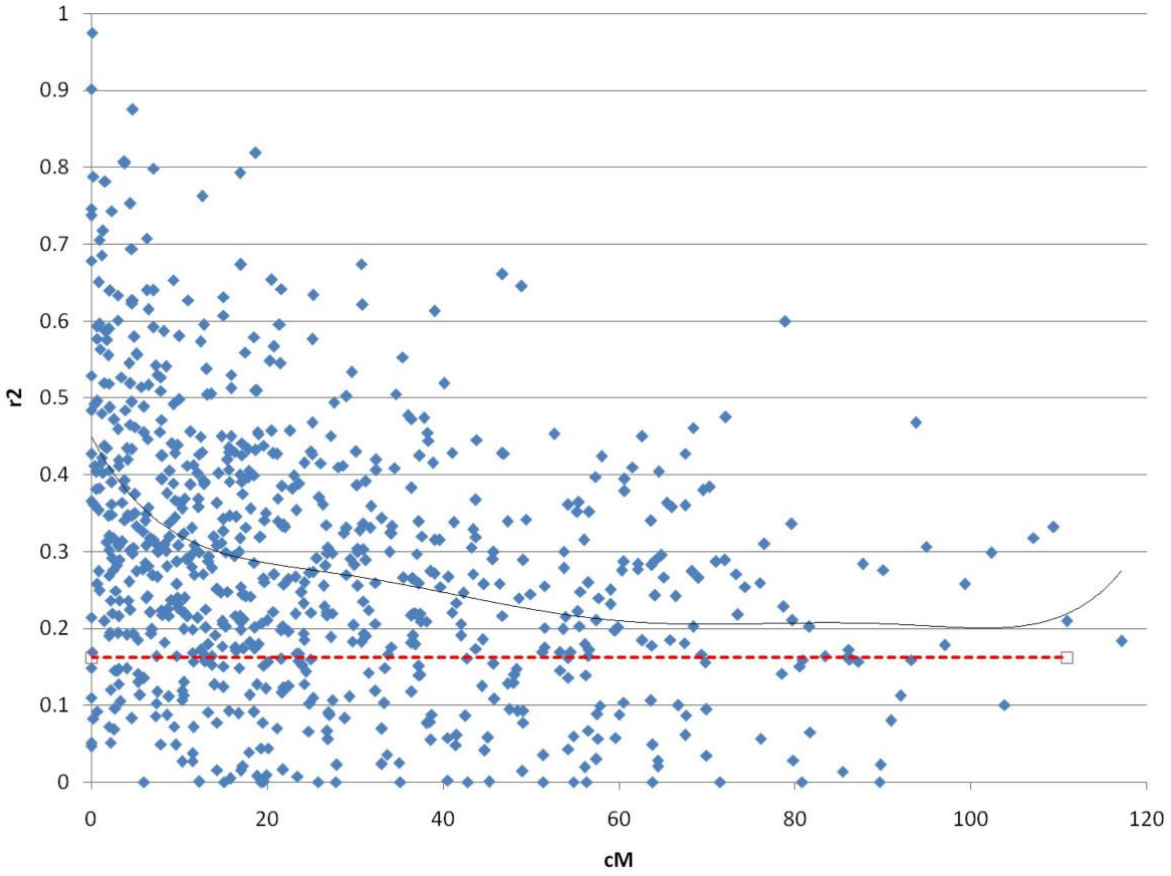


Figure 1.2: An example of an LD decay plot. Levels of LD (r^2) between syntenic markers are plotted against genetic distance (cM). The solid line represents a 6th degree polynomial trendline for best fit to the data, while the broken red line is the average level of LD between non-syntenic markers (0.16). Figure taken from Moen *et al*. 2008.

An additional means of characterising the rate at which LD decays with distance is by determining the coefficient of decay, b_j (Heifetz *et al*. 2005; Meadows *et al*. 2008; Rexroad *et al*. 2009). This parameter, which increases as LD decays more quickly, can be estimated by fitting the observed data to a re-expression of the model developed by Sved (1971):

$$LD_{ij} = (1 + 4b_j d_{ij})^{-1} + e_{ij} \quad (1.8)$$

(Heifetz *et al*. 2005), where LD_{ij} is the observed LD between marker pair i of population j, separated by distance d_{ij} (in cM), b_j describes the decline of LD with genetic distance within population j, and e_{ij} is the model redidual. Therefore, rather than describing the theoretical relationship between LD and N_e, this equation then describes the extent and decline of LD with genetic distance.

A second related measure of LD that can be employed when data for more than one population are available is the partitioning of LD into contributions within and between populations (Ohta 1982; Slatkin 2008). Similar to the manner in which Wright's F-statistics (*i.e.* F_{IS}, F_{ST}, F_{IT}) partition genetic variation based on the deviation from Hardy-Weinberg expected frequencies, the partitioning of LD involves the segregation of the total LD over a subdivided population, D_{IT}^2, into the average LD within each subpopulation, D_{IS}^2, and the LD due to divergent allele frequencies between subpopulations, D_{ST}^2. However, the use and interpretation of these measures are not entirely analogous to those of F-statistics. Unique to D-statistics is the ability to distinguish between the two main causes of LD: epistatic natural selection and random genetic drift. For this purpose, a second set of values is also calculated for LD within and between subpopulations, namely, D'_{IS}^2 and D'_{ST}^2, respectively. By interpreting the ratios of D_{ST}^2 / D_{IS}^2 and D'_{IS}^2 / D'_{ST}^2 based on Ohta's model of LD in finite populations at equilibrium, it is possible to determine which of the two factors is primarily responsible for the observed LD between each pair of loci (Whittam *et al*. 1983; Barton & Clark 1990; Volis *et al*. 2003; Yu *et al*. 2003; Matala *et al*. 2004). For example, using these ratios, Yu *et al*. (2003) determined that selection for adaptation to differing environments, rather than genetic drift, was predominantly responsible for maintaining non-random associations between multiple loci within a biologically and geographically diverse study population of rice (*Oryza sativa*). This conclusion was further supported by the region-specific patterns of genomic diversity observed.

The final related measure to be discussed here is the characterisation of multi-locus LD. This measure is based on the principle that when more than two loci are considered at a time, it is possible to observe sets of loci that are in LD with one other to an extent that is not fully accounted for by the consideration of only their pairwise comparisons (Slatkin 2008). Such sets of loci, referred to as haplotype blocks, have been observed over a wide

range of species (Slatkin 2008), and there are now a number of different methods for investigating the phenomenon (Geiringer 1944; Thomson & Baur 1984; Hayes *et al*. 2003; Albrechtsen *et al*. 2007; Kim *et al*. 2008), two of which will be introduced here. The first method is simply a modification of the original LD coefficient equation to describe higher-order disequilibria, here for three loci:

$$D_{ABC} = p_{ABC} - p_A D_{BC} - p_B D_{AC} - p_C D_{AB} - p_A p_B p_C \quad (1.9)$$

(Geiringer 1944; Thomson & Baur 1984; Slatkin 2008), where D_{ABC} is the three-way interaction term quantifying the level of association that is not accounted for by the pairwise coefficients, D_{AB}, D_{BC}, and D_{AC}.

The second method uses a somewhat different approach, although still making use of the difference between an observed value and that expected under linkage equilibrium. The index of association, I_A, utilises the variances in number of alleles that differ between haplotypes in pairwise comparisons, normalised by the expected value:

$$I_A = \frac{(V_O - V_E)}{V_E} \quad (1.10)$$

(Brown *et al*. 1980; Mueller 2004), where V_O is the observed variance of pairwise distances, and V_E is the expected variance under linkage equilibrium. This measure therefore tests the extent to which two haplotypes, being the same at one locus, are more likely than random to also be the same at additional loci.

1.3) Applications of Linkage Disequilibrium

1.3.1) Understanding population genetic dynamics and micro-evolution

Within the field of population genetics, linkage disequilibrium data represents a highly versatile and informative statistical resource. Its usefulness in this regard stems primarily from its sensitivity to a wide variety of both locus-specific and demographic factors, as well as the well-defined and predictable manner in which it responds to these factors. Equipped with an understanding of the manner in which LD affects and is affected by these factors, it is therefore possible to draw conclusions concerning the demographic and evolutionary histories of target populations by examining the magnitude, extent and patterns of LD therein (Slatkin 2008). Fortunately, the population genetic theory of LD is well developed

and much is already known about how LD behaves under the influence of various micro-evolutionary processes, as previously discussed.

Based on the inverse relationship between effective population size and the overall level of LD within a population, one common application of LD data is to estimate effective population size. As significant changes in effective population size are often associated with major demographic events, such as population bottlenecks or admixture events, estimates of effective population size are widely regarded as some of the most important parameters in both evolutionary (Charlesworth 2009) and conservation biology (Luikart *et al*. 2010; Robinson & Moyer 2013). The relationship between LD and effective population size was initially characterised by Sved (1971), who formulated the expectation of LD (r^2) between a given pair of loci as a function of effective population size (N_e) and the recombination rate (c) (commonly replaced by recombination distance in Morgans):

$$E(r^2) = (1 + 4N_e c)^{-1} \quad (1.11)$$

Using this relationship, Hill (1981) then derived an equation for inferring N_e from the level of LD. While this method was initially dismissed by many because of a severe bias introduced when the sample size was less than the true effective size (England *et al*. 2006), this and other issues have largely been resolved through extensive optimisation over the last decade (Waples 2006; Waples & Do 2008, 2010; Waples & England 2011; Peel *et al*. 2013). In addition to estimating contemporary N_e (*i.e.* within the time period encompassed by the sampling effort), the LD method can also be used to estimate more historical N_e. While, LD between closely linked loci would tend to reflect more ancient population history, longer range LD would tend to reflect more recent events, as the closer two loci are, the longer the time required for the LD to be broken down (Hill 1981). As such, changes in N_e over time can be investigated by examining the level of LD across different genetic distances, with the LD over a specific distance, c (in Morgans), reflecting the ancestral N_e $(2c)^{-1}$ generations ago (Hayes *et al*. 2003). Estimates of historical N_e can therefore be of particular interest when investigating the demographic history of populations. For example, in a study on linkage disequilibrium in a large, mildly selected cattle population from Western Africa (*Bos indicus* x *Bos taurus*), Thévenon *et al*. (2007) reported a decreasing trend in historical N_e, despite relatively large estimates for contemporary N_e as compared with other livestock populations (Farnir *et al*. 2000; McRae *et al*. 2002; Tenesa *et al*. 2003a; Nsengimana *et al*. 2004; Harmegnies *et al*. 2006). In

explanation, the authors suggested either continuous admixture events, or a selective sweep for resistance to the protozoan parasite, *Trypanosoma brucei*, as possible causative factors, as both alternatives could result in an increase in LD, which would deflate estimates of N_e based on the LD method. However, as estimates of genetic diversity (heterozygosity and mean number of alleles) remained high within the study population, and historical and continuous hybridisation events between *Bos indicus* and *Bos taurus* have been widely reported (MacHugh *et al*. 1997; Hanotte *et al*. 2002; Freeman *et al*. 2004), it was concluded that admixture, rather than selective pressure, was the predominant causative factor in decreasing N_e. In contrast, estimates of contemporary N_e can be highly instrumental in predicting the extinction risk of potentially endangered populations (Luikart *et al*. 2010). As an example, Saura *et al*. (2014) recently assessed the conservation status of a closed population of Iberian pigs (Guadyerbas strain) by estimating contemporary N_e from LD data. The herd was established from only 24 individuals (4 males and 20 females) in 1944, and is now believed to be in danger of extinction, a hypothesis which was confirmed by the critically low estimate of current N_e, 36. Alternatively, estimates of N_e can also find application in the evaluation and optimisation of selective breeding strategies in domesticated species (Corbin *et al*. 2010; Daetwyler *et al*. 2010; Qanbari *et al*. 2010), particularly in the event of incomplete or unavailable pedigree data. For example, Corbin *et al*. (2010) evaluated the extent and decay of LD within a sample population of 817 Thoroughbred horses to determine the feasibility of the proposed genomic selection strategy using the marker set available. As part of their assessment, the authors estimated the current N_e of the study population (~180), which they then used to determine the potential accuracy of selection according to the equations derived by Daetwyler *et al*. (2010), concluding that the available marker panel was sufficient for effective genomic selection within Thoroughbred populations.

In addition to estimating effective population size, LD data can also be used to estimate the age of proposed demographic events, such as population divergence, admixture or a population bottleneck (Risch *et al*. 1995; Stephens *et al*. 1998; Abecasis *et al*. 2001; Sankararaman *et al*. 2012). As LD decays over time as more opportunities for recombination arise, the magnitude and extent of LD still present within the population can be used to determine how long ago an event that created LD occurred (Ardlie *et al*. 2002). For example, Sankararaman *et al*. (2012) used this approach to distinguish between the two opposing hypotheses for explaining why non-Africans share more genetic variants with Neandertals than Africans do (Green *et al*. 2010). The first suggests that ancient

substructure within the common ancestral population of humans and Neandertals is responsible for the discrepancy, while the second suggests that a more recent history of gene flow between Neandertals and modern humans after they left Africa is responsible. As shorter range LD generally indicates that more time has passed since the LD was created, and *vice versa*, the authors used the extent of LD still present within shared regions to determine how recently the last gene flow between humans and Neandertals occurred, finding convincing evidence to support the recent admixture hypothesis in the form of extensive long range LD. Similarly, local patterns of LD can also be used to date the appearance of a specific allele within the population. In conjunction with coalescence theory (Hudson & Kaplan 1986; Hudson 1990), which models the genealogy of a particular haplotype based on inter-haplotype variation and local recombination rates (Nordborg 2002), Stephens *et al*. (1998) used the extent of LD surrounding a deletion mutation within the macrophage chemokine receptor gene, *CCR5*, to determine its likely age. The *CCR5-Δ32* allele, which they estimated to be approximately 700 years old, is known to occur exclusively within Caucasian populations, where it is highly prevalent, and has been positively associated with heightened resistance to HIV-1 infection. As such, the authors speculate that the *CCR5-Δ32* allele may also infer resistance to additional pathogens that utilise a similar infection mechanism to HIV-1, and that the rapid increase in frequency was caused by a selective sweep for resistance to another widespread epidemic occurring at the time, such as the bubonic plague (650 years ago).

With regards to the effects of selection on LD, there is now increasing interest in the detection of genomic regions containing genes that are, or have been, under selective constraint. In particular, such information can provide valuable insights into the genetic mechanisms responsible for the divergence of phenotypes between populations, the elucidation of which represents a foremost objective within evolutionary biology (Feder & Mitchell-Olds 2003; Storz 2005; Gholami *et al*. 2015). As selection events tend to create distinctive 'signatures' of LD surrounding loci under selection, which differ based on the type and age of the selection event, exploiting the effects of selection on the level of LD across shorter distances represents an ideal strategy for identifying these candidate regions under selection (Storz 2005; Kuhn *et al*. 2014; Qanbari & Simianer 2014; Gholami *et al*. 2015; Wollstein & Stephan 2015). As such, a wide variety of statistics for detecting these signatures has been developed (reviewed by Qanbari & Simianer 2014). For example, the extended haplotype homozygosity (EHH) method, developed by Sabeti *et al*. (2002), takes advantage of the genetic hitch-hiking phenomenon by looking for extended

regions of homozygosity, called core haplotypes. Using neutral theory as the null model, which states that all changes in allele frequency are driven by genetic drift alone, selectively neutral mutations will require many generations in order to reach a high frequency within the population, thus allowing the initially high levels of surrounding LD to decay significantly in that time (Kimura 1984). However, if the new mutation is beneficial, positive selection will cause a rapid increase in frequency, resulting in the persistence of strong, long-range LD surrounding the new variant. Therefore, while extended regions of homozygosity might also be observed surrounding a newly formed neutral mutation, candidate regions under selection can be distinguished by searching preferentially for core haplotypes that are already at high frequencies within the population. Using this approach, Bomba *et al*. (2015) was able to identify a number of candidate regions for recent directional selection within a multi-breed population cohort comprised of dairy, beef and dual purpose cattle breeds. In total, 82 and 87 candidate regions were identified within dairy and beef breeds, respectively, from which 244 and 232 genes could be identified using bioinformatics analysis, many of which are associated with milk or meat production.

However, as mentioned previously, selection can also create non-random associations between loci that are distinctly further away from each other, or even non-syntenic. Although it is possible for such associations to be caused by chance alone, or as a result of demographic events, such as admixture (Farnir *et al*. 2000; Williams *et al*. 2001; Tenesa *et al*. 2003b), non-syntenic LD may also represent selection for a particular multi-locus allelic combination within a gene network (Tong *et al*. 2004; Chan *et al*. 2010a, 2010b; Rhode *et al*. 2013). While simple traits that are primarily determined by a single gene with a large effect on phenotypic outcome are relatively prevalent [*e.g.* monogenic disorders such as phenylketonuria, cystic fibrosis and Huntington's disease (Chial 2008)], the genetic architectures of most traits are far more complex (Templeton 2000; Mackay 2001). In accordance with the complexity of most biological systems, the vast majority of traits tend to be associated with a variety of biological processes, represented by a multitude of genes, each with varying degrees of effect (Mackay 2001; Goddard & Hayes 2009). These include many traits that are important in agriculture (*e.g.* growth rate or yield traits), medicine (*e.g.* disease susceptibility), and evolution (*e.g.* clutch size of birds) (Goddard & Hayes 2009). As such, the effects of selection operating on such traits could potentially extend over many or all of the genes involved, creating non-random associations between advantageous or functionally linked alleles, irrespective of their relative genomic locations (Chan *et al*. 2010a, 2010b). Within this context, non-syntenic LD could therefore potentially

represent functional relationships between genes, and thus aid in the elucidation of the genetic architecture of complex traits of interest. A prominent example of this is the interpretation of non-syntenic LD between the *MAM* and *AOP* genes of the model glucosinolate (GSL) pathway in *Arabidopsis thaliana* by Chan *et al*. (2010a). The study aimed to further investigate the genomic architecture of plant-insect interaction traits by performing a genome-wide association study for GSL phenotype in 96 natural *A. thaliana* accessions. Both the *MAM* and *AOP* genes are known to represent important regulatory elements within the GSL pathway, and interactions between them are well characterised (reviewed by Kliebenstein (2009)). As such, the non-random association between the two genes was interpreted as evidence of their functional relationship, rather than as a false positive association caused by population structure, and may suggest that the associated biochemical phenotypes are under selective pressure.

1.3.2) Identifying genotype-phenotype associations

In addition to furthering our understanding of the processes that shape diversity within biological populations, another key endeavour of genetics is the identification and characterisation of the genetic elements that make up traits of interest. However, this is not an easy goal to achieve in most cases, as the vast majority are complex traits, which greatly complicates the process of deciphering their genetic architectures (Goddard & Hayes 2009; Yue 2014). In addition to this, their expression is also affected by environmental conditions and epistatic interactions between genes, all of which combine to produce a continuous distribution of phenotypes (Mackay 2001). As such, they are also referred to as quantitative traits, with the genes or genomic regions that determine them being referred to as quantitative trait loci (QTLs). The task of locating the most relevant genes for a given trait, referred to as QTL mapping, is therefore complicated. A variety of mapping strategies exist, all of which begin with the construction of a genetic linkage map indicating the relative chromosomal locations of all mapping markers. These maps act as a framework for the identification and mapping of QTLs, and are constructed by tracking the inheritance of polymorphic markers within a set of families, or mapping populations. Markers that consistently co-segregate are grouped together within linkage groups, which represent (putative) chromosomes. Because the probability of recombination occurring between adjacent loci is a function of the distance between them, the proportion of recombinants observed between each pair of syntenic markers is used as an indication of the relative distances between them (measured in centiMorgans; cM) (Korol *et al*. 2007;

Roodt-Wilding & Brink 2011). These mapped markers can then be used to narrow down, or map, the locations of QTLs of interest by looking for evidence of marker-QTL co-segregation in the form of marker-phenotype associations, for which a number of strategies are available (Stinchcombe & Hoekstra 2007; Goddard & Hayes 2009; Yue 2014).

In addition to its utility in investigating population genetic processes, LD data has also become increasingly popular as a means of mapping QTLs (Stinchcombe & Hoekstra 2007; Meadows *et al*. 2008; Goddard & Hayes 2009). The technique, known as LD- or association mapping, is based on the likelihood that one or more of the study markers will be located close enough to a causal variant to be in LD with it, which can then be detected in an appropriately designed genome-wide association study, or GWAS (Stinchcombe & Hoekstra 2007). As such, a typical GWAS requires a densely saturated set of molecular markers spanning the entire genome, so as to ensure a reasonable chance of detecting marker-QTL co-segregation. However, while this approach has the advantage that no *a priori* knowledge of the trait's genetic architecture is necessary (Goddard & Hayes 2009), the large number of markers required to map QTLs accurately has been a prohibitive factor, particularly within species for which relatively few markers have been developed. The problem was circumvented by instead using linkage data from carefully constructed pedigrees, which made use of the limited recombination occurring within a family to keep blocks of linkage large, and therefore minimise the number of markers required (Meadows *et al*. 2008). However, this mapping strategy brought with it a number of distinct disadvantages that made linkage mapping a somewhat less desirable method for locating QTLs. Firstly, the construction of large, multigenerational mapping families was often an impractical task, particularly when dealing with natural populations or species with long generation times (Stinchcombe & Hoekstra 2007). Secondly, although larger blocks of LD meant fewer markers were required, it also increased the confidence intervals for mapped QTLs. A typical genome scan maps QTLs to a sizeable region of 20 cM or more, within which hundreds of genes might be located, thus making the definitive identification of causal variants problematic (McRae *et al*. 2002; Goddard & Hayes 2009; Yue 2014). Lastly, because linkage analysis only follows the segregation of marker alleles within a single lineage at a time, the possibility is always present that any QTLs identified could be specific to a particular lineage, and therefore have limited applicability to the greater population (Risch 2000; Massault *et al*. 2008; Goddard & Hayes 2009; Yue 2014). In contrast, LD mapping makes use of population data, which does not require controlled

breeding experiments, is able to map QTLs more precisely, and greatly increases the chance that any QTLs identified will be applicable to the broader population, and not just the familial cohort in which they were identified (Stinchcombe & Hoekstra 2007; Goddard & Hayes 2009). As such, following the recent advancement in molecular techniques and genomic tools that facilitated the development of large numbers of markers for both model and non-model species, LD mapping has re-emerged as the method of choice for locating QTLs in both wild and domesticated populations (Slate & Pemberton 2007; Stinchcombe & Hoekstra 2007; Goddard & Hayes 2009; Li & Merilä 2010a).

However, in addition to scanning the entire genome in search of marker-phenotype associations, LD mapping can also be applied on a smaller scale by focusing on one or more previously identified candidate regions/loci/genes, obtained *via* various means (Long & Langley 1999; Mackay 2001; Stinchcombe & Hoekstra 2007; Poelstra *et al*. 2013). For example, in cases where linkage data is readily available, a combined linkage-LD mapping approach has been used, where QTLs are first mapped broadly using pedigree analyses, and then "fine-mapped" using LD mapping targeted at the identified candidate regions (Mackay 2001; Meuwissen *et al*. 2002; Stinchcombe & Hoekstra 2007; Goddard & Hayes 2009). This approach has been particularly successful in locating QTLs for economically important traits within a number of species (*e.g.* sheep, dairy cattle, pigs and salmon), where target QTLs could be mapped to between 0.04 and 10 cM of their true locations (Meuwissen *et al*. 2002; Meuwissen *et al*. 2004; Olsen *et al*. 2004; Uleberg *et al*. 2005; Hayes *et al*. 2006; Olsen *et al*. 2008; Marshall *et al*. 2009). Alternatively, it is also possible to identify candidate loci based on distinctive patterns of LD. Assuming that some form of selection for the trait of interest is already underway, promising candidate loci can then be singled out by looking for the expected patterns of LD that are left behind as a result of being under selective constraint, as discussed in the previous subsection (Storz 2005; Kuhn *et al*. 2014; Qanbari & Simianer 2014; Gholami *et al*. 2015; Wollstein & Stephan 2015). In the case of domesticated species, the effectiveness of this approach can be improved further by comparing data from wild progenitor populations with those from commercially bred ones, as the more recent, *i.e.* post-domestication, selection signals will not be present within the wild populations (Rubin *et al*. 2010; Li *et al*. 2013; Rhode *et al*. 2013; Carneiro *et al*. 2015; Evin *et al*. 2015). Alternatively, relatively recent selection events can also be distinguished from more historical ones based on the extent of LD still surrounding the site, as 'older' sites will tend to have less extensive surrounding LD. The process of finding QTLs can thus be streamlined further by filtering out loci that are less

likely to be QTLs for the trait/s under selection. These candidate genes/variants can then be further investigated using bioinformatics analyses to identify the most promising QTL candidates for association studies.

However, the ultimate application of confirmed QTLs, particularly within species of economic importance, is to not only study them, but also improve them (Dekkers & Hospital 2002; Dekkers 2004; Stinchcombe & Hoekstra 2007; Goddard & Hayes 2009; Yue 2014). While conventional phenotypic selection, which is already in place for the majority of domesticated species, remains a relatively effective strategy for the genetic improvement of animal populations, this approach poses a number of distinct disadvantages. In particular, phenotypic selection, which is based on the assessment of an individual's own performance, is less effective when considering traits that are difficult to measure, such as those where the individual would need to be sacrificed, *e.g.* carcass traits (Hayes *et al*. 2007; Yue 2014). Moderate to low heritability traits are also problematic for this type of selection, as individuals with a superior phenotype are not guaranteed to have a superior genotype (Knapp 1998). Furthermore, many traits are only measurable after an animal has reached a certain age, *e.g.* longevity or reproduction traits (Dekkers & Hospital 2002; Goddard & Hayes 2009), which greatly increases the time and resources required to determine an individual's breeding value. In contrast, the use of genetic enhancement strategies, such as gene-assisted selection (GAS) and marker-assisted selection (MAS), would allow the assessment of an individual's performance potential at the earliest opportunity and with little to no harm to the animal itself (Sonesson & Meuwissen 2009; Yue 2014). These techniques make use of molecular markers (direct or indirect, respectively) to identify superior individuals for a given trait based on their genetic makeup at those markers. The development of performance evaluation markers has therefore become a primary objective in the effort to improve commercially exploited populations, of which the mapping of QTLs and identification of causal variants forms an integral part (Dekkers & Hospital 2002; Dekkers 2004; Liu & Cordes 2004; Nieuwhof *et al*. 2008).

1.4) Abalone: Biology and Commercial Importance

1.4.1) Overview of biology, ecology and evolution

Abalones are a group of marine gastropod molluscs that are prevalent along most rocky shores within temperate to tropical waters. At present, there are 56 recognised species

and 18 subspecies, all of which fall within a single genus, *Haliotis*, the only member of the family, *Haliotidae* (Geiger & Owen 2012). Unlike most gastropods, abalones possess a relatively flat shell, although it is still distinctly spiralled (Figure 1.3). The dorsal surface of the shell is often dull and highly textured with bumps and ridges (Figure 1.3a), while the smooth ventral surface is coated with a thick layer of iridescent nacre (mother of pearl) (Figure 1.3c). The overall shape varies between species, from almost entirely round to highly elongated, as does the colour, which can be affected by diet. Several open respiratory holes, which are characteristic of the genus, line the edge of the shell mantel and allow for ventilation of the gills and the release of gametes during spawning. The body is dominated by the large fleshy foot, the dorsal portion of which, the epipodium or mantel, is frilled and protrudes out below the edge of the shell, largely covering the lower portion of the foot, which the animal uses to attach itself to rocks or other hard substrates (Figure 1.4a – c). The sole of the foot is a simple, flat structure that is usually a creamy white, but can be pigmented. The head is comprised of short, moveable eye stalks, cephalic tentacles, and a semi-mobile snout (Figure 1.4d – f) (Geiger & Owen 2012).

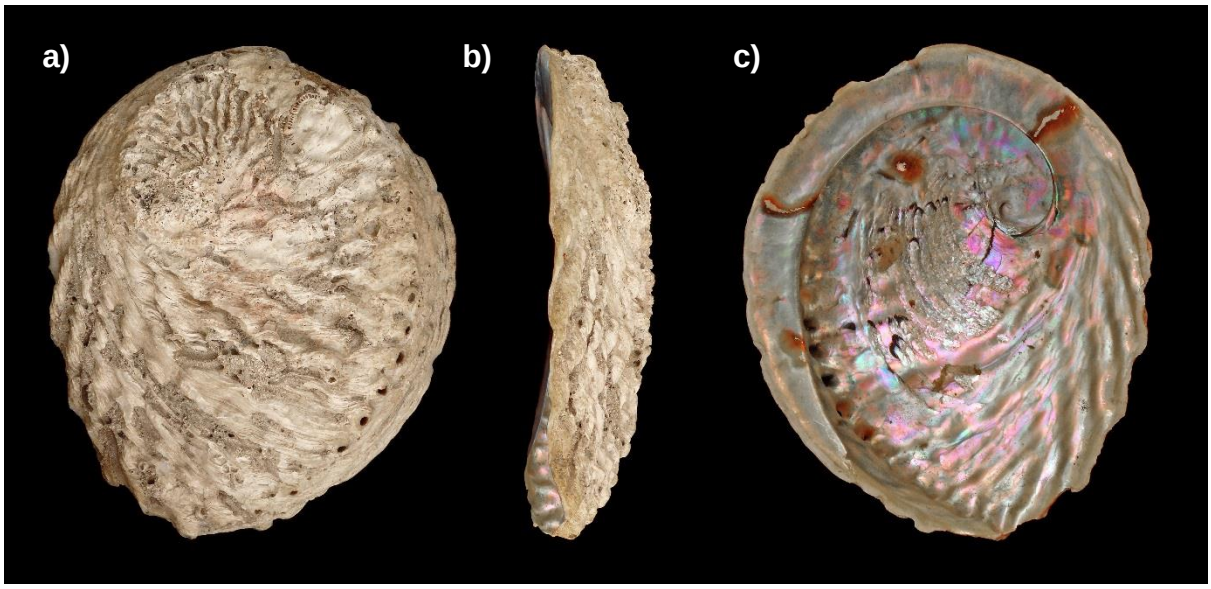


Figure 1.3: Images of the **a)** dorsal, **b)** lateral (right side), and **c)** ventral surfaces of an abalone shell (*Haliotis midae*). Photograph by H. Zell, distributed under a CC BY-SA 3.0 license.

Abalones are dioecious broadcast spawners and generally mature within three to seven years (Tarr 1995; Geiger & Owen 2012). Spawning is largely dependent on seasonal fluctuations in a variety of environmental factors, *e.g.* water temperature and photoperiod, as well as the presence of conspecific gametes, during which millions of eggs and sperm are released into open water for external fertilisation (Geiger & Owen 2012). For the first one to two weeks, offspring are carried as lecithotrophic (non-feeding) larvae by ocean

currents, potentially travelling hundreds of kilometres before being prompted to settle by chemicals associated with encrusting coralline algae (Wood & Buxton 1996; Roberts 2001). In general, larvae settle preferentially on encrusting corallines, where they are more or less invisible, and feed primarily on diatoms and bacteria (Day & Branch 2000).

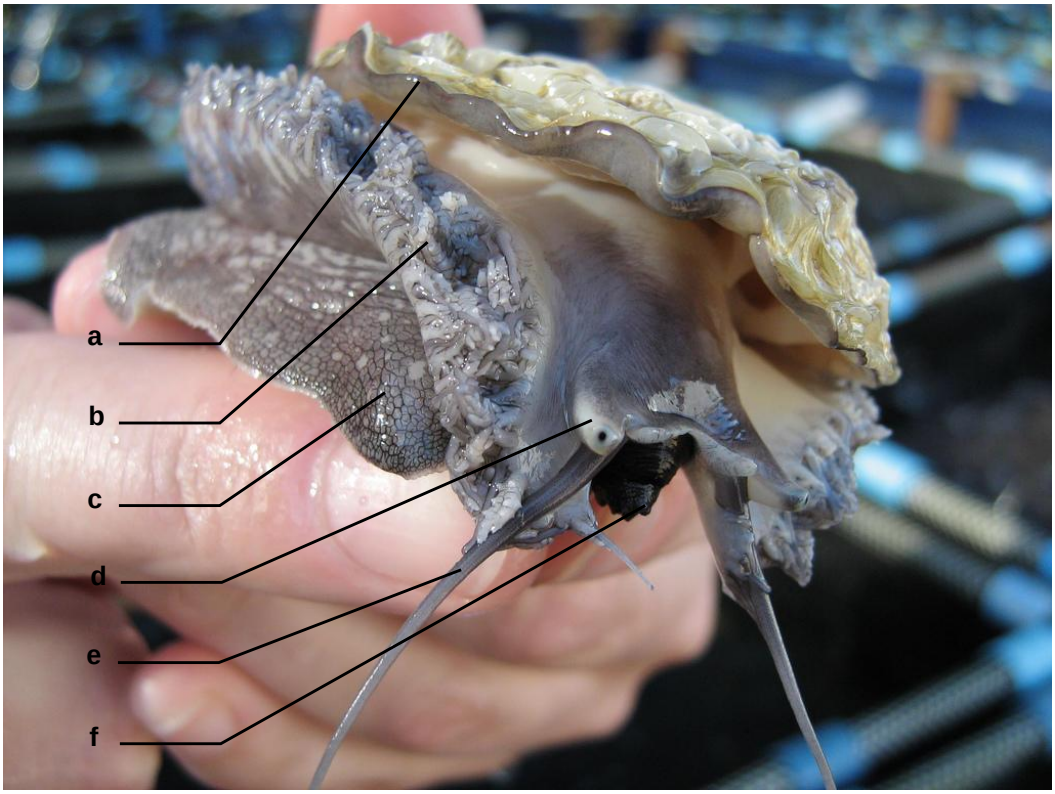


Figure 1.4: Image of a cultured *Haliotis midae* individual, illustrating the basic morphology of the body and head structures, *i.e.* the edge of the shell (**a**), the riffled mantel (**b**), the lower portion of the foot (**c**), the short, moveable eye stalks (**d**), the long, downward protruding cephalic tentacles (**e**), and the semi-mobile snout (**f**). Photograph by A. Roux, distributed under a CC BY-ND 2.0 license.

However, as their diet changes from micro- to macro-algae and their shells become larger and darker, juvenile abalone migrate to more sheltered habitats for protection, such as boulders and crevices, and even beneath other larger sedentary sea organisms, such as adult sea urchins (Day & Branch 2000). Such protection is required until they are sufficiently large to be able to rely on their own adhesion to the rock substrate to protect them from detachment by wave action and predators (Geiger & Owen 2012).

In terms of distribution, both widespread and endemic species are prevalent, although no species of global distribution currently exists. As a rule, species that prefer more temperate waters (*e.g. H. midae*, *H. rubra*, and *H. rufescens*) tend towards endemism, while the more tropical species (*e.g. H. asinia*, *H. ovina*, and *H. diversicolor*) have a much wider distribution (Geiger & Owen 2012). Four main regions of endemism have been identified,

namely, South Africa (five species), Australia (nine species), New Zealand (three species), and western North America (seven species) (Geiger 2000), while the more widespread species are predominantly distributed within the Indo-Pacific region. In terms of origin, three potential models have been proposed to explain the current morphological and geographic distribution of the *Haliotis* genus, namely, the Pacific Rim model, the Indo-Pacific model, and the Tethys model (Geiger & Owen 2012). The first suggests a star-like dispersal pattern emanating from a narrow region spanning between Japan and north-eastern Australia (Talmadge 1963). The second suggests that the region with the highest present day diversity, the Indo-Malayan area, represents the centre of radiation for the family (Lindberg 1992). The third, which is based on chromosomal count data, suggests that the evolution of halotids was characterised by a progressive increase in chromosome number, implying that the earliest species (represented today by *H. tuberculata*, 2*n* = 28) originated within the ancient Tethys Sea (of which the Mediterranean is a remnant), and subsequently radiated eastwards through the Indo-Pacific (2*n* = 32), to the North Pacific (2*n* = 36) (Geiger & Groves 1999). While there is no general consensus as yet concerning which of these models is the most likely, phylogenetic evidence is largely in support of either the Indo-Pacific (Briggs 1999; Geiger 2000; Bester-van der Merwe *et al*. 2012) or Tethys (Lee & Vacquier 1995; Bester-van der Merwe *et al*. 2012) models.

Five abalone species readily occur within South African waters, namely *H. midae*, *H. spadicea*, *H. parva*, *H. queketti*, and *H. alfredensis*, all of which are endemic (Geiger 2000) (Figure 1.5). *Haliotis midae*, or *perlemoen*, is the largest and most abundant of the five, and also the only species viable for commercial exploitation, the other species being either too small or too scarce (Roodt-Wilding & Brink 2011; Bester-van der Merwe *et al*. 2012). The natural range of *H. midae*, which is likely determined by temperature, covers roughly

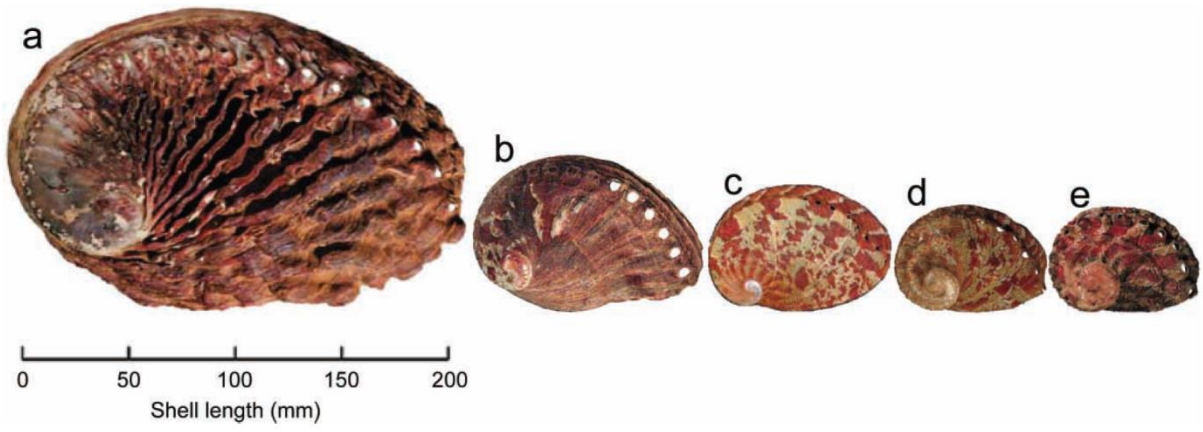


Figure 1.5: Images of shells from the five endemic South African abalone species, **a)** *Haliotis midae*, **b)** *Haliotis spadicea*, **c)** *Haliotis alfredensis*, **d)** *Haliotis parva* and **e)** *Haliotis queketti*, demonstrating their relative maximum shell lengths (Geiger & Owen 2012). Images courtesy of B. Owen.

two thirds of the South African coastline, extending from St Helena Bay, on the west coast, to Port St Johns, on the east coast (Lindberg 1992; Britz *et al*. 1997; Geiger 2000). Within this range, some level of genetic divergence has been observed between east and west coast populations on either side of Cape Agulhas, where the retroflection of the warm southward Agulhas current (east coast) caused by the cold northward Benguela current (west coast) is thought to act as a natural barrier to larval dispersal, and therefore gene flow (Bester-van der Merwe *et al*. 2011).

1.4.2) Commercial importance and exploitation in South Africa

The South African abalone industry, which is built around a single endemic abalone species, *Haliotis midae*, is an emerging, but already highly successful aquaculture production industry (Raemaekers *et al*. 2011; DAFF 2012). Abalones are highly prized for their fleshy foot, which is considered a delicacy in many countries (Oakes & Ponte 1996; Roodt-Wilding & Brink 2011), and the success of the industry has grown considerably since the establishment of the first farms in 1990 (Sales & Britz 2001), with the total production of 1036.01 tons in 2011 being valued at approximately R357 million (DAFF 2012), or 94% of the entire marine aquaculture sector. The most recent estimates indicate a total production of 1100 tons in 2013, currently valued at approximately R520 million (US$41.7 million) (FAO 2015). However, up until relatively recently, the abalone industry in South Africa was primarily fisheries based. Beginning in 1949, the fishery was initially well managed and an assessment by Tarr (1992) concluded that the prognosis for both the fishery and the resource were good. Operations were divided into seven commercial fishing zones (A-G) spread out along the southwest coast, each of which were allocated an appropriate total allowable catch (TAC), which was revised annually (Tarr 1992). Since then, a number of major changes have been effected regarding the management of participants in the fishery, as detailed in the Marine Living Resources Act (RSA 1998), to allow for more equitable access to South Africa's marine resources. Unfortunately, the unexpected boom in illegal exploitation (*i.e.* poaching) since 1994, as well as long under-regulated recreational fishing, have put populations in many of the zones under immense pressure (Tarr 2000). In addition to this, the collapse of urchin populations (*Parechinus angulosus*) within certain zones, likely the result of increased predation by the rock lobster (*Jasus lalandii*), resulted in unusually high mortality rates for juvenile abalone, which rely heavily on the shelter provided by adult urchins to survive to adulthood (Day & Branch 2000). As such, despite efforts to stabilise the fishery through decreased TACs, the fishery

was finally banned in 2008 to prevent the total collapse of wild populations (Raemaekers *et al*. 2011). A small subsistence fishery has since been reopened in 2010, but in terms of large scale commercial production, aquaculture is now the sole source of abalone in South Africa (Roodt-Wilding & Brink 2011).

1.5) Research Opportunities in Abalone

1.5.1) Genetic improvement of commercial stock

At present, *Haliotis midae* aquaculture is the largest and most lucrative of the aquaculture sectors, with 14 operational farms throughout South Africa (DAFF, 2012). However, abalone cultivation is a relatively labour- and cost intensive process, and as such, has been the subject of a number of studies aimed at improving and optimising both the aquaculture processes and the end product (Roodt-Wilding & Brink 2011). In particular, genetic improvement represents an area where significant financial gains might be achieved. As such, a breeding programme employing quantitative breeding techniques was initiated in 2006 with the aim of developing superior abalone strains for the South African abalone industry (Roodt-Wilding & Brink 2011).

However, in addition to the more general problems associated with traditional selection methods that were discussed in the previous subsection, it's application in aquaculture is made more problematic by the reproductive mode of most aquaculture species. Abalones are broadcast spawners, which makes maintaining accurate pedigree records difficult within large scale commercial operations; as a result, most facilities fall back on practicing mass selection. However, because this method does not take into account the degree to which superior individuals might be related, mass selection often leads to excessive inbreeding in aquaculture species, made worse by their characteristically high fecundity (Bentsen & Olesen 2002; Rhode *et al*. 2014). In contrast, the addition of MAS/GAS to the existing selection programme would allow for the simultaneous identification of both genetically related individuals, as well as genetically superior individuals (Dekkers & Hospital 2002), which would facilitate the large scale application of breeding programmes within the South African abalone industry. Furthermore, by estimating breeding values directly from marker genotypes, rather than indirectly from the phenotype, the accuracy of breeding value predictions will increase, resulting in improved selection efficiency (Dekkers & Hospital 2002; Liu & Cordes 2004; Yue 2014). In addition, resources could be allocated to individuals identified as breeding candidates *via* MAS at a much earlier stage to

promote rapid growth, maturation and spawning, thereby decreasing the generation interval and maximising genetic and -economic gain (Baranski *et al*. 2008).

However, despite the obvious benefits that molecular assisted breeding strategies offer, such technologies have been somewhat slow to be incorporated within the various production industries thus far (Dekkers 2004; Nieuwhof *et al*. 2008; Yue 2014). This has been particularly true within the aquaculture industry, which lags significantly behind other agricultural production industries, such as beef- and dairy cattle, sheep, pig, poultry and wheat, which have already begun benefitting significantly from the incorporation of molecular assisted breeding strategies into their breeding programmes (Dekkers 2004; Liu & Cordes 2004; Nieuwhof *et al*. 2008). However, large numbers of molecular markers, as well as linkage maps, are now available for many aquaculture species (Liu & Cordes 2004; Massault *et al*. 2008; Dunham *et al*. 2014; Yue 2014), including the South African abalone (Vervalle *et al*. 2013), ensuring that significant progress in developing the necessary tools for molecular assisted breeding in these species can be made.

1.5.2) Population genetic research

As discussed in the previous subsection, because so many factors are affected by and affect linkage disequilibrium, the levels and extent of LD within a population can provide much insight into the demographic and evolutionary processes associated with that population. Genome-wide LD patterns may aid in elucidating the population history, breeding system, and patterns of geographic subdivision of populations, whereas local LD reflects the influences of natural selection and mutation, both of which represent highly relevant areas of research within a species such as *Haliotis midae* that is currently undergoing domestication. In contrast with some other cultured abalone species, *e.g.* the Pacific abalone (*H. discus hannai*) and Thai abalone (*H. asinina*) (Li *et al*. 2004; Praipue *et al*. 2010), domestication of *H. midae* is still in the initial phases, with many aquaculture operations only recently starting to close the reproductive cycle by replacing wild broodstock with superior F_1-generation individuals, to produce improved F_2-generation animals for the international market. For this reason in particular, this species represents an ideal opportunity for further investigating the genetic effects and mechanisms associated with the process of domestication (López *et al*. 2014), as it allows the observation of how the combined effects of genetic drift, natural and/or artificial selection and reproductive mode affect the large and small scale patterns of LD within a shellfish

species. Significant population differentiation has already been observed between wild and cultured populations (*e.g.* Slabbert *et al*. 2009; Rhode *et al*. 2012, 2014), which will likely have had a marked effect on the nature of LD within cultured populations. Furthermore, Rhode *et al*. (2013) recently identified a number of microsatellite markers as being under directional selection between wild and cultured populations, however, it is still unknown as to whether this selection is the result of a recent selection event (domestication), or a more historic selection event. In this case, the extent of syntenic LD surrounding these markers might give an indication of how recently the event occurred, as LD tends to decay over time as more opportunities arise for recombination between linked loci. Additionally, as large-scale LD data becomes available for other halotid species as well, across species comparisons will present new opportunities to elucidate the demographic and evolutionary histories of the *Haliotis* genus, which is as yet not certain (Degnan *et al*. 2006; Bester-van der Merwe *et al*. 2012; Geiger & Owen 2012). As such, the availability of LD data for this species would contribute significantly towards the further characterisation of the abalone species, and a greater understanding of how genetic and environmental factors can and have shaped both wild and cultured populations.

1.6) Study Rationale, Aims and Objectives

1.6.1) Problem statement

Prior to 2004, very little was available in terms of the genetic resources for this species, however, much progress has been made in this regard. Basic genomic information, such as somatic chromosome number (2*n* = 36; Van der Merwe & Roodt-Wilding 2008) and genome size (~1400 cM, Franchini *et al*. 2010), has been determined, and a sizeable collection of both microsatellite (Bester *et al*. 2004; Slabbert *et al*. 2008; Hepple 2010; Rhode 2010; Slabbert *et al*. 2010; Jansen 2012; Slabbert *et al*. 2012) and single nucleotide polymorphism (SNP) markers (Bester *et al*. 2008; Rhode *et al*. 2008; Rhode 2010; Blaauw 2012; Du Plessis 2012) are now also available. Recently, an integrated *Haliotis midae* linkage map constructed from 116 microsatellite and 70 SNP markers has been completed (Vervalle *et al*. 2013), with an average marker spacing of 6.88 cM, which has been determined as sufficient for QTL studies within aquaculture species (Massault *et al*. 2008). However, no large scale LD data has as yet been generated, which is not only a prerequisite for performing association analyses, but would also provide a large number of opportunities for population genetic research in this species.

1.6.2) Project aims and objectives

In light of the numerous benefits inherent in having genome-wide LD data available, the primary focus of this study was, therefore, to quantify and characterise the levels of genome-wide LD within the South African abalone, *Haliotis midae*, so as to make this genetic resource available for the diverse applications for which it can be used within the context of abalone. For this purpose, the following objectives were set:

1) Calculate the extent of genome-wide LD between syntenic marker pairs from the *H. midae* linkage map within two wild and two cultured cohorts of abalone, and evaluate the decay of LD with genetic distance.
2) Calculate the extent of LD among non-syntenic marker pairs within each study cohort, and determine baseline levels of LD within these cohorts.
3) Compare the levels of syntenic and non-syntenic LD within each cohort and determine the extent to which their differing demographic histories have influenced the magnitude and extent of LD therein.
4) Determine the feasibility of association studies within these cohorts based on current levels of LD.

A secondary aim was to apply the phenomenon of LD in an investigation to determine whether the set of markers identified by Rhode *et al*. (2013) as being under putative divergent selection between wild and cultured *H. midae* are more likely to be associated with natural- or artificial selection events during domestication. To achieve this aim, the following objectives were set:

1) Phenotype a population of first generation cultured abalone for growth-related traits, *i.e.* live weight, shell length and shell width, and genotype the top and bottom 15% based on live weight using the candidate loci under selection.
2) Test for genotype-phenotype associations between allelic variants and size using a selection of association tests.
3) Determine whether artificial selection, as opposed to natural selection, is the dominant selective force in causing the observed signatures of selection at these loci.

1.6.3) Thesis layout

The four objectives for the primary aim will be addressed within the first experimental chapter (Chapter 2), which will deal with quantifying the levels of genome-wide LD within the four abalone cohorts, and discussing the results within the context of their respective demographic/evolutionary histories, as well as how this affects the feasibility of association studies within these cohorts. The second experimental chapter (Chapter 3) will address the objectives of the second aim, which is to determine whether a small set of microsatellite markers, previously determined as being under divergent selection between wild and cultured populations of *Haliotis midae*, are more likely to be associated with artificial selection events occurring during domestication, or with natural selection for adaptation to the new aquaculture environment. The final chapter (Chapter 4) will summarise the various findings of the previous chapters, and discuss the results within the broader context of the study aims, highlighting how the current study demonstrates the utility of LD data within population genetic studies. Lastly, this chapter will also discuss the limitations of the current study and provide suggestions for future research.

References

Abecasis G.R., Noguchi E., Heinzmann A., Traherne J.A., Bhattacharyya S., Leaves N.I., Anderson G.F., Zhang Y., Lench N.J., Carey A., Cardon L.R., Moffatt M.F. & Cookson W.O. (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *The American Journal of Human Genetics* 68, 191-197.

Albrechtsen A., Castella S., Anderson G., Hansen T., Pederson O. & Nielsen R. (2007) A Bayesian multilocus association method: Allowing for higher-order interaction in association studies. *Genetics* 176, 1197-1206.

Ardlie K.G., Kruglyak L. & Seielstad M. (2002) Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics* 3, 299-309.

Auton A., Fledel-Alon A., Pfeifer S., Venn O., Ségurel L., Street T., Leffler E.M., Bowden R., Aneas I., Broxholme J., Humburg P., Iqbal Z., Lunter G., Maller J., Hernandez R.D., Melton C., Venkat A., Nobrega M.A., Bontrop R., Myers S., Donnelly P., Przeworski M. & McVean G. (2012) A fine-scale chimpanzee genetic map from population sequencing. *Science* 336, 193-198.

Auton A., Li Y.R., Kidd J., Oliveira K., Nadel J., Holloway J.K., Hayward J.J., Cohen P.E., Greally J.M., Wang J., Bustamante C.D. & Boyko A.R. (2013) Genetic recombination is targeted towards gene promoter regions in dogs. *PLoS Genetics* 9, e1003984.

Backström N., Forstmeier W., Schielzeth H., Mellenius H., Nam K., Bolund E., Webster M.T., Öst T., Schneider M., Kempenaers B. & Ellegren H. (2010) The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Research* 20, 485-495.

Baird S.J. (2015) Exploring linkage disequilibrium. *Molecular Ecology Resources* 15, 1017-1019.

Balloux F., Lehmann L. & de Meeûs T. (2003) The population genetics of clonal and partially clonal diploids. *Genetics* 164, 1635-1644.

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.

Barton N. & Clark A. (1990) Population structure and process in evolution. In: Wohrmann K. & Jain S.K. (eds), *Population Biology: Ecological and Evolutionary Viewpoints*. Berlin: Springer Science & Business Media. pp 115-173.

Bateson W., Saunders E.R. & Punnett R.C. (1905) Experimental studies in the physiology of heredity. *Reports to the Evolution Committee of the Royal Society 2*, 1-55, 80-99.

Baudat F., Buard J., Grey C., Fledel-Alon A., Ober C., Przeworski M., Coop G. & De Massy B. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327, 836-840.

Ben-Ari G., Zenvirth D., Sherman A., David L., Klutstein M., Lavi U., Hillel J. & Simchen G. (2006) Four linked genes participate in controlling sporulation efficiency in budding yeast. *PLoS Genetics* 2, e195.

Bentsen H.B. & Olesen I. (2002) Designing aquaculture mass selection programs to avoid high inbreeding rates. *Aquaculture* 204, 349-359.

Berg I.L., Neumann R., Lam K.W.G., Sarbajna S., Odenthal-Hesse L., May C.A., & Jeffreys A.J. (2010) *PRDM9* variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genetics* 42, 859-863.

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.

Bester-van der Merwe A.E., D'Amato M.E., Swart B.L. & Roodt-Wilding R. (2012) Molecular phylogeny of South African abalone, its origin and evolution as revealed by two genes. *Marine Biology Research* 8, 727-736.

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

Blaauw S. (2012) SNP screening and validation in *Haliotis midae*. Unpublished MSc thesis, Stellenbosch University, South Africa.

Blixt S. (1975) Why didn't Gregor Mendel find linkage? *Nature* 256, 206.

Bomba L., Nicolazzi E.L., Milanesi M., Negrini R., Mancini G., Biscarini F., Stella A., Valentini A. & Ajmone-Marsan P. (2015) Relative extended haplotype homozygosity signals across breeds reveal dairy and beef specific signatures of selection. *Genetics Selection Evolution* 47, 25.

Briggs J.C. (1999) Coincident biogeographic patterns: Indo-west Pacific Ocean. *Evolution* 53, 326-335.

Britz P.J., Hecht T. & Mangold S. (1997) Effect of temperature on growth, feed consumption and nutritional indices of *Haliotis midae* fed a formulated diet. *Aquaculture* 152, 191-203.

Brown A.H.D., Feldman M.W. & Nevo E. (1980) Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* 96, 523-536.

Carneiro M., Piorno V., Rubin C.-J., Alves J.M., Ferrand N., Alves P.C. & Andersson L. (2015) Candidate genes underlying heritable differences in reproductive seasonality between wild and domestic rabbits. *Animal Genetics* 46, 418-425.

Chan E.K., Rowe H.C. & Kliebenstein D.J. (2010a) Understanding the evolution of defense metabolites in *Arabidopsis thaliana* using genome-wide association mapping. *Genetics* 185, 991-1007.

Chan E.K., Rowe H.C., Hansen B.G. & Kliebenstein D.J. (2010b) The complex genetic architecture of the metabolome. *PLoS Genetics* 6, e1001198.

Charlesworth B. (2009) Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* 10, 195-205.

Charlesworth D. & Wright S. (2001) Breeding systems and genome evolution. *Current Opinion in Genetics and Development* 11, 685-690.

Chial H. (2008) Mendelian genetics: Patterns of inheritance and single-gene disorders. *Nature Education 1*, 3775-3781.

Choi K., Zhao X., Kelly K.A., Venn O., Higgins J.D., Yelina N.E., Hardcastle T.J., Ziolkowski P.A., Copenhaver G.P., Franklin F.C.H., McVean G. & Henderson I.R. (2013) *Arabidopsis* meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nature Genetics* 45, 1327-1336.

Corbin L.J., Blott S.C., Swinburne J.E., Vaudin M., Bishop S.C. & Woolliams J.A. (2010) Linkage disequilibrium and historical effective population size in the Thoroughbred horse. *Animal Genetics* 41, 8-15.

Correns C. (1900) Gregor Mendel's Regel über das Verhalten der Nachkommenschaft der Bastarde. *Berichte der Deutschen Botanischen Gesellschaft* 18, 158-168.

Cramér H. (1946) *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Culleton R., Martinelli A., Hunt P. & Carter R. (2005) Linkage group selection: rapid gene discovery in malaria parasites. *Genome Research* 15, 92-97.

Daetwyler H.D., Pong-Wong R., Villanueva B. & Woolliams J.A. (2010) The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185, 1021-1031.

Daly M.J., Rioux J.D., Schaffner S.E., Hudson T.J. & Lander E.S. (2001) High-resolution haplotype structure in the human genome. *Nature Genetics* 29, 229-232.

Day E. & Branch G.M. (2000) Evidence for a positive relationship between juvenile abalone *Haliotis midae* and the sea urchin *Parechinus angulosus* in the South-Western Cape, South Africa. *South African Journal of Marine Science* 22, 145-156.

De Vries H. (1900) Das Spaltungsgesetz der Bastarde. *Berichte der Deutschen Botanischen Gesellschaft* 18, 83-90.

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.

Dekkers J.C.M. (2004) Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. *Journal of Animal Science* 82, E-Suppl:E313-328.

Dekkers J.C.M. & Hospital F. (2002) The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics* 3, 22-32.

Department of Agriculture, Forestry and Fisheries (DAFF) (2012) Aquaculture Yearbook 2012. South Africa.

Drouaud J., Camilleri C., Bourguignon P.Y., Canaguier A., Bérard A., Vezon D., Giancola S., Brunel D., Colot V., Prum B., Quesneville H. & Mézard C. (2006) Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination “hot spots”. *Genome Research* 16, 106-114.

Du Plessis J. (2012) High-throughput SNP genotyping and linkage mapping in *Haliotis midae*. Unpublished MSc thesis, Stellenbosch University, South Africa.

Dunham R.A., Taylor J.F., Rise M.L. & Liu Z. (2014) Development of strategies for integrated breeding, genetics and applied genomics for genetic improvement of aquatic organisms. *Aquaculture* 420, 121-123.

England P.R., Cornuet J.M., Berthier P., Tallmon D.A. & Luikart G. (2006) Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conservation Genetics* 7, 303-308.

Evin A., Dobney K., Schafberg R., Owen J., Vidarsdottir U.S., Larson G. & Cucchi T. (2015) Phenotype and animal domestication: A study of dental variation between domestic, wild, captive, hybrid and insular *Sus scrofa*. *BMC Evolutionary Biology* 15, 6.

Farnir F., Coppieters W., Arranz J.J., Berzi P., Cambisano N., Grisart B., Karim L., Marcq F., Moreau L., Mni M., Nezer C., Simon P., Vanmanshoven P., Wagenaar D. & Georges M. (2000) Extensive genome-wide linkage disequilibrium in cattle. *Genome Research* 10, 220-227.

Feder M.E. & Mitchell-Olds T. (2003) Evolutionary and ecological functional genomics. *Nature Reviews Genetics* 4, 649-699.

Flint-Garcia S.A., Thornsberry J.M. & Buckler IV E.S. (2003) Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* 54, 357-374.

Food and Agriculture Organisation (FAO) Global Aquaculture Production 1950-2013. Accessed: 18/08/2015 http://www.fao.org/figis/servlet/SQServlet?file=/work/FIGIS/prod/webapps/figis/temp/hqp_5749472520183685249.xml&outtype=html.

Franchini P., Slabbert R., van der Merwe M., Roux A. & Roodt-Wilding R. (2010) Karyotype and genome size estimation of *Haliotis midae*: estimators to assist future studies on the evolutionary history of Haliotidae. *Journal of Shellfish Research* 29, 945-950.

Freeman A.R., Meghen C.M., Machugh D.E., Loftus R.T., Achukwi M.D., Bado A., Sauveroche B. & Bradley D.G. (2004) Admixture and diversity in West African cattle populations. *Molecular Ecology* 13, 3477-3487.

Frisse L., Hudson R.R., Bartoszewicz A., Wall J.D., Donfack J. & Di Rienzo A. (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *The American Journal of Human Genetics* 69, 831-843.

Fullerton S.M., Carvalho A.B. & Clark A.G. (2001) Local rates of recombination are positively correlated with GC content in human genome. *Molecular Biology and Evolution* 18, 1139-1142.

Gaut B.S. & Long A.D. (2003) The lowdown on linkage disequilibrium. *The Plant Cell* 15, 1502-1506.

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

Geiger D.L. & Groves L.T. (1999) Review of fossil abalone (Gastropoda: Vetigastropoda: Haliotidae) with comparison to recent species. *Journal of Paleontology* 73, 872-885.

Geiger D.L. & Owen B. (2012) *Abalone: Worldwide Haliotidae.* Hackenheim: ConchBooks.

Geiringer H. (1944) On the probability theory of linkage in Mendelian heredity. *Annals of Mathematical Statistics* 15, 25-57.

Gholami M., Reimer C., Erbe M., Preisinger R., Weigend A., Weigend S., Servin B. & Simianer H. (2015) Genome scan for selection in structured layer chicken populations exploiting linkage disequilibrium information. *PLoS ONE* 10, e0130497.

Giraut L., Falque M., Drouaud J., Pereira L., Martin O.C. & Mézard C. (2011) Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along chromosomes. *PLoS Genetics* 7, e1002354.

Goddard M.E. & Hayes B.J. (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics* 10, 381-391.

Goldstein D.B. (2001) Islands of linkage disequilibrium. *Nature Genetics* 29, 109-111.

Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M.H.-Y., Hansen N.F., Durand E.Y., Malaspinas A.-S., Jensen J.D., Marques-Bonet T., Alkan C., Prüfer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-Petri A., Butthof A., Höber B., Höffner B., Siegemund M., Weihmann A., Nusbaum C., Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic D., Kucan Z., Gušic I., Doronichev V.B., Golovanova L.V., Lalueza-Fox C., de la Rasilla M., Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D., Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Pääbo S. (2010) A draft sequence of the Neandertal genome. *Science* 328, 710-722.

Gregersen J.W., Kranc K.R., Ke X., Svendsen P., Madsen L.S., Thomsen A.R., Cardon L.R., Bell J.I. & Fugger L. (2006) Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature* 443, 574-577.

Groenen M.A., Wahlberg P., Foglio M., Cheng H.H., Megens H.J., Crooijmans R.P., Besnier F., Lathrop M., Muir W.M., Wong G.K.-S., Gut I. & Andersson L. (2009) A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Research* 19, 510-519.

Hanotte O., Bradley D.G., Ochieng J.W., Verjee Y., Hill E.W. & Rege J.E. (2002) African pastoralism: genetic imprints of origins and migrations. *Science* 296, 336-339.

Harmegnies N., Farnir F., Davin F., Buys N., Georges M. & Coppieters W. (2006) Measuring the extent of linkage disequilibrium in commercial pig populations. *Animal Genetics* 37, 225-231.

Hassold T., Sherman S. & Hunt P. (2000) Counting cross-overs: characterizing meiotic recombination in mammals. *Human Molecular Genetics* 9, 2409-2419.

Hayes B.J., Baranski M., Goddard M.E. & Robinson N. (2007) Optimisation of marker assisted selection for abalone breeding programs. *Aquaculture* 256, 61-69.

Hayes B.J., Gjuvsland A. & Omholt S. (2006) Power of QTL mapping experiments in commercial Atlantic salmon populations, exploiting linkage and linkage disequilibrium and effect of limited recombination in males. *Heredity* 97, 19-26.

Hayes B.J., Visscher P.M., McPartlan H.C. & Goddard M.E. (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research* 13, 635-643.

Hedrick P.W. (1987) Gametic disequilibrium measures: proceed with caution. *Genetics* 117, 331-341.

Heifetz E.M., Fulton J.E., O'Sullivan N., Zhao H., Dekkers J.C. & Soller M. (2005) Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. *Genetics* 171, 1173-1181.

Hepple J.-A. (2010) An integrated linkage map of Perlemoen (*Haliotis midae*). Unpublished MSc thesis, Stellenbosch University, South Africa.

Hill E. (2009) Genetics 1900-1910. On-line publication. In: Cain J. (ed.), *First Class Essays*.

Hill W.G. (1975) Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology* 8, 117-126.

Hill W.G. (1981) Estimation of effective population size from data on linkage disequilibrium. *Genetical Research* 38, 209-216.

Hill W.G. & Robertson A. (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* 38, 226-231.

Hohenlohe P.A., Bassham S., Currey M. & Cresko W.A. (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society of London B* 367, 395-408.

Horn R.L., Kuehn R., Drechsel V. & Cowley D.E. (2014) Discriminating between the effects of founding events and reproductive mode on the genetic structure of *Triops* populations (Branchiopoda: Notostraca). *PLoS ONE* 9, e97473.

Hudson R.R. (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* 7, 44.

Hudson R.R. & Kaplan N.L. (1986) On the divergence of alleles in nested subsamples from finite populations. *Genetics* 113, 1057-1076.

Jansen S. (2012) Linkage mapping in *Haliotis midae* using gene-linked markers. Unpublished MSc thesis, Stellenbosch University, South Africa.

Kalantari B., Lari I., Rizzi A. & Simeone B. (1993) Sharp bounds for the maximum of the chi-square index in a class of contingency tables with given marginal. *Computational Statistics & Data Analysis* 16, 19-34.

Kim S., Plagnol V., Hu T.T., Toomajian c., Clark R.M., Ossowski S., Ecker J.R., Weigel D. & Nordborg M. (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* 39, 1151-1155.

Kim Y., Feng S. & Zeng Z.-B. (2008) Measuring and partitioning the high-order linkage disequilibrium by multiple order Markov chains. *Genetic Epidemiology* 32, 301-312.

Kimura M. (1984) *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.

Kliebenstein D.J. (2009) A quantitative genetics and ecological model system: understanding the aliphatic glucosinolate biosynthetic network *via* QTLs. *Phytochemistry Reviews* 8, 243-254.

Knapp S.J. (1998) Marker-assisted selection as a strategy for increasing the probability of selecting superior genotypes. *Crop Science* 38, 1164-1174.

Korol A., Shirak A., Cnaani A. & Hallerman E.M. (2007) Detection and analysis of quantitative trait loci (QTL) for economic traits in aquatic species. In: Liu Z.J. (ed.), *Aquaculture Genome Technologies*. Iowa: Blackwell Publishing. pp 169-197.

Kuhn A., Ong Y.M., Cheng C.Y., Wong T.Y., Quake S.R. & Burkholder W.F. (2014) Linkage disequilibrium and signatures of positive selection around LINE-1 retrotransposons in the human genome. *Proceedings of the National Academy of Sciences* 111, 8131-8136.

Lee Y.H. & Vacquier V.D. (1995) Evolution and systematics in Haliotidae (Mollusca: Gastropoda): inferences from DNA sequences of sperm lysin. *Marine Biology* 124, 267-278.

Lewontin R.C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49, 49-67.

Lewontin R.C. & Kojima K. (1960) The evolutionary dynamics of complex polymorphisms. *Evolution* 14, 458-472.

Li M., Tian S., Jin L., Zhou G., Li Y., Zhang Y., Wang T., Yeung C.K.L., Chen L., Ma J., Zhang J., Jiang A., Li J., Zhou C., Zhang J., Liu Y., Sun X., Zhao H., Niu Z., Lou P., Xian L., Shen X., Liu S., Zhang S., Zhang M., Zhu L., Shuai S., Bai L., Tang G., Liu H., Jiang Y., Mai M., Xiao J., Wang X., Zhou Q., Wang Z., Stothard P., Xue M., Gao X., Luo Z., Gu Y., Zhu H., Hu X., Zhao Y., Plastow G.S., Wang J., Jiang Z., Li K., Li N., Li X. & Li R. (2013) Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nature Genetics* 45, 1431-1438.

Li M.-H. & Merilä J. (2010a) Extensive linkage disequilibrium in a wild bird population. *Heredity* 104, 600-610.

Li M.-H. & Merilä J. (2010b) Sex-specific population structure, natural selection, and linkage disequilibrium in a wild bird population as revealed by genome-wide microsatellite analyses. *BMC Evolutionary Biology* 10, 66.

Li Q., Park Q., Endo T. & Kijima A. (2004) Loss of genetic variation at microsatellite loci in hatchery strains of the Pacific abalone (*Haliotis discus hannai*). *Aquaculture* 235, 207-222.

Lindberg D.R. (1992) Evolution, distribution and systematics of Haliotidae. In: Shepherd S.A., Tegner M.J. & Guzman del Proo S.A. (eds), *Abalone of the world. Biology, fisheries and culture*. Oxford: Fishing News Books. pp 3-18.

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

Long A.D. & Langley C.H. (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research* 9, 720-731.

López M.D., Guerra J.M. & Samuelsson T. (2010) Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLoS ONE* 5, e10654.

López M.E., Neira R. & Yáñez J.M. (2014) Applications in the search for genomic selection signatures in fish. *Frontiers in Genetics* 5, 458.

Lu Y., Zhang S., Shah T., Xie C., Hao Z., Li X., Farkhari M., Ribaut J.-M., Cao M., Rong T. & Xu Y. (2010) Joint linkage–linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. *Proceedings of the National Academy of Sciences* 107, 19585-19590.

Luikart G., Ryman N., Tallmon D.A., Schwartz M.K. & Allendorf F.W. (2010) Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conservation Genetics* 11, 355-373.

Maccaferri M., Sanguineti M.C., Demontis A., El-Ahmed A., Garcia del Moral L., Maalouf F., Nachit M., Nserallah N., Ouabbou H., Rhouma S., Royo C., Villegas D. & Tuberosa R. (2010) Association mapping in durum wheat grown across a broad range of water regimes. *Journal of Experimental Botany* 62, 409-438.

MacHugh D.E., Shriver M.D., Loftus R.T., Cunningham P. & Bradley D.G. (1997) Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics* 146, 1071-1086.

Mackay T.F.C. (2001) The genetic architecture of quantitative traits. *Annual Review of Genetics* 35, 303-339.

Marshall K., Maddox J.F., Lee S.H., Zhang Y., Kahn L., Graser H.U., Gondro C., Walkden-Brown S.W. & Van Der Werf J.H.J. (2009) Genetic mapping of quantitative trait loci for resistance to *Haemonchus contortus* in sheep. *Animal Genetics* 40, 262-272.

Massault C., Bovenhuis H., Haley C.S. & de Koning D.J. (2008) QTL mapping designs for aquaculture. *Aquaculture* 285, 23-29.

Matala A.P., Gray A.K., Heifetz J. & Gharrett A.J. (2004) Population structure of Alaskan shortraker rockfish, *Sebastes borealis*, inferred from microsatellite variation. *Environmental Biology of Fishes* 69, 201-210.

Maynard Smith J. & Haigh J. (1974) The hitch-hiking effect of a favourable gene. *Genetical Research* 23, 23-35.

McPeek M.S., & Speed T.P. (1995) Modelling interference in genetic recombination. *Genetics* 139, 1031-1044.

McRae A.F., McEwan J.C., Dodds K.G., Wilson T., Crawford A.M. & Slate J. (2002) Linkage disequilibrium in domestic sheep. *Genetics* 160, 1113-1122.

McVean G. (2007) The structure of linkage disequilibrium around a selective sweep. *Genetics* 175, 1395-1406.

McVean G.A.T., Myers S.R., Hunt S., Deloukas P., Bentley D.R. & Donnelly P. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* 304, 581-584.

Meadows J.R.S., Chan E.K.F. & Kijas J.W. (2008) Linkage disequilibrium compared between five populations of domestic sheep. *BMC Genetics* 9, 61.

Mendel G. (1866) Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines Brünn* 4, 3-47.

Meuwissen T.H.E. & Goddard M.E. (2004) Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics Selection Evolution* 36, 1-20.

Meuwissen T.H.E., Karlsen A., Lien S., Olsaker I. & Goddard M.E. (2002) Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* 161, 373-379.

Miko I. (2008) Epistasis: Gene interaction and phenotype effects. *Nature Education* 1, 197.

Moen T., Hayes B., Baranski M., Berg P.R., Kjøglum S., Koop B.F., Davidson W.S., Omholt S.W. & Lien S. (2008) A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers. *BMC Genomics* 9, 223.

Morgan T.H. (1910) Sex-limited inheritance in *Drosophila*. Science 132, 120-122.

Morgan T.H. (1911) Random segregation *versus* coupling in Mendelian inheritance. *Science* 34, 384.

Morton N.E., Zhang W., Taillon-Miller P., Ennis S., Kwok P.-Y. & Collins A. (2001) The optimal measure of allelic association. *Proceedings of the National Academy of Sciences of the USA* 98, 5217-5221.

Mueller J.C. (2004) Linkage disequilibrium for different scales and applications. *Briefings in Bioinformatics* 5, 355-364.

Myers S., Bowden R., Tumian A., Bontrop R.E., Freeman C., MacFie T.S., McVean G. & Donnelly P. (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* 327, 876-879.

Nieuwhof G.J., Winters M., Avendano S. (2008) The existing level of uptake of molecular techniques in animal breeding. In: Rosati A., Tewolde A., & Mosconi C. (eds), *Animal Production and Animal Science Worldwide: WAAP Book of the Year 2007*. Netherlands: Wageningen Academic Publishers. pp 195-201.

Nordborg M. (2000) Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* 154, 923-929.

Nordborg M. & Tavaré S. (2002) Linkage disequilibrium: what history has to tell us. *Trends in Genetics* 18, 83-90.

Nsengimana J., Baret P., Haley C.S. & Visscher P.M. (2004) Linkage disequilibrium in the domesticated pig. *Genetics* 166, 1395-1404.

Oakes F.R. & Ponte R.D. (1996) The abalone market: Opportunities for cultured abalone. *Aquaculture* 140, 187-195.

Ohta T. (1982) Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proceedings of the National Academy of Sciences of the USA* 79, 1940-1944.

Olsen H.G., Lien S., Svendsen M., Nilsen H., Roseth A., Opsal M.A. & Meuwissen T.H.E. (2004) Fine mapping of milk production QTL on BTA6 by combined linkage and linkage disequilibrium analysis. *Journal of Dairy Science* 87, 690-698.

Olsen H.G., Meuwissen T.H.E., Nilsen H., Svendsen M. & Lien S. (2008) Fine mapping of quantitative trait Loci on bovine chromosome 6 affecting calving difficulty. *Journal of Dairy Science* 91, 4312-4322.

Østman B., Hintze A. & Adami C. (2011) Impact of epistasis and pleiotropy on evolutionary adaptation. *Proceedings of the Royal Society of London B* 279, 247-256.

Ozkilinc H., Thomas K., Abang M. & Peever T.L. (2015) Population structure and reproductive mode of *Didymella fabae* in Syria. *Plant Pathology*, doi: 10.1111/ppa.12359.

Pan J., Sasaki M., Kniewel R., Murakami H., Blitzblau H.G., Tischfield S.E., Zhu X., Neale M.J., Jasin M., Socci N.D., Hochwagen A. & Keeney S. (2011) A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* 144, 719-731.

Pe'er I., Chretien Y.R., de Bakker P.I., Barrett J.C., Daly M.J. & Altshuler D.M. (2006) Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *The American Journal of Human Genetics* 78, 588-603.

Peel D., Waples R.S., Macbeth G.M., Do C. & Ovenden J.R. (2013) Accounting for missing data in the estimation of contemporary genetic effective population size (Ne). *Molecular Ecology Resources* 13, 243-253.

Poelstra J.W., Ellegren H. & Wolf J.B.W. (2013) An extensive candidate gene approach to speciation: diversity, divergence and linkage disequilibrium in candidate pigmentation genes across the European crow hybrid zone. *Heredity* 111, 467-473.

Praipue P., Klinbunga S. & Jarayabhand P. (2010) Genetic diversity of wild and domesticated stocks of Thai abalone, *Haliotis asinina* (Haliotidae), analyzed by single-strand conformational polymorphism of AFLP-derived markers. *Genetics and Molecular Research* 9, 1136-1152.

Qanbari S., Hansen M., Weigend S., Preisinger R. & Simianer H. (2010) Linkage disequilibrium reveals different demographic history in egg laying chickens. *BMC Genetics* 11, 103.

Qanbari S. & Simianer H. (2014) Mapping signatures of positive selection in the genome of livestock. *Livestock Science* 166, 133-143.

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.

Reich D.E., Cargill M., Bolk S., Ireland J., Sabeti P.C., Richter D.J., Lavery T., Kouyoumjian R., Farhadian S.F., Ward R. & Lander E.S. (2001) Linkage disequilibrium in the human genome. *Nature* 411, 199-204.

Republic of South Africa (RSA) (1998) Marine Living Resources Act (Act No. 18 of 1998). Government Gazette, South Africa 395(18930).

Rexroad C.E. & Vallejo R.L. (2009) Estimates of linkage disequilibrium and effective population size in rainbow trout. *BMC Genetics* 10, 83.

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

Rhode C., Hepple J., Jansen S., Davis T., Vervalle J., Bester-van der Merwe A.E. & Roodt-Wilding R. (2012) A population genetics analysis of abalone domestication events in South Africa: Implications for the management of the abalone resource. *Aquaculture* 356–357, 235-242.

Rhode C., Maduna S.N., Roodt-Wilding R. & Bester-van der Merwe A.E. (2014) Comparison of population genetic estimates amongst wild, F1 and F2 cultured abalone (*Haliotis midae*). *Animal Genetics* 45, 456-459.

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

Rhode C., Vervalle J., Bester-van der Merwe A.E. & Roodt-Wilding R. (2013) Detection of molecular signatures of selection at microsatellite loci in the South African abalone (*Haliotis midae*) using a population genomic approach. *Marine Genomics* 10, 27-36.

Risch N.J. (2000) Searching for genetic determinants in the new millennium. *Nature* 405, 847-856.

Risch N.J., de Leon D., Ozelius L., Kramer P., Almasy L., Singer B., Fahn S., Breakefield X. & Bressman S. (1995) Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nature Genetics* 9, 152-159.

Roberts R. (2001) A review of settlement cues for larval abalone (*Haliotis* spp.). *Journal of Shellfish Research* 20, 571-586.

Robinson J.D. & Moyer G.R. (2013) Linkage disequilibrium and effective population size when generations overlap. *Evolutionary Applications* 6, 290-302.

Roodt-Wilding R. & Brink D. (2011) Selection and sea snails: the South African story. *Philosophical Transactions in Genetics* 1, 1-40.

Rubin C.-J., Zody M.C., Eriksson J., Meadows J.R., Sherwood E., Webster M.T., Jiang L., Ingman M., Sharpe T., Ka S., Hallböök F., Besnier F., Carlborg Ö., Bed'hom B., Tixier-Boichard M., Jensen P., Siegel P., Lindblad-Toh K. & Andersson L. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464, 587-591.

Sabeti P.C., Reich D.E., Higgins J.M., Levine H.Z., Richter D.J., Schaffner S.F., Gabriel S.B., Platko J.V., Patterson N.J., McDonald G.J., Ackerman H.C., Campbell S.J.,

Altshuler D. Cooper R., Kwiatkowski D., Ward R. & Lander E.S. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832-837.

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

Sankararaman S., Patterson N., Li H., Pääbo S. & Reich, D. (2012) The date of interbreeding between Neandertals and modern humans. *PLoS Genetics* 8, e1002947.

Saura M., Woolliams J.A., Tenesa A., Fernández A. & Villanueva B. (2014) Estimation of ancient and recent effective population size from linkage disequilibrium in a closed herd of Iberian pigs. In: *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production*, 17-22 August, Vancouver, Canada. Vancouver: American Society of Animal Science.

Schwartz M.K., Luikart G. & Waples R.S. (2007) Genetic monitoring as a promising tool for conservation and management. *Trends in Ecology and Evolution* 22, 25-33.

Singhal S., Leffler E., Sannareddy K., Turner I., Venn O., Hooper D., Strand A., Li Q., Raney B., Balakrishnan C., Griffith S., McVean G. & Przeworski M. (2015) Stable recombination hotspots in birds. bioRxiv doi: 10.1101/023101.

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

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

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

Slate J. & Pemberton J.M. (2007) Admixture and patterns of linkage disequilibrium in a free-living vertebrate population. *Journal of Evolutionary Biology* 20, 1415-1427.

Slatkin M. (1994) Linkage disequilibrium in growing and stable populations. *Genetics* 137, 331-336.

Slatkin M. (2008) Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* 9, 477-485.

Smagulova F., Gregoretti I.V., Brick K., Khil P., Camerini-Otero R.D. & Petukhova G.V. (2011) Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature* 472, 375-378.

Smýkal P. (2014) Pea (Pisum sativum L.) in biology prior and after Mendel's discovery. *Czech Journal of Genetics and Plant Breeding* 50, 52-64.

Sonesson A.K. & Meuwissen T.H. (2009) Testing strategies for genomic selection in aquaculture breeding programs. *Genetics Selection Evolution* 41, 37.

Stapper A.P., Beerli P. & Levitan D.R. (2015) Assortative mating drives linkage disequilibrium between sperm and egg recognition protein loci in the sea urchin *Strongylocentrotus purpuratus*. *Molecular Biology and Evolution* 32, 859-870.

Stephens J.C., Reich D.E., Goldstein D.B., Shin H.D., Smith M.W., Carrington M., Winkler C., Huttley G.A., Allikmets R., Schriml L., Gerrard B., Malasky M., Ramos M.D., Morlot S., Tzetis M., Oddoux C., di Giovine F.S., Nasioulas G., Chandler D., Aseev M., Hanson M., Kalaydjieva L., Glavac D., Gasparini P., Kanavakis E., Claustres M., Kambouris M., Ostrer H., Duff G., Baranov V., Sibul H., Metspalu A., Goldman D., Martin N., Duffy D., Schmidtke J., Estivill X., O'Brien S.J. & Dean M. (1998) Dating the origin of the *CCR5-Δ32* AIDS-resistance allele by the coalescence of haplotypes. *The American Journal of Human Genetics* 62, 1507-1515.

Stinchcombe J.R. & Hoekstra H.E. (2007) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* 100, 158-170.

Storz J.F. (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* 14, 671-688.

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* 14, 43-59.

Sved J.A. (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* 2, 125-141.

Talmadge R.R. (1963) Insular haliotids in the western Pacific (Mollusca: Gastropoda). *Veliger* 5, 129-139.

Tarr R.J.Q. (1992) The abalone fishery of South Africa. In: Shepherd S.A., Tegner M.J. & Guzman del Proo S.A. (eds), *Abalone of the world. Biology, fisheries and culture*. Oxford: Fishing News Books. pp 438-447.

Tarr R.J.Q. (1995) Growth and movement of the South African abalone *Haliotis midae*: A reassessment. *Marine and Freshwater Research* 46, 583-590.

Tarr R.J.Q. (2000) The South African abalone (*Haliotis midae*) fishery: A decade of challenges and change. *Canadian special publication of Fisheries and Aquatic Sciences* 130, 32-40.

Templeton A.R. (2000) Epistasis and complex traits. In: Wolf J.B., Brodie E.D. & Wade M.J. (eds), *Epistasis and the Evolutionary Process*. New York: Oxford University Press. pp 41-57.

Tenesa A., Knott S.A., Ward D., Smith D., Williams J.L. & Visscher P.M. (2003a) Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. *Journal of Animal Science* 81, 617-623.

Tenesa A., Knott S.A., Carothers A.D. & Visscher P.M. (2003b) Power of linkage disequilibrium mapping to detect a quantitative trait locus (QTL) in selected samples of unrelated individuals. *Annals of Human Genetics* 67, 557-566.

Tenesa A., Navarro P., Hayes B.J., Duffy D.L., Clarke G.M., Goddard M.E. & Visscher P.M. (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Research* 17, 520-526.

Terwilliger J.D., Zöllner S., Laan M. & Pääbo S. (1998) Mapping Genes through the Use of Linkage Disequilibrium Generated by Genetic Drift: 'Drift Mapping' in Small Populations with No Demographic Expansion. *Human Heredity* 48, 138-154.

Thévenon S., Dayo G.K., Sylla S., Sidibe I., Berthier D., Legros H., Boichard D., Eggen A. & Gautier M. (2007) The extent of linkage disequilibrium in a large cattle population of western Africa and its consequences for association studies. *Animal Genetics* 38, 277-286.

Thomson G. & Baur M.P. (1984) Third order linkage disequilibrium. *Tissue Antigens* 24, 250-255.

Thomson G. & Single R.M. (2014) Conditional asymmetric linkage disequilibrium (ALD): Extending the biallelic r^2 measure. *Genetics* 198, 321-331.

Tong A.H.Y., Lesage G., Bader G.D., Ding H., Xu H., Xin X., Young J., Berriz G.F., Brost R.L., Chang M., Chen Y., Cheng X., Chua G., Friesen H., Goldberg D.S., Haynes J., Humphries C., He G., Hussein S., Ke L., Krogan N., Li Z., Levinson J.N., Lu H., Ménard P., Munyana C., Parsons A.B., Ryan O., Tonikian R., Roberts T., Sdicu A.-M., Shapiro J., Sheikh B., Suter B., Wong S.L., Zhang L.V., Zhu H., Burd C.G., Munro S., Sander C., Rine J., Greenblatt J., Peter M., Bretscher A., Bell G., Roth F.P., Brown G.W., Andrews B., Bussey H. & Boone C. (2004) Global mapping of the yeast genetic interaction network. *Science* 303, 808-813.

Tschermak E. (1900) Über künstliche Kreuzung bei *Pisum sativum*. *Berichte der Deutschen Botanischen Gesellschaft* 18, 232-239.

Uimari P. & Tapio M. (2011) Extent of linkage disequilibrium and effective population size in Finnish Landrace and Finnish Yorkshire pig breeds. *Journal of Animal Science* 89, 609-614.

Uleberg E., Widerøe I.S., Grindflek E., Szyda J., Lien S. & Meuwissen T.H.E. (2005) Fine mapping of a QTL for intramuscular fat on porcine chromosome 6 using combined linkage and linkage disequilibrium mapping. *Journal of Animal Breeding and Genetics* 122, 1-6.

Van der Merwe M. & Roodt-Wilding R. (2008) Chromosome number of the South African abalone *Haliotis midae*. *African Journal of Marine Science* 30, 195-198.

Vervalle J., Hepple J., Jansen S., du Plessis J., Wang P., Rhode C. & Roodt-Wilding R. (2013) Integrated linkage map of *Haliotis midae* Linnaeus based on microsatellite and SNP markers. *Journal of Shellfish Research* 32, 89-103.

Volis S., Shulgina I., Ward D. & Mendlinger S. (2003) Regional subdivision in wild barley allozyme variation: Adaptive or Neutral? *Journal of Heredity* 94, 341-351.

von Tschermak E. (1900) Über künstliche Kreuzung bei *Pisum sativum*. *Habilitationsschrift Zeitschrift für das landwirtschaftliche Versuchswesen in Österreich* 3, 465-555.

Wahlund S. (1928) Zusammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas* 11, 65-106.

Wakeley J. & Lessard S. (2003) Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. *Genetics* 164, 1043-1053.

Waples R.S. (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics* 7, 167-184.

Waples R.S. & Do C. (2008) LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources* 8, 753-756.

Waples R.S. & Do C. (2010) Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications* 3, 244-262.

Waples R.S. & England P.R. (2011) Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. *Genetics* 189, 633-644.

Weir B.S. (1996) *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Associates, Sunderland, MA.

Weir B.S. & Hill W.G. (1980) Effect of mating structure on variation in linkage disequilibrium. *Genetics* 95, 477-488.

Weiss K.M. & Clark A.G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics* 18, 19-24.

Whittam T.S., Ochman H. & Selander R.K. (1983) Geographic components of linkage disequilibrium in natural populations of *Escherichia coli*. *Molecular Biology and Evolution* 1, 67-83.

Williams R.W., Gu J., Qi S. & Lu L. (2001) The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. *Genome Biology* 2, 1-18.

Wollstein A. & Stephan W. (2015) Inferring positive selection in humans from genomic data. *Investigative Genetics* 6, 5.

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

Yamazaki T. (1977) The effects of overdominance on linkage in a multilocus system. *Genetics* 86, 227-236.

Yu S.B., Xu W.J. Vijayakumar C.H.M., Ali J., Fu B.Y., Xu J.L. Jiang Y.Z., Marghirang R., Domingo J., Aquino C., Virmani S.S. & Li Z.K. (2003) Molecular diversity and multilocus organisation of the parental lines used in the International Rice Molecular Breeding Program. *Theoretical and Applied Genetics* 108, 131-140.

Yue G.H. (2014) Recent advances of genome mapping and marker-assisted selection in aquaculture. *Fish and Fisheries* 15, 376-396.

Zhao H., Nettleton D., Soller M. & Dekkers J.C.M. (2005) Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genetical Research* 86, 77-87.

Chapter 2

Genome-wide Linkage Disequilibrium in Abalone

Abstract

The phenomenon of linkage disequilibrium is central to a wide variety of demographic and evolutionary processes. As such, LD data can be applied, and is even required, for numerous theoretical and practical applications within genetics. The aim of this chapter was therefore to characterise LD in two wild and two cultured cohorts of the South African abalone and to investigate how their respective population histories have shaped LD across the genome and on a locus-specific scale. One hundred and twelve mapped microsatellite markers were used to genotype the four cohorts. Levels of genetic diversity across the cohorts were similarly high, although significant population structure was evident. The decay of LD with genetic distance was present in both cultured cohorts, but almost absent within the wild cohorts. Levels of LD were high as compared with other aquaculture species, which could have been an artefact of microsatellite markers and limited sample size, but could also be explained by various population demographic events related to the over-exploitation of the species in the wild, domestication, and the reproductive strategy of abalone. When syntenic LD was visualised, a number of interesting patterns were observed surrounding candidate loci under selection, including possible hitch-hiking events and epistatic interactions between loci. In terms of the potential for association studies, the persistence of significant LD over relatively far distances within these cohorts suggested that fewer markers would be sufficient to ensure marker-QTL co-segregation, although the use of generationally discrete cohorts is advised.

2.1) Introduction

Linkage disequilibrium (LD) forms a central topic within the field of population genetics; while LD across the genome reflects the impact of demographic events, such as population bottlenecks and admixture, LD across single chromosomes or smaller genomic regions can reveal the effects of locus-specific factors, such as selection, recombination, and mutation. In particular, the tendency of selection to produce distinctive signatures, or patterns, of LD can provide important clues in the search for regions/genes that are involved in adaptation *via* natural selection, or that are being targeted by artificial selection (Qanbari & Simianer 2014). These candidate regions can then be further investigated using the LD mapping approach to find genotype-phenotype correlations, which relies on the co-segregation of mapped markers with the quantitative trait loci (QTLs) of interest to reveal the latter's location (Stinchcombe & Hoekstra 2007; Goddard & Hayes 2009).

In this regard, the nature of LD across the genome is also of great importance, as the success of a mapping study is largely dependent on taking into account the levels of LD within the target population (Gaut & Long 2003; Balding 2006; Goddard & Hayes 2009). For example, as the LD mapping approach is based on the assumption that LD decays with genetic distance, it is important to first confirm that the target population adheres to this expectation. Populations in which LD is no longer decaying with distance would tend to produce more false positive associations, as LD between a marker and QTL would no longer represent close linkage. Furthermore, in order to maximise the potential success of a mapping study, it is necessary to determine the level of marker saturation required to ensure that the majority of QTLs co-segregate with at least one marker. In situations where LD decays rapidly overall, blocks of linkage across the genome would tend to be relatively small (under 5 cM), and a large number of markers would be required in order to successfully detect most QTL through marker-QTL co-segregation. However, if LD decays slowly, causing blocks of linkage to be much larger (tens of cM), a relatively low marker density would be sufficient to capture most associations (Meadows *et al.* 2008; Goddard & Hayes 2009).

Given its numerous theoretical and practical applications, efforts to characterise the magnitude and extent of genome-wide LD within species of interest have become widespread, with numerous examples in both terrestrial livestock (Tenesa *et al.* 2003; Heifetz *et al.* 2005; Harmegnies *et al.* 2006; Meadows *et al.* 2008) and crop (Maccaferri *et*

al. 2005; Stich *et al.* 2005; Malysheva-Otto *et al.* 2006) species. In comparison, LD has been characterised within relatively few aquaculture species, with Atlantic salmon (Moen *et al.* 2008) and rainbow trout (Rexroad & Vallejo 2009) being the first, and the pacific white shrimp (Du *et al.* 2010) and silver-lipped pearl oyster (Jones *et al.* 2013) being added more recently. However, this number is expected to increase rapidly as higher density linkage maps become available for more species. To date, no large-scale LD data has yet been generated for any halotid species, although the most recently generated *Haliotis midae* linkage map (Vervalle *et al.* 2013) is sufficiently saturated for this purpose (average of 5.6 markers per linkage group [min = 2; max = 15] and average marker spacing of 9.21 cM [min = 0.2 cM; max = 40.8 cM]).

The aim of this chapter was, therefore, to quantify the levels of genome-wide LD within two wild and two cultured population cohorts of *H. midae*, for the primary purpose of investigating how their respective population histories have shaped LD on both a genome-wide and locus-specific scale. As this species is currently within the earlier stages of domestication, this represented a unique opportunity to observe the behaviour of LD under the influence of the various demographic and evolutionary factors associated with such an event. In addition, the data generated was also used to provide a preliminary assessment of the feasibility of association studies of complex traits within this species by critically evaluating the level and extent of LD within the study cohorts.

2.2) Materials and Methods

2.2.1) Study populations

A total of 128 *Haliotis midae* individuals were used for this study, comprising two wild and two cultured population cohorts of 32 animals each. Cultured cohorts were obtained from two aquaculture facilities located on the west (Atlantic Sea Farm, ASF) and east (Wild Coast Abalone, WCA) coasts of South Africa, respectively. Both cohorts were comprised of F_1 individuals (three years of age) produced by random mating of wild broodstock under semi-natural conditions, and reared communally according to standard commercial practice. For comparative purposes, wild cohorts were obtained from the respective progenitor populations of the cultured stocks, *i.e.* Saldana Bay (SAL) for the west coast population, and Riet Point (RP) for the east coast population. Animals were sampled at random from the respective populations, but were restricted to the minimum legal catch size (*i.e.* only sexually mature adults). Muscle and/or gill tissue was collected from each

individual and stored in 70% ethanol at -20°C. DNA was extracted using the standard CTAB protocol of Saghai-Maroof *et al.* (1984).

2.2.2) Markers and genotyping

The current study included 112 of the 116 microsatellite markers from the *Haliotis midae* integrated linkage map (Vervalle *et al.* 2013), which span 17 of the 18 linkage groups (no microsatellite markers were mapped for linkage group 11) (Figure S2.1; Table S2.1). Markers were used to genotype all four abalone cohorts. For amplification *via* polymerase chain reaction (PCR), markers were divided into multiplexes according to primer annealing temperature, product length and fluorescent dye colour, and amplified using the Qiagen® Multiplex PCR kit according to the manufacturer's specifications. PCR products were verified on 1.5% agarose gel electrophoresis and separated *via* capillary electrophoresis. Alleles were scored using Peak Scanner^{TM} v1.0 (Life Technologies) and binned using AutoBin v0.9 (Salin 2010). Markers that could not be scored reliably were excluded from the data sets. Null allele frequencies were estimated using the maximum likelihood method in GenePop v4.2 (Rousset 2008). Microsatellite Toolkit (Park 2001) and PGDSpider v2.0.8.2 (Lischer & Excoffier 2012) were used to convert the data to the necessary input file formats.

2.2.3) Analysis of genetic diversity and population differentiation

Basic genetic diversity statistics were calculated per locus and cohort (ASF, SAL, WCA and RP) using GenAlEx v6.501 (Peakall & Smouse 2006, 2012). These included: number of alleles per marker (A_n), effective number of alleles per marker (A_e), allele frequencies, heterozygosities (H_o and unbiased H_e), Shannon's information index (I), and per locus F_{IS}. A Kruskal-Wallis (KW; $P < 0.05$) test was used to test for significant differences in number of alleles, effective number of alleles, Shannon's information index and unbiased heterozygosity between each cohort in XLStatistics v12.11.22 (Carr 2012). Deviation from Hardy-Weinberg (HW) equilibrium (exact probability test, 10000 dememorisation, 20 batches, 5000 iterations per batch) was calculated for each locus within each cohort in GenePop. Effective population size (N_e) was estimated for each cohort using the heterozygote excess method, temporal method, and LD method (random mating, lowest allowed frequency: 0.02), in NeEstimator v2.01 (Do *et al.* 2013). For N_e calculations, a subset of unlinked markers was chosen based on adherence to HW equilibrium and level of diversity (Table S2.2). Estimates using the temporal method could only be calculated for

the cultured cohorts, as the wild cohorts were used as the ancestral generation for the respective cultured cohorts, and no ancestral samples were available for the wild cohorts. Mean relatedness (*r*) was also calculated for each cohort using the method of Ritland & Lynch 1999 (999 permutations, standardised for max = 1) in GenAlEx. The Ewens-Watterson (EW) test for neutrality *via* Slatkin's exact test (1994) (10000 replicates), and an F_{ST}-outlier test (hierarchical island model, 10000 simulations, 100 demes, 10 groups, significance: $P < 0.05$), were used to identify potential loci under selection, performed in PyPop v0.7.0 (Lancaster *et al.* 2007) and Arlequin v3.5.1.3 (Excoffier & Lischer, 2010), respectively. The Ewens-Watterson test was conducted for each cohort separately and interpreted as a two-tailed test against the alternative hypotheses of balancing ($P < 0.025$) or directional ($P > 0.975$) selection (significant at the 5% level). For the F_{ST}-outlier test, cohorts were grouped according to region (*i.e.* East *versus* West) using the Arlequin Structure Editor to facilitate comparisons between wild and cultured cohorts within the respective groups (F_{SC}), as well as between all cohorts (F_{ST}). To assess population differentiation and the partitioning of genetic variation between cohorts, pairwise F_{ST} values (1000 permutations, significance: $P < 0.05$) were calculated and a locus-by-locus Analysis of Molecular Variance (AMOVA; 10000 permutations, significance: $P < 0.05$) performed in Arlequin, as well as a principal coordinate analysis (PCoA) in GenAlEx. For the AMOVA, cohorts were again grouped according to region, and the analysis was done across all markers, as well as for each linkage group separately.

2.2.4) Analysis of linkage disequilibrium

For estimating LD, two multi-allelic measures, D' (Equation 2.1) and $\chi^{2\prime}$ (Equation 2.2), were calculated for each pair of syntenic markers using PyPoP v0.7.0 (1000 permutations, significance: $P < 0.05$). D' was included primarily to preserve comparability with other studies that used this statistic.

$$D' = \sum_{i=1}^{k} \sum_{j=1}^{m} p(A_i)p(B_j)\left|D'_{ij}\right| \tag{2.1}$$

$$\chi^{2\prime} = \frac{\chi^2}{2N(l-1)} \tag{2.2}$$

Linkage disequilibrium between a subset of non-syntenic markers (Table S2.3) was also calculated for each cohort, from which two empirical baseline levels of significant LD were

determined, as in Meadows *et al.* (2008). The first was calculated as the average LD value over all non-syntenic marker pairs and served as the lower baseline level. To account for any inadequacies in the lower baseline to filter out non-significant associations between loci, and to investigate whether the most significant LD behaved differently, a second more stringent baseline was also calculated using an approach similar to that of Heifetz *et al.* (2005): All non-syntenic pairwise values were first ranked according to *P*-value, and then the average taken over all $\chi^{2\prime}$ values with $P < 0.05$. To visualise the decay of LD with genetic distance, syntenic LD (D' and $\chi^{2\prime}$) was plotted against genetic distance (cM) for each cohort in Microsoft Excel (2013). This was done for all pairwise comparisons (D' and $\chi^{2\prime}$), as well as for only those values that exceeded the respective baseline levels ($\chi^{2\prime}$ only) (*i.e.* three subsets of LD values). The decay of LD in each cohort and for each subset of values was quantified by fitting the model for LD decay (Equation 2.3) to the observed LD values ($\chi^{2\prime}$) and calculating the LD decay coefficient, b_j. Equation parameters were optimised by minimising the sum of squared differences (SSD) between the observed and model values for LD using the Solver add-in in Microsoft Excel.

$$LD_{ij} = (1 + 4b_j d_{ij})^{-1} + e_{ij} \qquad (2.3)$$

Finally, to investigate locus-specific patterns of LD, significant associations between syntenic marker pairs were visualised on heat map tables. As any particular patterns of LD would be easiest to identify within the context of a higher marker density, only those linkage groups with at least 7 – 8 markers were examined in this way, and preference was given to those that contributed the most to the variation between wild and cultured cohorts within groups, as determined by the AMOVA. Only LD values greater than or equal to the 5% baseline level were considered significant, although values that only exceeded the lower baseline level were also included to give a more comprehensive view of possible association “climates”.

2.3) Results

2.3.1) Markers

Three of the markers (*HmS104*, *H.rub13F06* and *HmidILL1.46687*) located on different linkage groups (LG2, -4 and -9, respectively) failed to amplify. One marker, *HmLCS37*, was excluded due to excessive missing data in the ASF and WCA cohorts. As a result, a total of 109 markers spread across 17 of the 18 linkage groups were available for further analyses (108 for the two cultured cohorts). The number of markers per linkage group

ranged from 2 – 15, with an average of 5.6 markers per linkage group and an average marker spacing of 18.74 cM. The majority of markers were in HW equilibrium within the respective cohorts ($P > 0.05$), with a slightly higher percentage within the wild cohorts (Table S2.4). Non-zero null allele frequencies (lower bound >0) were present within all cohorts for 32-38% of markers (Table S2.5).

2.3.2) Genetic diversity and population structure

Based on all estimates, genetic diversity was found to be moderate to high across all population cohorts (Figure 2.1, Table S2.4), with levels being consistent with those previously reported for these populations (Rhode *et al*. 2013). Between cohorts, all estimates were highly comparable, with cultured cohorts being only slightly lower, but not significantly so (KW test: $P > 0.05$). Despite this, significant differentiation was observed between all cohorts based on estimates of pairwise F_{ST}, with values ranging from 0.0077 ($P = 0.0000$) between the wild cohorts, to 0.0327 ($P = 0.0000$) between the two cultured cohorts. This level of differentiation was also reflected in the clustering of the cohorts primarily within separate quadrants of the PCoA plot (Figure 2.2), although the high level of within population variance (97.81%, AMOVA across all markers) was still evident from the significant dispersal of individuals within the clusters. The AMOVAs run for each linkage group separately indicated that certain linkage groups contributed significantly more to the variation among cohorts within groups (*i.e.* between the respective wild and cultured cohorts), with percentages ranging from 0.12% for LG4 to 7.81% for LG9 (Table 2.1). Many of the same linkage groups were also highlighted by the EW test for neutrality and F_{ST}-outlier test as harbouring possible loci under selection (Table S2.6).

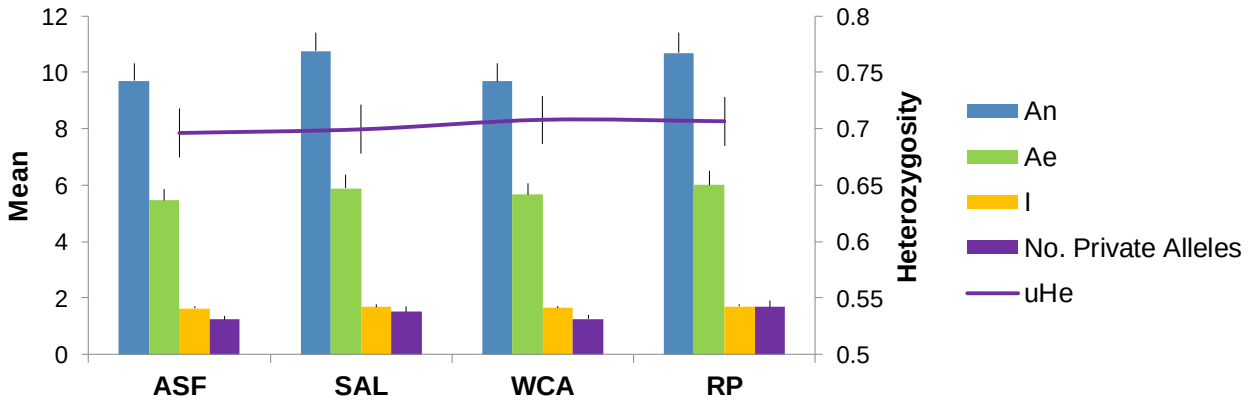


Figure 2.1: Summary of genetic diversity statistics across the four cohorts. These include mean number of alleles (A_n), mean number of effective alleles (A_e), mean for Shannon's Information Index (I), mean number of private alleles and mean unbiased expected heterozygosities (uH_e). Error bars indicate standard error.

Mean relatedness was significantly different between all cohorts; values for both wild cohorts were effectively zero, while a low level of relatedness was present within both cultured cohorts (Figure 2.3). Estimates of effective population size varied greatly between cohorts, but also between methods (Table 2.2). Point estimates for the heterozygote excess method were the highest (∞ for all cohorts), while estimates using the temporal method were substantially lower. Estimates using the LD method were intermediate between the two, with both west coast cohorts having significantly higher estimates than either of the east coast cohorts.

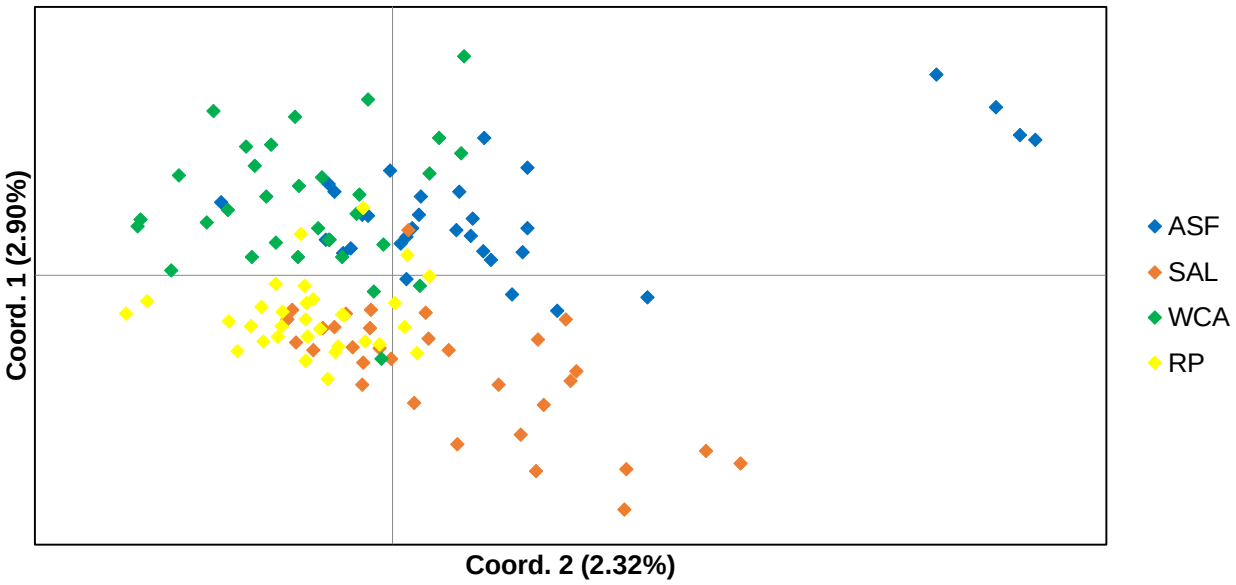


Figure 2.2: Principal coordinate analysis (PCoA) of the four cohorts using the first and second coordinates.

2.3.3) Linkage disequilibrium analyses

As expected, estimates of LD based on D' were consistently higher than those based on $\chi^{2\prime}$ across all cohorts, with average D' levels being inflated by at least 73% in all cases (Figure 2.4A – D). However, the overall trends in the extent of LD (*i.e.* the shapes of the trend line graphs) were comparable between estimates, with only minor differences in the initial slopes of the graphs being observed for the two wild cohorts. The points at which LD became non-significant (*i.e.* where the trend lines intersected the lower baseline levels) differed quite substantially, however, although no clear trends could be identified. As in other studies (*e.g.* Heifetz *et al.* 2005; Zhao *et al.* 2005; Meadows *et al.* 2008), the overall conclusion was that $\chi^{2\prime}$ represents a more reliable estimate of LD, and all subsequent analyses were conducted using $\chi^{2\prime}$ only.

Table 2.1: Estimates of F_{ST} (between all populations), F_{SC} (between wild and cultured cohorts within groups), and percentage variance among populations from the global AMOVA across all markers, and for each linkage group separately.

Dataset	F_{ST}	F_{SC}	% var. among pops
All	0.022**	0.023**	2.293
LG1	0.012**	0.011**	1.116
LG2	0.011*	0.007	0.692
LG3	0.003	0.006	0.636
LG4	0.002	0.001	0.123
LG5	0.041**	0.043**	4.317
LG6	0.031**	0.021**	2.035
LG7	0.016*	0.017	1.665
LG8	0.014**	0.014**	1.396
LG9	0.041**	0.075**	7.808
LG10	0.022**	0.007	0.717
LG12	0.008*	0.009*	0.950
LG13	0.008**	0.012**	1.237
LG14	0.019**	0.018**	1.790
LG15	0.014**	0.012*	1.238
LG16	0.014**	0.008*	0.767
LG17	0.005	0.010	1.045
LG18A-D	0.043**	0.050**	5.003

* Significant at the 5% level
** Significant at the 1% level

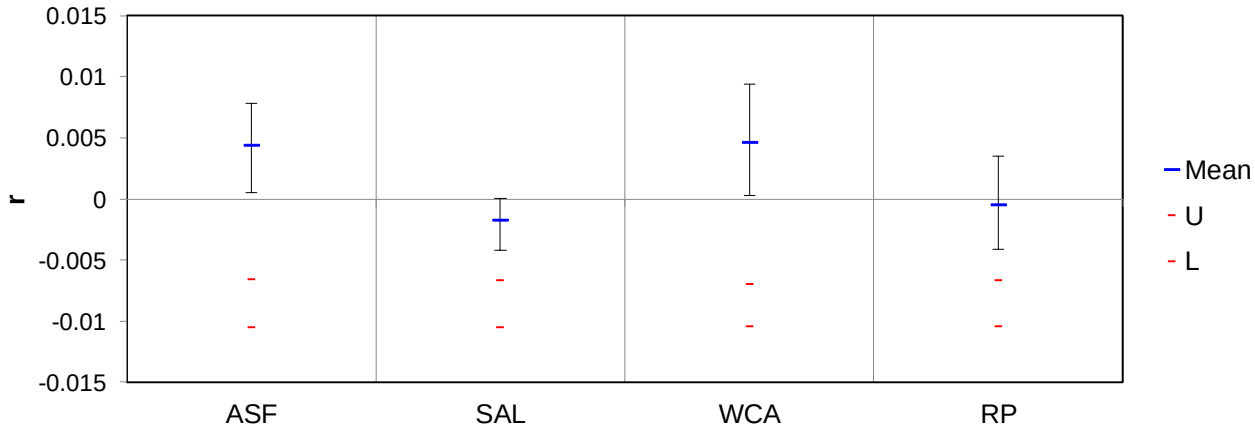


Figure 2.3: Estimates of mean relatedness among the four cohorts. Error bars indicate 95% confidence intervals about the respective means. Upper (U) and lower (L) bounds in red indicate 95% confidence intervals for the null hypothesis of no difference between the cohorts.

In terms of overall magnitude, the strength of non-random associations between syntenic loci was similarly high within all four cohorts of *H. midae* (Figure 2.4A – D). Values of $\chi^{2\prime}$ seldom dropped lower than 0.3 in any cohort (as evidenced by the levelling out of the trend

Table 2.2: Estimates of effective population size (N_e) and 95% confidence intervals (CIs) for the four cohorts, based on the heterozygote excess, LD and temporal methods. Estimates using the temporal method could only be calculated for the two cultured cohorts (ASF and WCA), as the wild cohorts were used as the ancestral samples for the cultured cohorts and no ancestral samples were available for the wild cohorts.

Method	ASF	SAL	WCA	RP
Heterozygote excess	∞	∞	∞	∞
[95% CI]	[28.6 - ∞]	[86.6 - ∞]	[∞]	[∞]
Linkage disequilibrium	393.7	931.5	145.4	145.8
[95% CI]	[87.3 - ∞]	[126.0 - ∞]	[60.7 - ∞]	[72.5 - 1732.6]
Temporal	36.7	-	64.3	-
[95% CI]	[14.9 - 490.0]		[19.2 - ∞]	

lines just below 0.40), with only 26 – 30% of values being ≤0.20, and only 9 – 10% being ≤0.10. Lower baseline levels were comparable between all four cohorts (Table 2.3A), although a substantial disparity appeared when the 5% baseline levels were calculated, with values for both wild cohorts increasing only slightly to around $\chi^{2\prime}$ = 0.37, while the values for the cultured cohorts increased sharply to $\chi^{2\prime}$ = 0.53 and 0.57 for WCA and ASF, respectively (Table 2.3B). As a result, a considerably higher percentage of values were found to still be significant within the wild cohorts after the more stringent baselines were applied (34 – 36%), while values for the cultured cohorts dropped significantly, with 17% still being significant in WCA, and only 7% in ASF (Table 2.3B). Interestingly, despite the higher cut-off point, significant marker-marker associations could still be found over relatively far distances in all cohorts (but particularly within the wild cohorts), although the vast majority (>50%) were observed between loci 0 – 20 cM apart (Figure 2.5A – D).

Regarding the decay of LD with distance as visualised by the logarithmic trend lines fitted to the respective LD decay scatter plots for all pairwise comparisons (Figure 2.4A – D), the most noticeable difference was observed between the wild and cultured cohorts, with the initial slopes being significantly steeper in the cultured cohorts, indicating a more pronounced decay of LD with distance. LD overall decayed most rapidly with distance within the first 2 cM in all cohorts, after which it decayed relatively slowly, reaching the lower baseline at around 40 cM in both cultured cohorts, and not at all in the wild cohorts. When fitting the model for LD decay to the data, it was found that the dataset containing all pairwise comparisons did not fit the model very well (SSD values of 10.3 – 13.1).

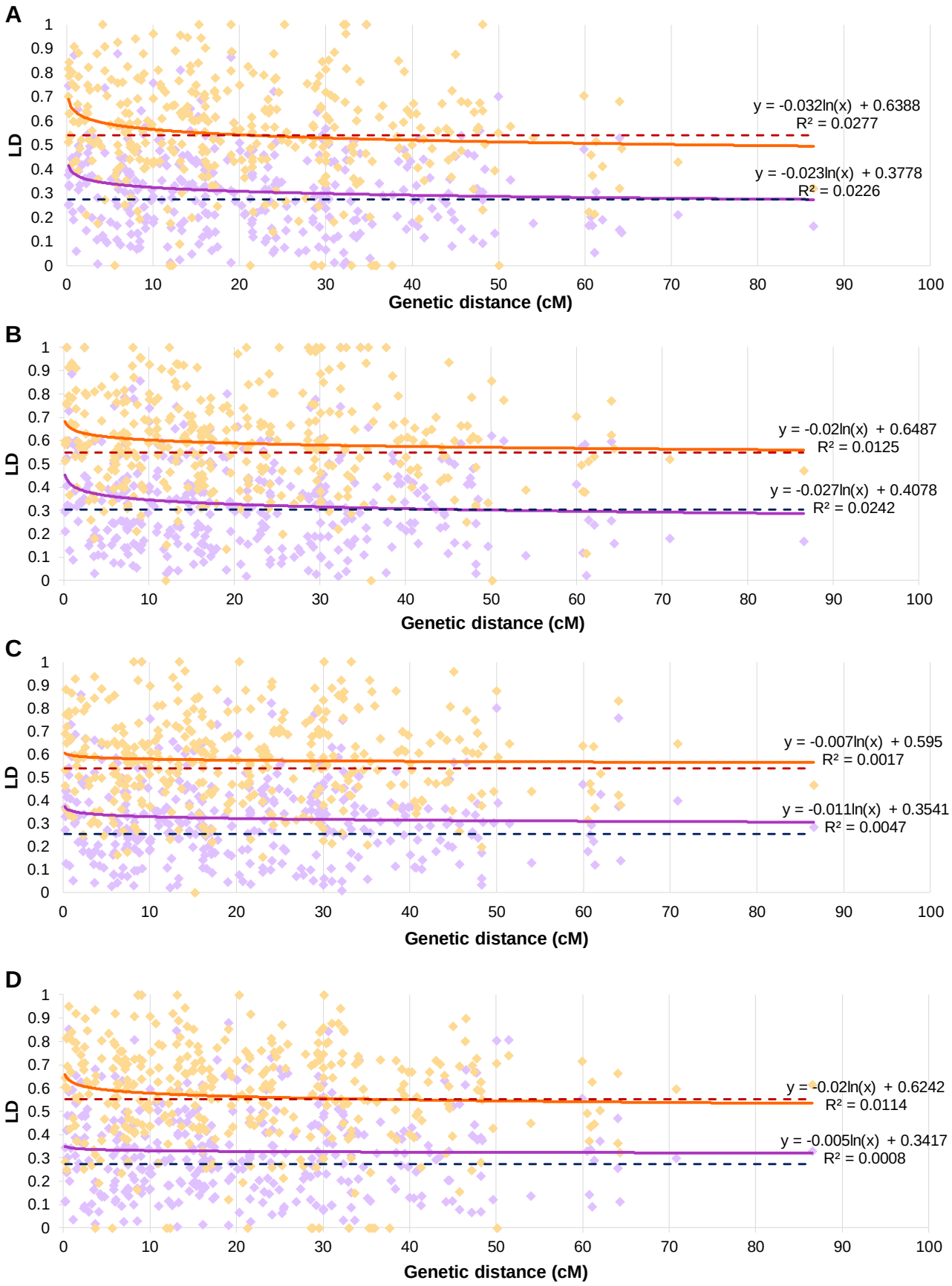


Figure 2.4A – D: Scatter plots with logarithmic trend lines comparing the decay of $\chi^{2\prime}$ (purple) and D' (orange) with genetic distance (cM) within the ASF (**A**), WCA (**B**), SAL (**C**) and RP (**D**) cohorts. Horizontal dashed lines indicate the lower baseline levels for D' and $\chi^{2\prime}$, respectively. Equations for the logarithmic trend lines and associated R^2-values are also displayed.

Table 2.3A – B: Descriptive statistics for the extent and decay of significant LD after applying the lower (**A**) and 5% (**B**) baselines. These include: Baseline (Bsl) values, percentage pairwise comparisons still significant, maximum distances of significant LD, coefficients of LD decay (b_j), and the model sum of squared differences (SSD).

A

Cohort	Lower Bsl	% significant	Max distance (cM)	b_j	Model SSD
ASF	0.275	56%	~65	23.69	3.356
SAL	0.254	64%	>85	269246.39	4.405
WCA	0.306	53%	~65	4.72	4.019
RP	0.274	59%	>85	269246.39	4.432

B

Cohort	5% Bsl	% significant	Max distance (cM)	b_j	Model SSD
ASF	0.568	7%	50	18.92	0.328
SAL	0.373	34%	~70	7312.98	1.850
WCA	0.529	17%	~65	5.14	0.941
RP	0.372	36%	~65	1.94E+07	2.320

However, when the model was fitted to only the significant LD values ($\chi^{2\prime} \geq$ lower and 5% baselines, respectively), the fit noticeably improved (Figure 2.5A – D, Table 2.3A – B), suggesting that significant LD is more inclined to decay with distance within these cohorts, although still not substantially within the wild cohorts. Importantly, while $\chi^{2\prime}$ values were obviously biased upwards when the more stringent baseline was applied, the overall trends in LD did not change appreciably between the two baselines (Figure 2.5A – D, Table 2.3A – B). Significant LD in WCA decayed consistently slower with genetic distance than in ASF (*i.e.* blocks of particularly strong LD were more extensive), and both wild cohorts reached equilibrium levels almost immediately (*i.e.* the strength of associations had little dependence on the distance between loci).

To investigate how these trends might be expressed in terms of locus-specific effects on LD, the magnitude and extent of syntenic LD was examined visually using heat maps (Figures 2.6 – 2.10). For this purpose, linkage groups that contributed the most towards the genetic variation between wild and cultured cohorts (F_{SC}, Table 2.1) were targeted, and limited to those with at least 7 – 8 markers (*i.e.* LG1, -5, -6, -8, and -9). In general, the extent of LD differed significantly between cohorts, as expected from the vastly different LD decay plots. However, in a number of cases, the markers identified as being under selection *via* the Ewens-Watterson and F_{ST}-outlier tests (Table S2.6) were consistently surrounded by small to medium blocks of significant LD (*e.g.* LG1 and -6; Figures 2.6 – 2.7).

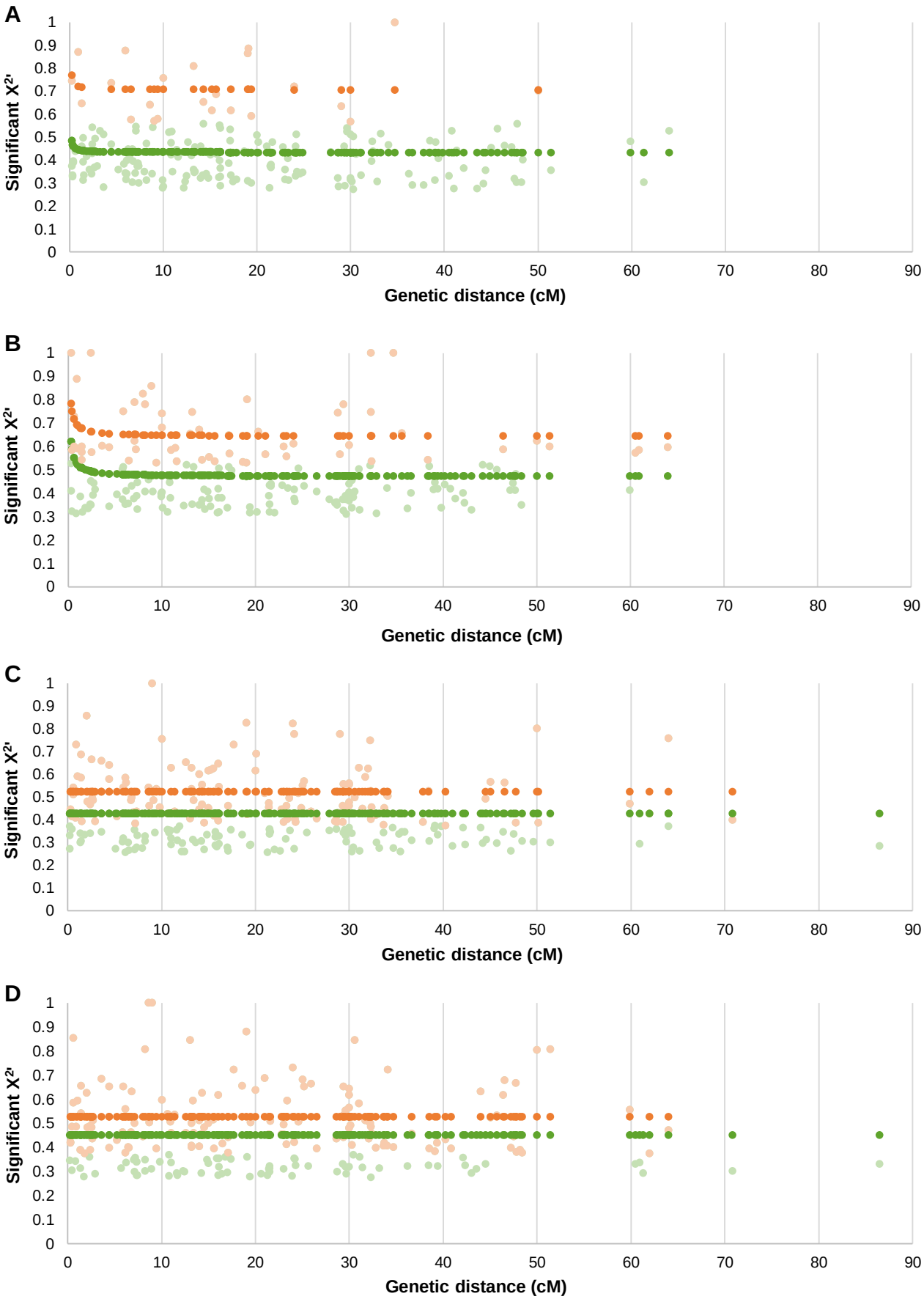


Figure 2.5A – D: Scatter plots showing the decay of significant LD (green: $\chi^{2\prime}$ ≥ lower baseline; orange: $\chi^{2\prime}$ ≥ 5% baseline) with genetic distance (cM) within the ASF (**A**), WCA (**B**), SAL (**C**) and RP (**D**) cohorts. The model for LD decay was fitted to both sets of values; empirical values are shaded lighter, while decay model values are shaded darker.

There was also some variation in the overall levels of LD across the different linkage groups. For example, in comparison with LG1, relatively few significant associations were observed in any of the cohorts between loci on LG9 (Figure 2.8A – D), despite their similar marker densities. Interestingly, patterns of LD on LG8 (Figure 2.9A – D) differed considerably between wild and cultured cohorts, with wild cohorts demonstrating far more extensive LD than either of the cultured cohorts. A similar scenario was observed on LG5 (Figure 2.10A – D), although to a slightly lesser extent.

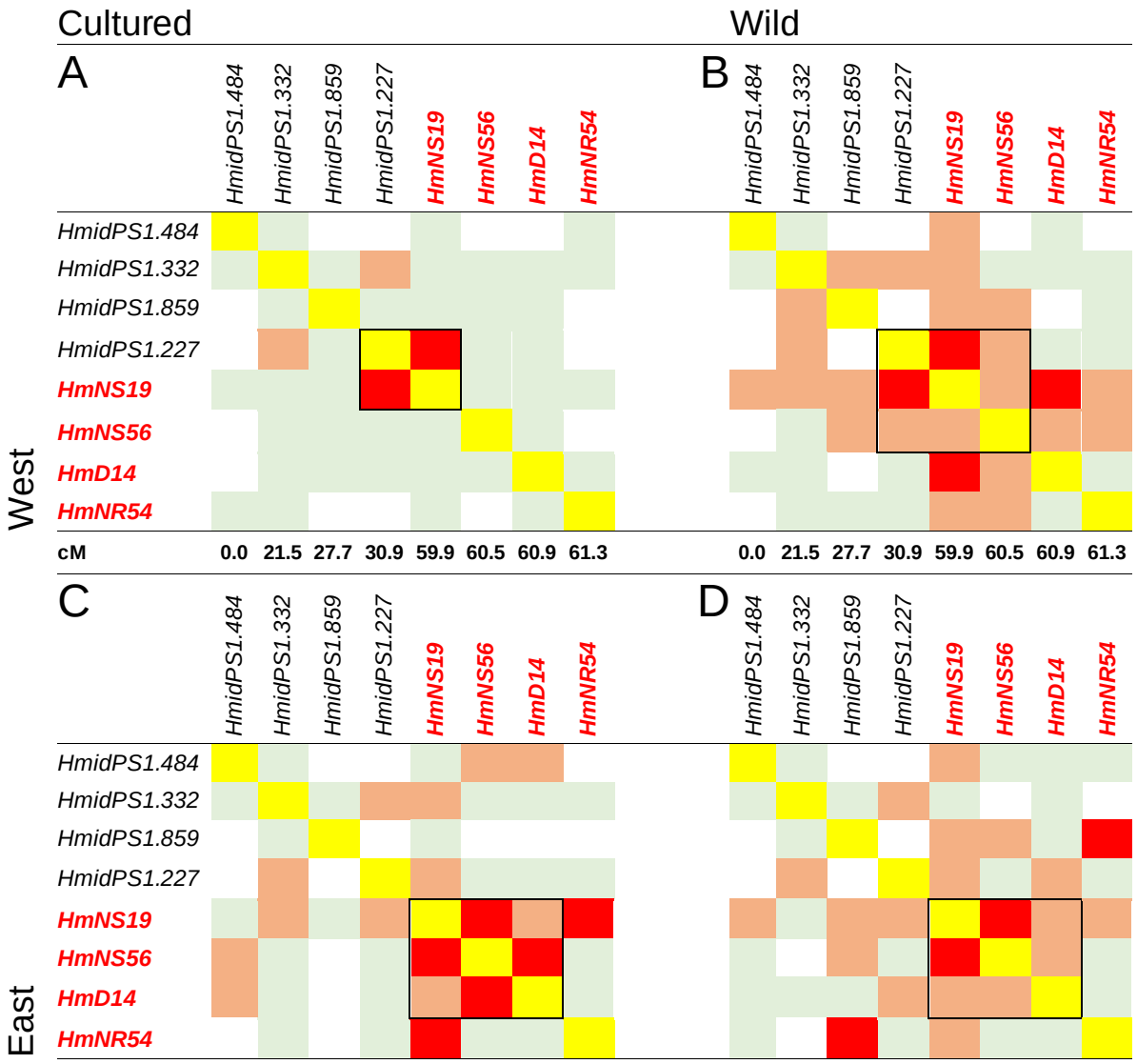


Figure 2.6: Heat map of pairwise comparisons between loci on LG1 within the ASF (**A**), SAL (**B**), WCA (**C**) and RP (**D**) cohorts. Yellow blocks indicate redundant comparisons, blue blocks indicate $\chi^{2\prime}$ ≥ lower baseline, pink blocks indicate $\chi^{2\prime}$ ≥ 5% baseline, and red blocks indicate $\chi^{2\prime}$ ≥ 5% baseline where $P < 0.05$. Cumulative genetic distances (cM) are indicated for each column. Candidate markers under selection are coloured red and blocks of linkage are highlighted with black borders.

2.4) Discussion

2.4.1) Linkage disequilibrium across the Haliotis midae *genome*

Overall, the most prominent feature of LD within the groups of abalone studied here was that levels of both syntenic and non-syntenic LD were particularly high. Although the maximum values for syntenic LD were comparable with those in other studies (*e.g.* Heifetz *et al*. 2005; Rexroad & Vallejo 2009; Du *et al.* 2010), average $\chi^{2\prime}$ values tended to remain well above 0.20 within all cohorts, even at further distances (>50 cM). In contrast, LD in other species tends to reach equilibrium levels closer to 0.10, with many decaying almost to zero. This had a significant impact on how the levels and extent of significant LD within these cohorts were determined, as an r^2 value of 0.2 – 0.25 (or $\chi^{2\prime}$ in the case of a multi-allelic marker) is commonly held as the minimum level of LD expected between two loci that are significantly associated (Meuwissen *et al*. 2001). As such, a number of studies have used this value as the cut-off point in order to investigate the properties of significant LD within a population (*e.g.* Meadows *et al.* 2008; Rexroad & Vallejo 2009). However, given the unusually high level of background LD present within the current study cohorts (lower baseline values of 0.254 – 0.306), the use of such an approach did not yield particularly informative results. As such, rather than using the theoretical minimum value for significant LD, two empirical minimums were calculated for each cohort from levels of non-syntenic LD, which were taken to represent the level expected under the null hypothesis of no LD. Perhaps the closest comparison to the current study is that of Atlantic salmon (Moen *et al.* 2008), where the average $\chi^{2\prime}$ value at distances greater than 50 cM was 0.20, and the empirically established baseline level (equivalent to this study's lower baseline) was 0.16. However, although somewhat less pronounced than in other studies, the expected decay of LD with genetic distance was still evident within both cultured cohorts, particularly when only the most significant associations were examined, although almost absent within the wild cohorts. Significant LD extended as far as 50 cM and ~65 cM within the ASF and WCA cohorts, respectively, the latter of which is somewhat higher than in most other cultured populations of aquaculture species, such as the Atlantic salmon (~50 cM, Moen *et al.* 2008), rainbow trout (40 cM, Rexroad & Vallejo 2009), and silver-lipped pearl oyster (~50 cM, Jones *et al.* 2013).

One important factor that could at least partially account for the elevated levels of LD in this study relative to others is the use of different marker types. In comparing the relative levels of LD between SNP-, microsatellite- and SNP-microsatellite pairs, Moen *et al.*

(2008) observed a marked difference in the overall levels of LD between the different marker types. Although LD observed between microsatellite pairs was only slightly higher than that between SNPs and microsatellites, estimates of LD between pairs of SNPs was distinctly lower than either pairing. This difference was in part attributed to the differences in heterozygosity between the different marker types (Sham *et al.* 2000; Varilo *et al.* 2003), but could also have been influenced by differences in the mutation rates of SNPs and microsatellites. By virtue of their larger size, microsatellite markers tend to accumulate

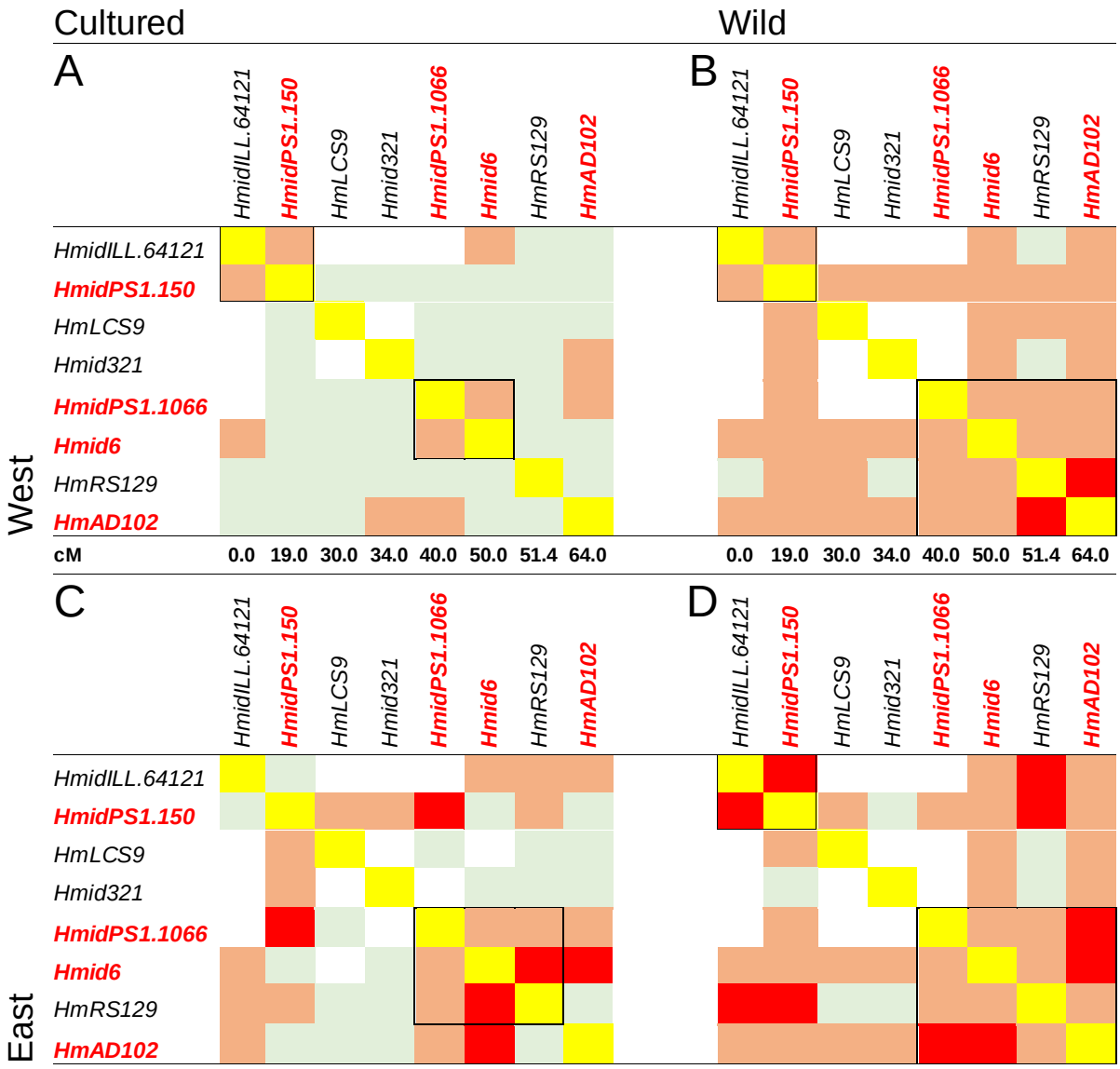


Figure 2.7: Heat map of pairwise comparisons between loci on LG6 within the ASF (**A**), SAL (**B**), WCA (**C**) and RP (**D**) cohorts. Yellow blocks indicate redundant comparisons, blue blocks indicate $\chi^{2\prime}$ ≥ lower baseline, pink blocks indicate $\chi^{2\prime}$ ≥ 5% baseline, and red blocks indicate $\chi^{2\prime}$ ≥ 5% baseline where $P < 0.05$. Cumulative genetic distances (cM) are indicated for each column. Candidate markers under selection are coloured red and blocks of linkage are highlighted with black borders.

mutations at a much faster rate than SNP markers (approximately 10^{-2} – 10^{-6} per locus per generation), which, as discussed in the previous chapter, can greatly increase levels of LD with surrounding loci (Ellegren 2000; Chistiakov *et al.* 2006). As such, direct comparisons between levels of LD in this study and those from studies using SNPs (*e.g.* Du *et al.* 2010; Jones *et al.* 2013) should be interpreted within the correct context. However, this reason alone cannot account for the differences, as a selection of the other studies also used microsatellites (Heifetz *et al.* 2005; Moen *et al.* 2008; Rexroad & Vallejo 2009). Alternatively, Ruzzante (1998) determined that small sample sizes ($n < 50$) can increase the likelihood of sampling errors, particularly when using highly polymorphic markers, such as microsatellites. As such, the inflated estimates of LD in the current study could also have been a factor of the relatively small sample sizes used, particularly as compared with studies that used much larger sample sizes (*e.g.* Du *et al*. 2010; Jones *et al*. 2013).

A further possible explanation for the overall high levels of LD within the cultured cohorts is that LD is still inflated as a result of the recent population bottleneck introduced when a finite number of broodstock individuals were sampled from the wild populations. Both cultured cohorts were comprised of F_1 individuals, and the increase in LD produced by the founder event is unlikely to have dissipated to any great extent after only a single generation. Additionally, the reproductive strategy of abalone could also have contributed to the inflation of over-all levels of LD. Like many marine species, abalones are broadcast spawners that rely on high fecundity to counteract the high mortality rate of offspring during the early stages of development (Hedgecock & Pudovkin 2011). However, as successful fertilisation is also heavily reliant on favourable oceanographic conditions, the number of individuals that reproduce successfully during any given spawning event can be highly variable, referred to as “sweepstakes reproductive success” (Hedgecock & Pudovkin 2011). As such, unequal parental contributions during spawning, particularly under aquaculture conditions, have been widely reported in broadcast spawning species (Lind *et al*. 2009; Slabbert *et al*. 2009; Van den Berg & Roodt-Wilding 2010), and may have contributed to a further reduction in the effective number of breeders during the most recent spawning event. With regards to differences in levels of LD within the study, this phenomenon could also account for the significant disparity observed between levels of LD within the two cultured cohorts, WCA and ASF. If parental contributions were particularly skewed during the spawning event that produced the WCA F_1 population, the further decrease in number of genetic contributors could have resulted in an even more

pronounced upward bias of overall LD within the following generation. This hypothesis is consistent with the estimates and confidence intervals (CIs) for effective population size (or in this case, the effective number of breeders) based on the LD method (Table 2.2), which indicate a significantly lower point estimate and lower bound for WCA (145.4 [60.7 – ∞]) than for ASF (393.7 [87.3 – ∞]). It is unexpected, however, that such an event did not also result in a noticeable decrease in genetic diversity relative to either the wild progenitor cohort (RP), or the other cultured cohort, although the highly polymorphic nature of the study markers may have masked this effect to some extent (Väli *et al*. 2008). Furthermore, the pairwise comparisons of mean relatedness did indicate that levels of relatedness within each cohort were significantly different from each other (*i.e.* CI fell outside the upper and lower bounds for the null hypothesis of no difference) (Figure 2.3).

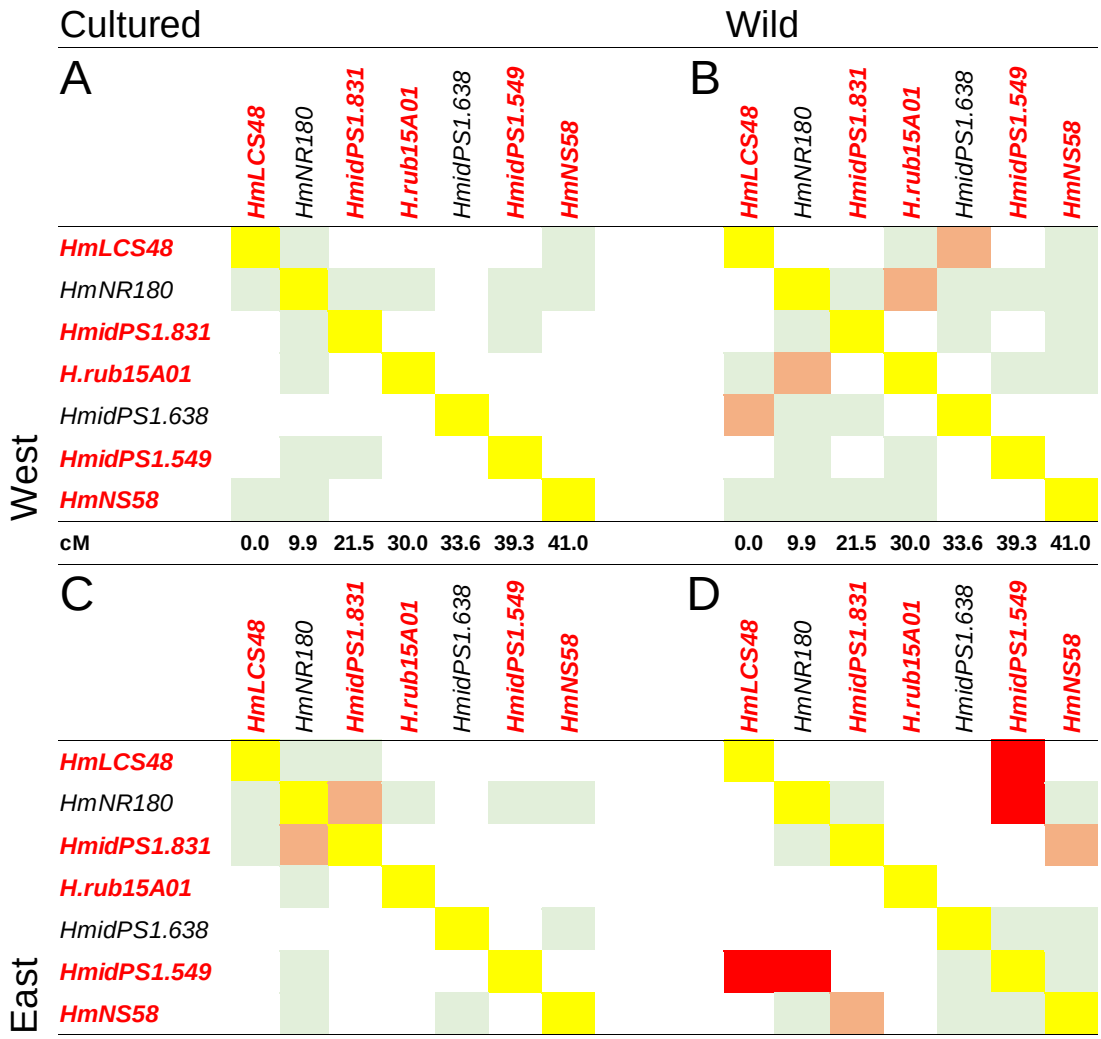


Figure 2.8: Heat map of pairwise comparisons between loci on LG9 within the ASF (**A**), SAL (**B**), WCA (**C**) and RP (**D**) cohorts. Yellow blocks indicate redundant comparisons, blue blocks indicate $\chi^{2\prime}$ ≥ lower baseline, pink blocks indicate $\chi^{2\prime}$ ≥ 5% baseline, and red blocks indicate $\chi^{2\prime}$ ≥ 5% baseline where $P < 0.05$. Cumulative genetic distances (cM) are indicated for each column. Candidate markers under selection are coloured red.

As overall levels of LD were similarly high within all cohorts, when looking at the LD decay plots for all pairwise comparisons, the differences between the wild and cultured cohorts appeared very slight (Figure 2.4A – D). However, two major differences became clear when looking at only the significant LD. Firstly, although the lower baseline levels were fairly comparable (with the possible exception of WCA), the 5% baselines were markedly different, showing a clear divide between the lower wild cohort values and the higher cultured cohort values (Table 2.3A – B). As wild populations are expected to have vastly greater effective population sizes than cultured ones, it is also expected that the levels of background LD would be somewhat lower in wild populations. This explanation was consistent with the estimates of effective population size for SAL (931.5 [126 – ∞]), but much less so for RP (145.8 [72.5 – 1732.6]), which had consistently lower point estimates than SAL and ASF, and was comparable with those of WCA (Table 2.2). As such, it is somewhat surprising that the levels of diversity and background LD for RP were still comparable with those of SAL. Secondly, when significant LD (according to either baseline value) was plotted against genetic distance, the almost complete lack of LD decay with distance became clearly evident in both wild cohorts, with values of b_j in excess of 400-fold higher than those for the cultured cohorts (Table 2.3A – B). The distribution of the most significant associations was also distinctly more dispersed within the wild cohorts; while the range within which at least 50% of significant values fell was approximately 15 cM in the cultured cohorts (max distance = 50 – 65 cM), this distance was closer to 20 cM in both wild cohorts (max distance = 65 – 70 cM), further illustrating the disconnect between the strength of locus associations and distance between loci.

Nevertheless, relative to what might be expected from a cohort representative of a wild population, levels of LD were still unusually high within both wild cohorts. Average $\chi^{2\prime}$ values across consecutive distance ranges (all pairwise comparisons) were often close to comparable with those within the cultured cohorts, with values up to 20 cM being only slightly lower than cultured values, after which the trend reversed. One possible explanation for these elevated values could be drawn from the known recent demographic histories of the source populations from which the wild cohorts were derived. The rampant over-exploitation of wild abalone populations across the South African coastline subsequent to 1994 would have severely reduced the number of breeding individuals within the respective populations, and therefore introduced considerable population bottlenecks (Raemaekers *et al*. 2011). As a result, a significant reduction in genetic diversity would likely have occurred, causing an inflation in both syntenic and non-syntenic

LD. As it is impossible to know how diverse the populations were before the bottlenecks occurred, this hypothesis cannot be confirmed; however, the uncharacteristically low estimates of effective population size for the RP cohort (Table 2.2) suggest that it is a distinct possibility. Alternatively, or perhaps in addition to this, the overall high levels of LD within the wild cohorts could also have been a factor of the sampling method used to collect individuals from the wild populations. Abalones are iteroparous, *i.e.* they have multiple reproductive cycles over the course of their lifetimes, which means that, at any given time, the mature cohort within a natural population would likely include members from at least two or more over-lapping generations (*i.e.* individuals produced by roughly the same parental cohort, but during different spawning events). As such, the restriction of sampling to only mature adults would not have prevented the sampling of individuals from multiple generations. As the occurrence of unequal parental contributions during spawning can cause allele frequencies to fluctuate from one generation to the next (exacerbated by the high fecundity of the species) (Chapman *et al*. 2002; Robainas *et al*. 2005; Lee & Boulding 2009), the presence of over-lapping generations would therefore produce an "admixture-like" effect, thus inflating LD within the sampled cohort (Waples 1991; Robinson & Moyer 2013). To this effect, Rhode (2013) observed significant genetic differentiation between temporal samples spanning approximately 1 – 2 generations taken from both the SAL and RP populations, thus indicating the likelihood of this scenario within these cohorts. Importantly, this also provides a possible explanation for the high levels of LD prevalent within the wild cohorts relative to those within the cultured cohorts, as the cultured individuals are known to represent only a single respective generation.

2.4.2) Locus-specific patterns of linkage disequilibrium

When significant associations between syntenic loci on the more saturated linkage groups was examined visually using heat maps, a number of interesting patterns in both close-range and longer-range LD emerged. Patterns of association were found to be extremely heterogeneous between the different linkage groups, which likely reflect at least in part the expected non-uniformity of recombination rates across the genome. For example, the relative sparsity of significant associations along LG9 (Figure 2.8A – D) would seem to suggest a somewhat higher rate of recombination overall (*i.e.* associations are broken up quickly by recombination), which could at least in part explain the noticeable absence of LD patterns despite the apparent functional significance of this linkage group (*i.e.* the AMOVA results and number of candidate loci under selection). In contrast, LG1, -5, -6 and

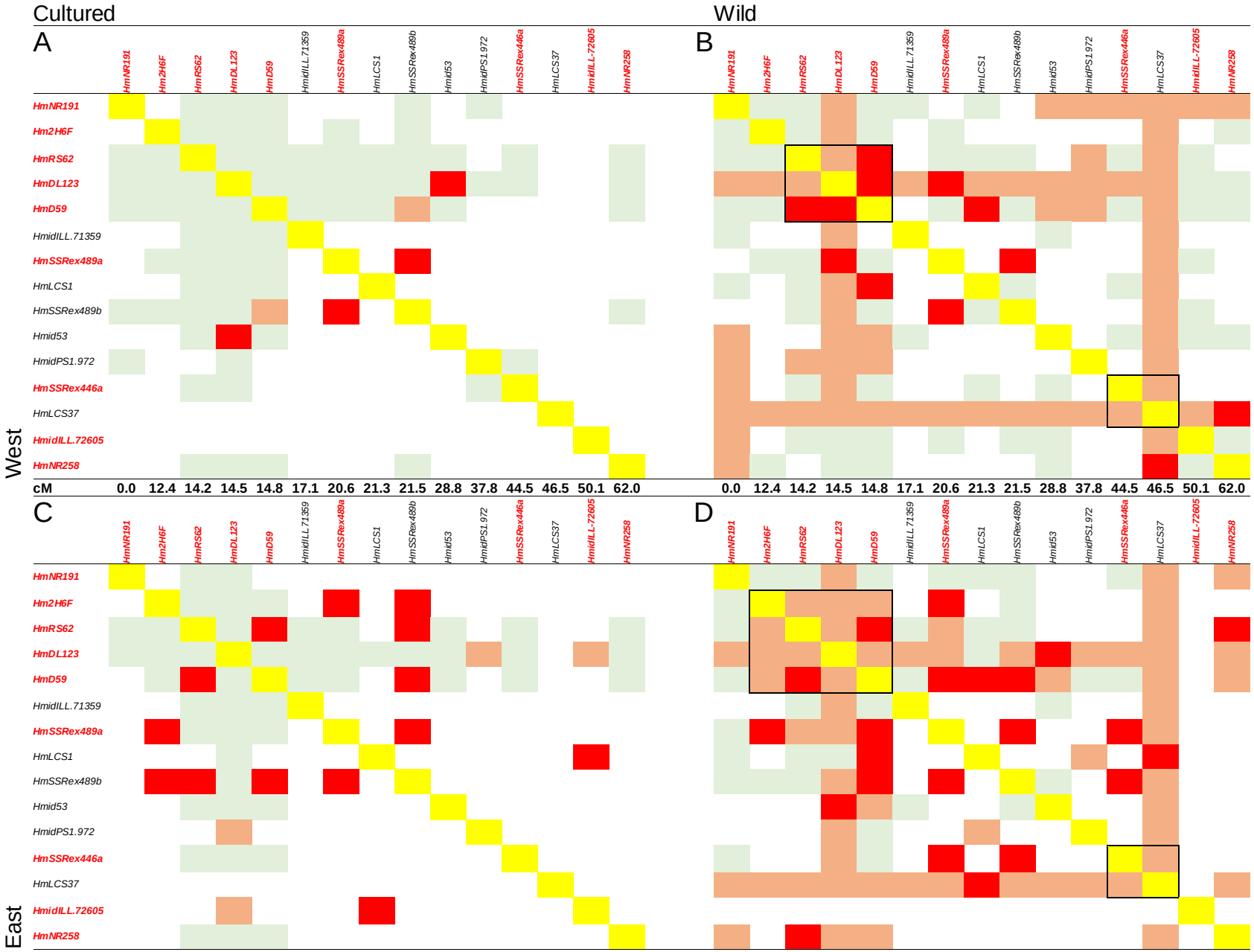


Figure 2.9: Heat map of pairwise comparisons between loci on LG8 within the ASF (**A**), SAL (**B**), WCA (**C**) and RP (**D**) cohorts. Yellow blocks indicate redundant comparisons, blue blocks indicate $\chi^{2\prime} \geq$ lower baseline, pink blocks indicate $\chi^{2\prime} \geq$ 5% baseline, and red blocks indicate $\chi^{2\prime} \geq$ 5% baseline where P < 0.05. Cumulative genetic distances (cM) are indicated for each column. Candidate markers under selection are coloured red and blocks of linkage are highlighted with black borders.

-8, all of which included at least four candidate loci under selection (Table S2.6), contained large numbers of significant marker-marker associations (Figures 2.6, 2.7, 2.9 and 2.10), although LG5 had unexpectedly small blocks of LD given the relatively high marker density, which might also indicate a higher recombination rate than the other linkage groups. Of particular interest, however, was that many of the associations involving the candidate loci were arranged in patterns of LD that were consistent with what might be

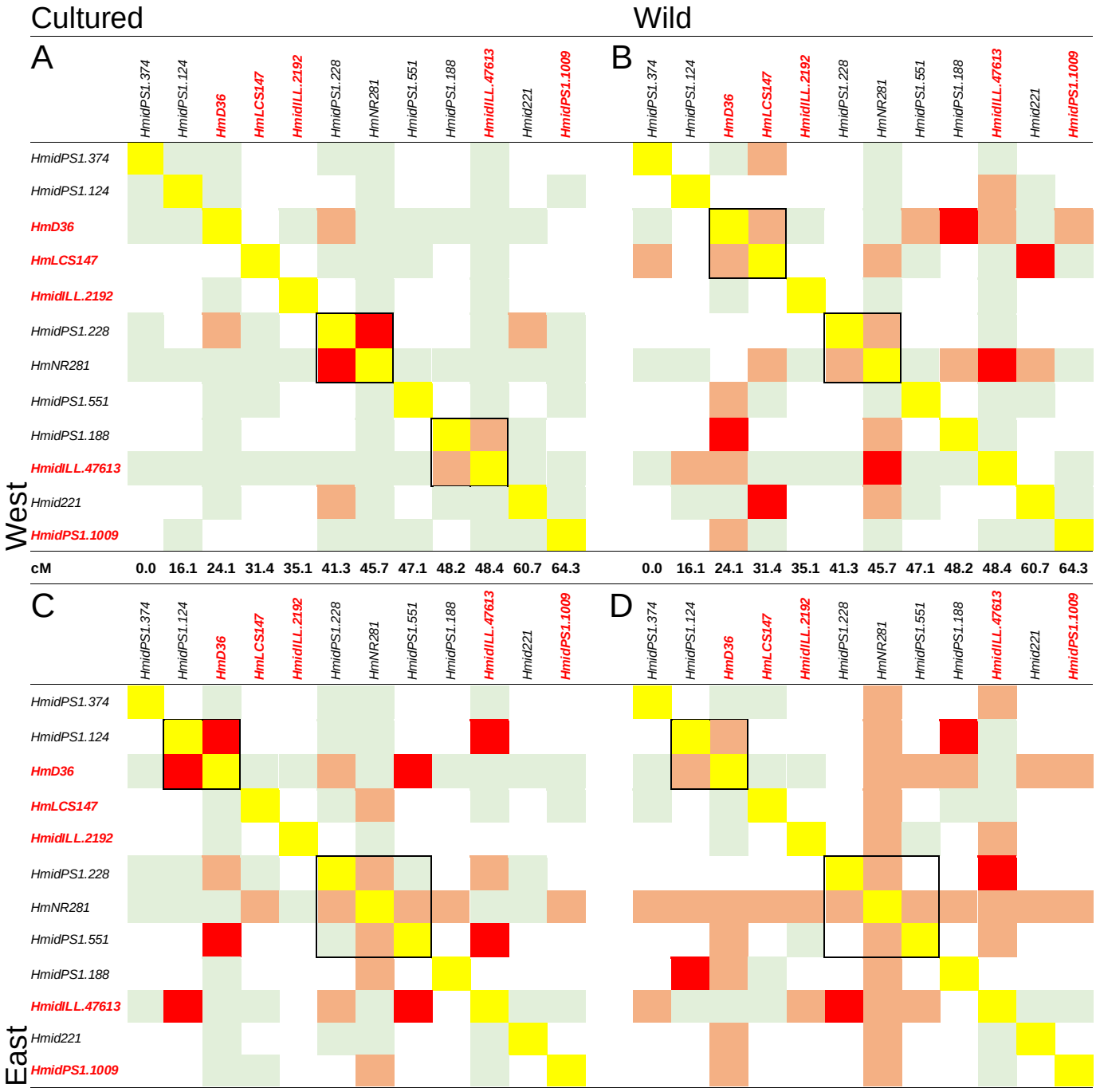


Figure 2.10: Heat map of pairwise comparisons between loci on LG5 within the ASF (**A**), SAL (**B**), WCA (**C**) and RP (**D**) cohorts. Yellow blocks indicate redundant comparisons, blue blocks indicate $\chi^{2\prime}$ ≥ lower baseline, pink blocks indicate $\chi^{2\prime}$ ≥ 5% baseline, and red blocks indicate $\chi^{2\prime}$ ≥ 5% baseline where $P < 0.05$. Cumulative genetic distances (cM) are indicated for each column. Candidate markers under selection are coloured red and blocks of linkage are highlighted with black borders.

expected in the event of locus-specific effects of selection and/or epistatic interactions between loci. For example, candidate markers on LG1, *HmNS19* and *HmNS56*, formed a core association group across three of the four cohorts, with some variation in LD with surrounding markers. Notably, all of the candidate loci were located within fairly close proximity of one other (0.4- and 0.6 cM apart, respectively), creating an arrangement that would tend to facilitate genetic hitch-hiking of surrounding loci in the event that one of the loci is under strong selective constraint. As such, it is possible that only one of the candidate loci is genuinely under selection, while the others only appear to be because of strong linkage with the causal locus (*e.g.* Chan *et al.* 2010). A second example of this could be located on LG8, where three candidate loci (*HmRS62*, *HmDL123* and *HmD59*) are located similarly close together. However, regardless of which locus is responsible, the association pattern on LG1 appears relatively consistent across all cohorts, and could therefore represent selection for a survival trait common to all environments, as opposed to one associated with adaptation to a particular environment. As such, the slight differences in marker composition of the linkage blocks between the east- and west coast cohorts could reflect the regional population structure evident from the PCoA plot (Figure 2.2), while the similarities might reflect a common selective pressure.

In contrast, patterns of association on LG8 appear vastly different between the wild and cultured cohorts, *i.e.* certain patterns that are highly prevalent within both wild cohorts are either totally absent, or almost completely so, within the cultured cohorts. As with LG1, many or most of these patterns are centred around candidate loci for selection (*e.g.* the *HmRS62-HmDL123-HmD59* cluster), and therefore, the rapid (but not yet complete) loss of these patterns within the cultured cohorts could represent a case of relaxed natural selection for one or more wild survival traits. Furthermore, rather than only a single locus (or group of closely linked loci) showing patterns of selection, consistent long-range associations (*i.e.* associations separated by stretches of non-significant LD and/or genetic distance) between otherwise independent blocks of LD or single loci suggest that multiple loci on LG8 might be similarly affected, either epistatically, or because they are functionally related (*i.e.* gene networks) (Rhode *et al*. 2013). Similar locus interaction patterns were also observed on LG5 and 6, suggesting that functionally related genes might not be randomly distributed across the abalone genome. While the clustering of functionally related genes is more often associated with prokaryotes (*i.e.* operons), some evidence for this has also been found in eukaryotic organisms (*e.g.* Boutanaev *et al.* 2002; Lercher *et al.* 2002; Lee & Sonnhammer 2003; Ben-Ari *et al.* 2006; López *et al.* 2010).

Lastly, despite the clear association between distinctive locus-specific patterns of LD and the loci identified as selection candidates that has been demonstrated thus far, unusually persistent patterns of LD were also observed in all four cohorts between supposedly neutral loci on LG5 (*HmidPS1.228*, *HmNR281* and *HmidPS1.551*), which were then also in long-range LD with one or more of the selection candidate loci. One possible explanation is that the three loci are co-segregating only because of close linkage. However, as significant associations on LG5 are relatively dispersed and not particularly common between adjacent loci, it is unlikely that these loci are in LD purely because of their relatively close proximity. Furthermore, while genetic linkage could explain the pattern of local association, it does not provide an adequate explanation for the persistent long-range LD. As such, an alternative hypothesis is that the neutral loci are not responsible for creating the observed patterns of LD, and that one or more unmapped loci located amongst the three mapped markers are functionally related to the other candidate loci. This hypothesis could account for both the long-range and local patterns of LD, although if such were the case, it might have been expected that at least one of the mapped markers would also have been detected as a selection candidate in a similar scenario to that proposed on LG1 and 8. However, one of the markers, *HmNR281*, was among those with relatively high null allele frequencies (>0.15, Table S2.5), which could be interpreted as some degree of the aforementioned effect, as the reduction in heterozygosity at a locus under directional selection can be interpreted as evidence of null alleles (Rhode *et al*. 2013).

2.4.3) Prospects for genotype-phenotype association/LD mapping in Haliotis midae

With regards to the potential for LD mapping/association studies in these cohorts, the prerequisite of the decline of LD with distance was fulfilled within both of the cultured cohorts, although to a much lesser extent within the wild cohorts, particularly when only significant LD was considered. As such, LD mapping is unlikely to be successful within the latter, and efforts should rather be focused towards generationally discrete cohorts (typically housed in aquaculture facilities) if such studies are to be attempted. In terms of the extent of LD within these cohorts, a relatively large proportion of significant associations were observed over considerable distances in both cultured cohorts (prevalent up to 20 cM), which would imply that LD mapping studies could be reasonably successful with a relatively low marker density, as fewer markers would be required to ensure marker-QTL co-segregation (Slate & Pemberton 2007; Li & Merilä 2010). However,

as large blocks of LD would make adding additional markers largely redundant, the extensive LD in these cohorts would also likely preclude efforts to fine-map QTLs once they have been localised to a broader genomic region (Li & Merilä 2010). Given the extent of LD, confidence intervals for identified QTLs might only be marginally better than those of conventional linkage mapping. However, LD mapping still has the advantage of using population data, rather than family data (*i.e.* no mapping families required), and any QTLs identified are likely to be common causative variants (or in LD with them), rather than family-specific ones. As such, it would still be beneficial to pursue the LD mapping approach for complex traits in these cohorts, particularly if the candidate gene/locus approach is used, rather than a genome-wide association study (Stinchcombe & Hoekstra 2007; Goddard & Hayes 2009; Li & Merilä 2010).

2.5) Conclusion

Although levels of LD within abalone were found to be somewhat elevated as compared with other species, particularly with regards to non-syntenic, or background levels of LD, these higher levels could be justified by drawing on various aspects of the respective population histories of the four cohorts and the reproductive biology of abalone. In particular, recent and historical population bottlenecks, as well as the prevalence of unequal parental contributions during spawning, have had marked effects on levels of LD within both wild and cultured abalone cohorts. Furthermore, despite the inflated values, the overall trends in genome-wide LD were consistent with what was expected based on theory and previous observations: LD appeared to decay with genetic distance, and to a much greater extent within cultured cohorts than wild ones. When significant associations were examined on a finer scale (*i.e.* per linkage group), a number of distinctive patterns suggestive of genetic hitch-hiking events were observed surrounding markers identified as candidate loci under selection, providing further support for their possible functional significance. In some cases, these patterns were observed across all cohorts, suggesting a common selective pressure, while in others, they were more specific to a particular group, suggesting differential selective pressures. Several patterns of long-range LD were also observed on some linkage groups, many between selection candidate loci, suggesting possible epistatic interactions or functional relationships. Regarding the prospects for LD mapping/association studies within these cohorts, it was suggested that studies within generationally discrete cohorts, such as the two cultured cohorts, would be the most successful, as the decay of LD with distance was suitably pronounced to facilitate

the mapping of QTLs. Furthermore, significant LD extended over relatively far distances within these cohorts, implying that reasonably few markers would be required to detect most QTLs.

References

Balding D.J. (2006) A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7, 781-791.

Ben-Ari G., Zenvirth D., Sherman A., David L., Klutstein M., Lavi U., Hillel J. & Simchen G. (2006) Four linked genes participate in controlling sporulation efficiency in budding yeast. *PLoS Genetics* 2, e195.

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

Boutanaev A.M., Kalmykova A.I., Shevelyov Y.Y. & Nurminsky D.I. (2002) Large clusters of co-expressed genes in the *Drosophila* genome. *Nature* 420, 666-669.

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

Chan E.K., Rowe H.C., Hansen B.G. & Kliebenstein D.J. (2010) The complex genetic architecture of the metabolome. *PLoS Genetics* 6, e1001198.

Chapman R.W., Ball A.O. & Mash L.R. (2002) Spatial homogeneity and temporal heterogeneity of red drum (*Sciaenops ocellatus*) microsatellites: effective population sizes and management implications. *Marine Biotechnology* 4, 589-603.

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.

Do C., Waples R.S., Peel D., Macbeth G.M., Tillet B.J. & Ovenden J.R. (2013) NeEstimator V2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Molecular Ecology Resources* 14, 209-214.

Du Z.Q., Ciobanu D.C., Onteru S.K., Gorbach D., Mileham A.J., Jaramillo G. & Rothschild M.F. (2010) A gene-based SNP linkage map for pacific white shrimp, *Litopenaeus vannamei*. *Animal Genetics* 41, 286-294.

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

Excoffier L. & Lischer H.E.L. (2010) Arlequin suite version 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* 10, 564-567.

Gaut B.S. & Long A.D. (2003) The lowdown on linkage disequilibrium. *The Plant Cell* 15, 1502-1506.

Goddard M.E. & Hayes B.J. (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics* 10, 381-391.

Harmegnies N., Farnir F., Davin F., Buys N., Georges M. & Coppieters W. (2006) Measuring the extent of linkage disequilibrium in commercial pig populations. *Animal Genetics* 37, 225-231.

Hedgecock D. & Pudovkin A.I. (2011) Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. *Bulletin of Marine Science* 87, 971-1002.

Heifetz E.M., Fulton J.E., O'Sullivan N., Zhao H., Dekkers J.C. & Soller M. (2005) Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. *Genetics* 171, 1173-1181.

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

Jones D.B., Jerry D.R., Khatkar M.S., Raadsma H.W. & Zenger K.R. (2013) A high-density SNP genetic linkage map for the silver-lipped pearl oyster, *Pinctada maxima*: a valuable resource for gene localisation and marker-assisted selection. *BMC Genomics* 14, 810.

Lancaster A.K., Single R.M., Solberg O.D., Nelson M.P. & Thomson G. (2007) PyPop update—a software pipeline for large-scale multilocus population genomics. *Tissue Antigens* 69, 192-197.

Lee H.J.E. & Boulding E.G. (2009) Spatial and temporal population genetic structure of four northeastern Pacific littorinid gastropods: the effect of mode of larval development on variation at one mitochondrial and two nuclear DNA markers. *Molecular Ecology* 18, 2165-2184.

Lee J.M. & Sonnhammer E.L. (2003) Genomic gene clustering analysis of pathways in eukaryotes. *Genome Research* 13, 875-882.

Lercher M.J., Urrutia A.O. & Hurst L.D. (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genetics* 31, 180-183.

Li M.-H. & Merilä J. (2010) Extensive linkage disequilibrium in a wild bird population. *Heredity* 104, 600-610.

Lind C.E., Evans B.S., Knauer J., Taylor J.J.U. & Jerry D.R. (2009) Decreased genetic diversity and a reduced effective population size in cultured silver-lipped pearl oysters (*Pinctada maxima*). *Aquaculture* 286, 12-19.

Lischer H.E.L. & Excoffier L. (2012) PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 28, 298-299.

López M.D., Guerra J.M. & Samuelsson T. (2010) Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLoS ONE* 5, e10654.

Lynch M. & Ritland K. (1999) Estimation of pairwise relatedness with molecular markers. *Genetics* 152, 1753-1766.

Maccaferri M., Sanguineti M.C., Noli E. & Tuberosa R. (2005) Population structure and long-range linkage disequilibrium in a durum wheat elite collection. *Molecular Breeding* 15, 271-289.

Malysheva-Otto L.V., Ganal M.W. & Roder M.S. (2006) Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). *BMC Genetics* 7, 6.

Meadows J.R.S., Chan E.K.F. & Kijas J.W. (2008) Linkage disequilibrium compared between five populations of domestic sheep. *BMC Genetics* 9, 61.

Meuwissen T.H., Hayes B.J. & Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819-1829.

Moen T., Hayes B., Baranski M., Berg P.R., Kjøglum S., Koop B.F., Davidson W.S., Omholt S.W. & Lien S. (2008) A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers. *BMC Genomics* 9, 223.

Park S.D.E. (2001) Trypanotolerance in West African Cattle and the population genetic effects of selection. Unpublished PhD thesis, University of Dublin, Dublin.

Peakall R. & Smouse P.E. (2006) GenAlEx 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6, 288-295.

Peakall R. & Smouse P.E. (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics* 28, 2537-2539.

Qanbari S. & Simianer H. (2014) Mapping signatures of positive selection in the genome of livestock. *Livestock Science* 166, 133-143.

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.

Rexroad C.E. & Vallejo R.L. (2009) Estimates of linkage disequilibrium and effective population size in rainbow trout. *BMC Genetics* 10, 83.

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

Rhode C. (2013) Signatures of selection in natural and cultured abalone (*Haliotis midae*): A population genomics study. Unpublished PhD thesis, Stellenbosch University, Stellenbosch.

Rhode C., Hepple J.A., Jansen S., Davis T., Vervalle J., Bester-van der Merwe A.E. & Roodt-Wilding R. (2012) A population genetic analysis of abalone domestication events in South Africa: Implications for the management of the abalone resource. *Aquaculture* 356, 235-242.

Rhode C., Vervalle J., Bester-van der Merwe A.E. & Roodt-Wilding R. (2013) Detection of molecular signatures of selection at microsatellite loci in the South African abalone (*Haliotis midae*) using a population genomic approach. *Marine Genomics* 10, 27-36.

Robainas B.A., López G.E., Hernández D. & García-Machado E. (2005) Temporal variation of the population structure and genetic diversity of *Farfantepenaeus notialis* assessed by allozyme loci. *Molecular Ecology* 14, 2933-2942.

Robinson J.D. & Moyer G.R. (2013) Linkage disequilibrium and effective population size when generations overlap. *Evolutionary Applications* 6, 290-302.

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

Ruzzante D.E. (1998) A comparison of several measures of genetic distance and population structure with microsatellite data: bias and sampling variance. *Canadian Journal of Fisheries and Aquatic Sciences* 55, 1-14.

Saghai-Maroof M.A., Soliman K.M., Jorgenson R.A. & Allard R.W. (1984) Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proceedings of the National Academy of Sciences of the United States of America* 81, 8014-8018.

Salin F. (2010) AutoBin v0.9. Available at http://www4.bordeaux-aquitaine.inra.fr/biogeco/Ressources/Logiciels/ Autobin [accessed 26 August 2013].

Sham P.C., Zhao J.H. & Curtis D. (2000) The effect of marker characteristics on the power to detect linkage disequilibrium due to single or multiple ancestral mutations. *Annals of Human Genetics* 64, 161-169.

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

Slabbert R., Hepple J.-A., Rhode C., Bester-Van der Merwe A.E. & Roodt-Wilding R. (2012) New microsatellite markers for the abalone *Haliotis midae* developed by 454

pyrosequencing and *in silico* analyses. *Genetics and Molecular Research* 11, 2769-2779.

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

Slate J. & Pemberton J.M. (2007) Admixture and patterns of linkage disequilibrium in a free-living vertebrate population. *Journal of Evolutionary Biology* 20, 1415-1427.

Slatkin M. (1994) An exact test for neutrality based on the Ewens sampling distribution. *Genetical Research* 64, 71-74.

Stich B., Melchinger A.E., Frisch M., Maurer H.P., Heckenberger M. & Reif J.C. (2005) Linkage disequilibrium in European elite maize germplasm investigated with SSRs. *Theoretical and Applied Genetics* 111, 723-730.

Stinchcombe J.R. & Hoekstra H.E. (2007) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* 100, 158-170.

Tenesa A., Knott S.A., Ward D., Smith D., Williams J.L. & Visscher P.M. (2003) Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. *Journal of Animal Science* 81, 617-623.

Väli Ü., Einarsson A., Waits L. & Ellegren H. (2008) To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Molecular Ecology* 17, 3808-3817.

Van den Berg N.-C. (2008) Microsatellite marker development and parentage assignment in *Haliotis midae*. Unpublished MSc thesis, Stellenbosch University, Stellenbosch.

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

Varilo T., Paunio T., Parker A., Perola M., Meyer J., Terwilliger J.D. & Peltonen L. (2003) The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. *Human Molecular Genetics* 12, 51-59.

Vervalle J., Hepple J., Jansen S., du Plessis J., Wang P., Rhode C. & Roodt-Wilding R. (2013) Integrated linkage map of *Haliotis midae* Linnaeus based on microsatellite and SNP markers. *Journal of Shellfish Research* 32, 89-103.

Waples R.S. (1991) Genetic methods for estimating the effective size of cetacean populations. *Report of the International Whaling Commission* 13, 279-300.

Zhao H., Nettleton D., Soller M. & Dekkers J.C.M. (2005) Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genetical Research* 86, 77-87.

Chapter 3

Association Analysis of Candidate Loci under Selection with Size in the South African Abalone

Abstract

Despite domestication still being within the initial stages, significant differentiation has been observed between wild and cultured populations of *Haliotis midae*. Genetic regions associated with the divergence of wild and cultured populations could represent loci determining biologically and economically important traits. Previous studies in this species identified several loci thought to be under divergent selection between wild and cultured populations. The aim of this study was therefore to investigate the influence of artificial selection on genetic variation by determining whether these candidate loci are associated with larger size (primary production trait) in a commercial F_1 population. Thirteen microsatellite markers, putatively identified as being under directional selection, were chosen for association analysis. Various statistical tests were used to detect significant genotype-phenotype associations within a family-bias corrected population cohort, and two family cohorts. Two loci demonstrated significant evidence for association with size, with both loci possessing alleles that correlated significantly with either increased or decreased size. As size is currently the only trait actively selected for in terms of production, the current results suggest that natural selection for adaptation to the novel aquaculture environment is the predominant selective force shaping genetic variation during the initial stages of domestication in abalone. Furthermore, whilst it is currently unclear as to whether these loci represent causative variants for size traits, they may be useful in future molecular-assisted breeding programmes for *H. midae*.

3.1) Introduction

In addition to investigating the processes that shape diversity within biological populations, a further key objective in genetics is the characterisation of the genetic architecture of complex traits of interest. As such, there is increasing interest in the detection of genomic regions containing genes that are, or have previously been, under selective constraint, as such information can provide valuable insights into the genetic mechanisms responsible for the divergence of phenotypes between populations (Feder & Mitchell-Olds 2003; Storz 2005; Gholami *et al*. 2015). In particular, genetic regions that are associated with the divergence of domesticated populations from wild progenitor populations could indicate the presence of loci involved in adaptation *via* natural selection, or that are being targeted by artificial selection, and thus serve as important clues in the search for genes determining biologically and economically important traits (Rubin *et al*. 2010; Rhode *et al*. 2013; Carneiro *et al*. 2015; Evin *et al*. 2015).

Significant genetic differentiation between cultured populations and their wild progenitor populations, as well as between each other, has been reported in a number of commercially important aquaculture species (*e.g.* Skaala *et al*. 2004; Hara & Sekino 2007; Dixon *et al*. 2008; Lind *et al*. 2009), including the South African abalone, *Haliotis midae* (Slabbert *et al*. 2009; Rhode *et al*. 2012, 2014). Despite domestication of this species still being within the initial phases relative to other cultured halotids (Li *et al*. 2004; Praipue *et al*. 2010), significant population differentiation has already been observed between wild and cultured (F_1 and F_2) populations of *H. midae* (Slabbert *et al*. 2009; Rhode *et al*. 2012, 2014). While this rapid divergence has predominantly been attributed to demographic effects associated with the founder event (further exacerbated by the reproductive mode of the species, Slabbert *et al*. 2009; Hedgecock & Pudovkin 2011), it has been suggested that a significant proportion of the differentiation observed between cultured- and wild populations of *H. midae* could be attributed to locus-specific effects of selection (Rhode *et al*. 2012, 2013; Vasemägi *et al*. 2012). During the domestication process, three facets of selection are at play: (i) Relaxed natural selection for traits important for survival in the wild; (ii) Increased natural selection for adaptation to a novel (captive) environment (not human-directed); and (iii) Increased artificial (human-directed) selection on traits of economic importance (Mignon-Grasteau *et al*. 2005). As regions affected by artificial selection could potentially be associated with quantitative trait loci (QTLs) for traits of

economic importance, such as growth rate (*e.g.* Stella *et al*. 2010; Moioli *et al*. 2013; Wilkinson *et al*. 2013), the identification of such regions may be of particular interest.

Using a population genomic approach, Rhode *et al*. (2013) identified a number of microsatellite markers within the South African abalone demonstrating signatures of selection associated with the divergence of cultured F_1 abalone populations from wild progenitor populations. Candidate markers under selection were identified based on the results of three different F_{ST}-outlier detection methods, as well as the Ewens-Watterson homozygosity test for neutrality. Approximately 9% of loci were found to be under directional selection, several of which could be linked to genes, or were derived from expressed sequences, which further supports a possible association with genomic regions responsible for phenotypic expression. As such, it was suggested that the signatures of selection at these loci may represent the effects of a selective sweep facilitated by the domestication event. However, it remains unknown as to whether this selection was caused by artificial selection (crude mass/phenotypic selection for faster growth rate is practiced), or whether natural selection for adaptation to the aquaculture environment is primarily responsible.

The aim of this chapter was, therefore, to determine whether artificial selection, as opposed to natural selection, is the dominant selective force in causing the observed signatures of selection at these loci. For this purpose, a candidate locus LD mapping approach was used to ascertain whether associations exist between allelic variants and size at three years of age (a proxy for growth rate), as size is currently the only trait under artificial selection. The largest and smallest individuals from a commercial F_1 abalone population, produced *via* a single spawning event, were genotyped at thirteen microsatellite loci putatively identified as being under directional selection (Rhode *et al*. 2013). The resulting allele frequencies and corresponding phenotypic data were subsequently analysed to investigate genotype-phenotype associations.

3.2) Materials and Methods

3.2.1) Study population

For this study, a group of 661 F_1 generation individuals was obtained from a single aquaculture facility (I&J Danger Point Abalone Farm). The F_1 animals (three years of age) were produced by random mating of wild broodstock under semi-natural conditions and

were reared communally according to standard commercial practice. All individuals were phenotyped for size-related traits (live weight, shell width and -length) and the top and bottom ±15% (102 large and 98 small individuals) were selected for further analysis and categorised as either large or small based on live weight (Figure 3.1). DNA was extracted from tentacle tissue using a standard CTAB method (Saghai-Maroof *et al*. 1984).

3.2.2) Markers and genotyping

A total of 13 microsatellite markers, putatively shown to be under directional selection (Rhode *et al*. 2013), were used in this study. The markers were divided into three multiplex reactions, and amplified using the Qiagen® Multiplex PCR kit according to the manufacturer's specifications, but using annealing temperatures specific for each multiplex (Table S3.1). Fragment analysis was performed via capillary electrophoresis and alleles were scored using Peak ScannerTM v1.0 (Life Technologies) and binned using AutoBin v0.9 (Salin 2010). Markers that could not be scored reliably, or that were monomorphic, were excluded from the data sets. Null allele frequencies were estimated using the maximum likelihood method in GenePop v4.2 (Rousset 2008). Microsatellite Toolkit (Park 2001) and CREATE (Coombs *et al*. 2008) were used to convert the data to the necessary input file formats.

3.2.3) Genetic data analyses

Initially, pedigree inference was performed in Kingroup v2.0 (Konovalov *et al*. 2004) using full sibship reconstruction (SIMPSON-assisted Descending Ratio algorithm, Konovalov 2006). A total of 23 families were inferred by KinGroup, from which a new data set was constructed by selecting, at random, a maximum of three large and three small individuals from each family, resulting in a family-bias corrected (FBC) cohort of 80 individuals (37 large, 43 small) (Figure 3.1). This was done in order to prevent the skewing of results due to unequal family sizes, as parental contributions in broadcast spawning animals are often highly variable (Slabbert *et al*. 2009). Separate data sets were also created for the two largest families (largest in number), allowing for within-family analyses: Family A (32 individuals; 16 large, 16 small) and Family B (27 individuals; 19 large, 8 small) (Figure 3.1).

Deviation from Hardy-Weinberg (HW) equilibrium (exact probability test, 10000 dememorisation, 20 batches, 5000 iterations per batch) was calculated for each size group

and cohort in GenePop. Basic genetic diversity statistics were calculated for each locus and cohort using GenAlEx v6.501 (Peakall & Smouse 2006, 2012). These included: number of alleles per marker (A_n), effective number of alleles per marker (A_e), allele frequencies, heterozygosities (H_o and unbiased H_e), Shannon's information index (I), and per locus F_{IS}. A Kruskal-Wallis (KW; $P < 0.05$) test was used to test for significant differences in number of alleles, effective number of alleles, Shannon's information index and heterozygosity between the large and small groups of each cohort. Pairwise linkage disequilibrium within the FBC cohort was calculated in GenePop. A locus-by-locus Analysis of Molecular Variance (AMOVA; 10000 permutations, significance: $P < 0.05$) was performed for the FBC cohort in GenAlEx to determine the extent of the genetic variance between the large and small groups attributed to each locus; as further measures of differentiation, per locus F_{ST}, and -G''$_{ST}$ were also calculated. Phenotypic variation was evaluated by calculating means, standard deviation and coefficients of variance for each size-related trait (shell width, -length, and live weight) within the three cohorts, and the total population of 661 individuals, using XLStatistics v12.11.22 (Carr 2012). A KW test was also performed to evaluate significant differences in the traits amongst the various cohorts and the total sampling population. Correlations between each pair of traits were also calculated.

A number of association tests were performed to determine whether any significant relationships exist between particular loci and size in the respective cohorts (Figure 3.1). Rather than applying the stringent Bonferroni adjustment of the significant *P*-value (Dunn 1961), corroborative results amongst the various association methods was used to minimise false positive detections. First, case-control analyses were performed, including: exact G-tests for allelic and genotypic differentiation (Goudet *et al*. 1996) (10000 dememorisation, 20 batches, 5000 iterations per batch), which were done in GenePop, and permutation-based distance tests for both allelic and genotypic association (using Prevosti distance estimate; 1000 permutations, significance: $P < 0.05$) (Prevosti *et al*. 1975; Nielsen & Weir 1999), in PowerMarker v3.25 (Liu & Muse 2005). Secondly, single-locus F-tests for associations with quantitative traits were performed in PowerMarker (significance: $P < 0.05$). For the case-control tests, individuals were typed as either large or small, while the F-tests utilised the three size-based quantitative traits, live weight (g), shell width (mm) and shell length (mm). A fourth association test, the multi-allelic trend test (Slager & Schaid 2001) (significance: $P < 0.05$), was also performed for the FBC cohort.

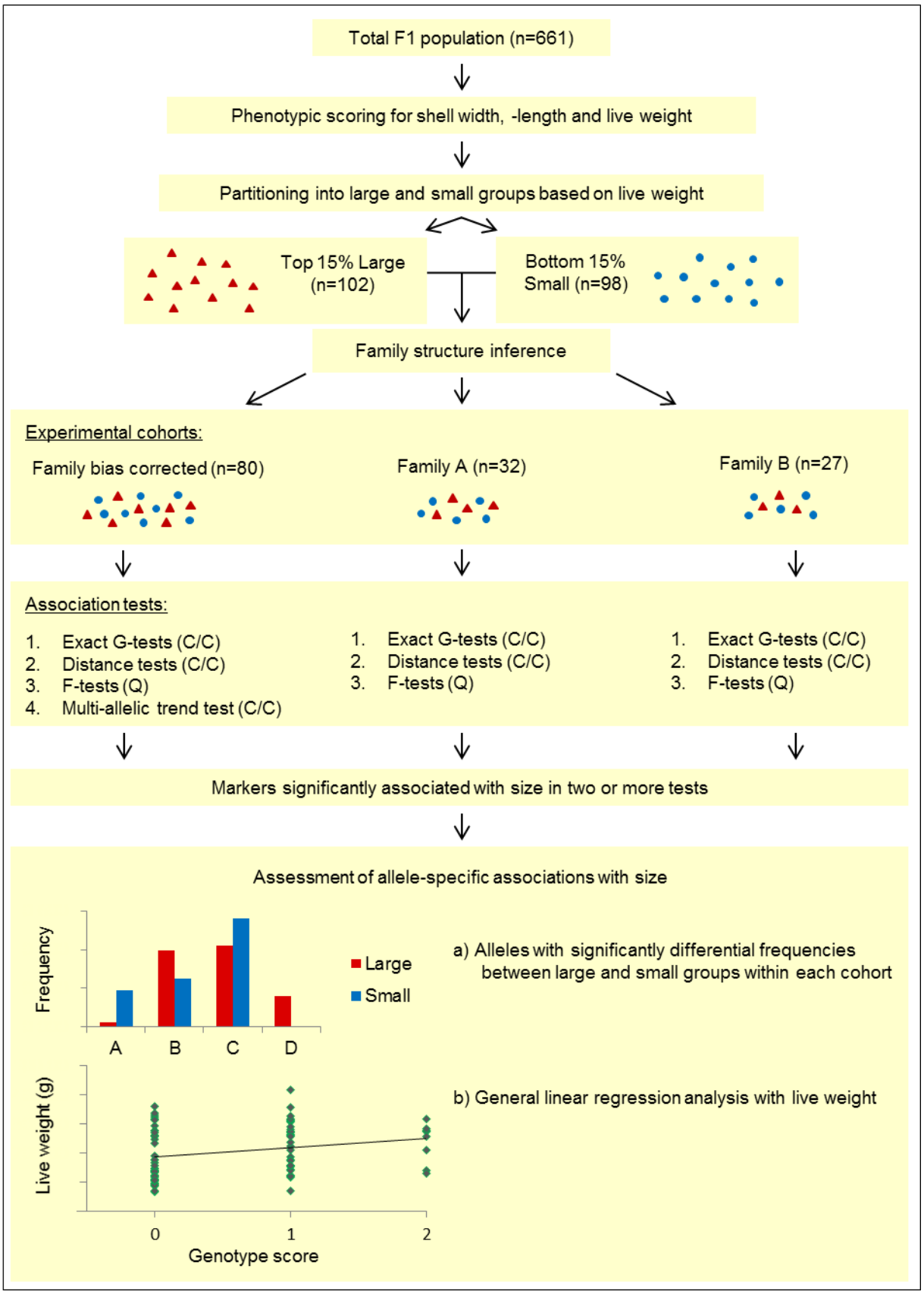


Figure 3.1: Graphical summary of the methodological approach, detailing the construction of the study populations, the association analyses performed for the various cohorts, and the assessment of allele-specific associations with size for significantly associated markers.

(C/C = Case control; Q = Quantitative)

This test, which makes use of a chi-squared statistic, could only be done for the FBC cohort due to sample size constraints for the family cohorts.

In terms of overall significance, markers within the same cohort that were found to be significantly associated with size in at least two of the association tests were considered potential growth related QTLs. To assess allele-specific associations for the identified candidate QTLs, the allelic distributions between large and small groups were individually assessed, within each cohort, in order to identify alleles with differential frequencies (Figure 3.1). The cohorts were then divided into individuals homozygous (genotype: A : A; genotypic score: 2), and heterozygous (genotype: A : -; genotypic score: 1) for the particular allele, and individuals with allelic combinations that excluded the allele of interested (genotype: - : -; genotypic score: 0). General linear regression analysis was then performed to detect association with live weight and particular allelic combinations (Figure 3.1). Where more than one allele (across different loci) had a significant association to size, allelic combinations of these "synergistic" alleles across respective loci was also tested. Finally, BLAST searches (Altschul *et al*. 1990) were conducted for all markers to determine whether they could be linked with any known genes. Sequences were compared with the Reference RNA and Non-redundant nucleotide databases using the BLASTn algorithm for somewhat similar sequences (*E*-value < 0.001), as well as the Non-redundant protein database using the BLASTx algorithm (*E*-value < 0.001).

3.3) Results

3.3.1) Marker efficiency evaluation

During allele size scoring, it was found that one of the markers, *HmNSS1*, could not be scored reliably (duplicated locus), and four markers (*HmidPS1.561*, *HmidILL.88398*, *HmidILL.37506*, and *HmidILL1.2192*) were monomorphic; these markers were therefore excluded from further analyses, with the exception of the BLAST searches. Across the three cohorts, the majority of markers appeared not to be in HW equilibrium ($P > 0.05$), although this number did decrease slightly at the 1% level (Table S3.2). Significant evidence for null alleles was found for three of the markers within the FBC cohort: *HmLCS5* (null allele frequency 97.5% CI: 0.4621-0.6354), *HmNS18* (null allele frequency 97.5% CI: 0.3827-0.5789), and *HmidILL.64192* (null allele frequency 97.5% CI: 0.3556-0.5453) (Table S3.3).

3.3.2) Phenotypic- and genetic diversity statistics

The phenotypic diversity statistics for the three cohorts and the original population are summarised in Table 3.1. The KW tests confirmed that all size groups were significantly different from the original population of 661 F_1 individuals for each trait at the 1% level. In general, the coefficients of variance for shell length and shell width were comparable, while live weight appeared to be the most variable trait. On average, the small groups also had higher coefficients of variance than the large groups for all three traits. The correlations between each pair of traits were similarly high, with R^2-values ranging between 0.9315 (shell width *versus* live weight) and 0.9610 (shell length *versus* live weight), and correlation coefficients between 0.9651 (P = 0.0000) and 0.9803 (P = 0.0000), respectively (Figures S3.1A – C).

Table 3.1: Basic phenotypic diversity statistics for shell length (mm), shell width (mm) and live weight (g) within the large and small groups of the FBC cohort, Family A, Family B, and the total population. These include: means with standard deviations (SD) and coefficients of variance (CV).

		Family-bias corrected (n=80)		**Family A (n=32)**		**Family B (n=27)**		**Total Population (n=661)**
		Large (n = 37)	**Small (n = 43)**	**Large (n = 16)**	**Small (n = 16)**	**Large (n = 19)**	**Small (n = 8)**	
Shell length	**Mean**	115.71	88.33	113.07	91.53	113.64	88.88	101.05
(mm)	**(±SD)**	(±6.28)	(±8.74)	(±5.91)	(±10.44)	(±4.97)	(±4.63)	(±14.68)
	CV	0.0542	0.0989	0.0522	0.1141	0.0438	0.0521	0.1453
Shell width	**Mean**	86.96	66.09	84.65	69.65	85.51	67.51	76.85
(mm)	**(±SD)**	(±5.41)	(±7.49)	(±5.26)	(±7.66)	(±4.36)	(±4.50)	(±11.70)
	CV	0.0622	0.1134	0.0622	0.1099	0.0510	0.0666	0.1522
Live weight	**Mean**	290.15	133.95	269.37	153.77	267.43	139.70	183.83
(g)	**(±SD)**	(±40.98)	(±35.52)	(±39.63)	(±50.37)	(±34.76)	(±13.77)	(±63.05)
	CV	0.1412	0.2652	0.1471	0.3276	0.1300	0.0986	0.3430

No significant differences (KW: $P > 0.05$), in genetic diversity within the FBC cohort, were observed between the large and small groups based on mean number of alleles, mean number of effective alleles, and unbiased heterozygosity (Figure 3.2, Table S3.2). Private alleles were reported for *HmidPS1.559*, *HmLCS48* and *HmidILL2.87955* within the large group, and *HmLCS48*, *HmidILL.146360* and *HmidILL2.87955* within the small group (Figures S3.3A – D). Within the FBC cohort as a whole, the number of alleles per marker ranged widely, with the highest being ten, for *HmNR106*, and the lowest being two, for *HmNS18* (Table S3.2). The majority of markers displayed relatively high heterozygosities,

with the notable exception of *HmidPS1.559*, which had an unbiased expected heterozygosity of 0.1910. In accordance with the high heterozygosities, individual F_{IS} values per marker for all other markers indicated that the majority of markers possessed an excess of heterozygotes. Estimates of pairwise linkage disequilibrium (LD) within the FBC cohort indicated that significant levels of LD exist between many of the markers, in particular, between *HmidILL2.87955* and *HmNR106* and almost all of the other markers (Table S3.4).

Within the family cohorts, genetic diversity estimates were again comparable between the large and small groups of each cohort, apart from the unbiased expected heterozygosities for Family B, where the expected heterozygosity for the small group was distinctly lower than that of the large group (Figure 3.2, Table S3.2). When comparing the size groups within each family separately, all groups, with the exception of the large group from Family A and the small group from Family B, possessed private alleles (Figures S3.2B – C). The number of alleles per marker were similar within each family, with the highest being four, for *H midILL-146360* and *HmNR106*, and the lowest being one, for *HmidPS1.559*, within both families (Table S3.2). Heterozygosities were high, with the exception again of *HmidPS1.559*, which was monomorphic within both families. Per marker F_{IS} values were negligible for all other markers.

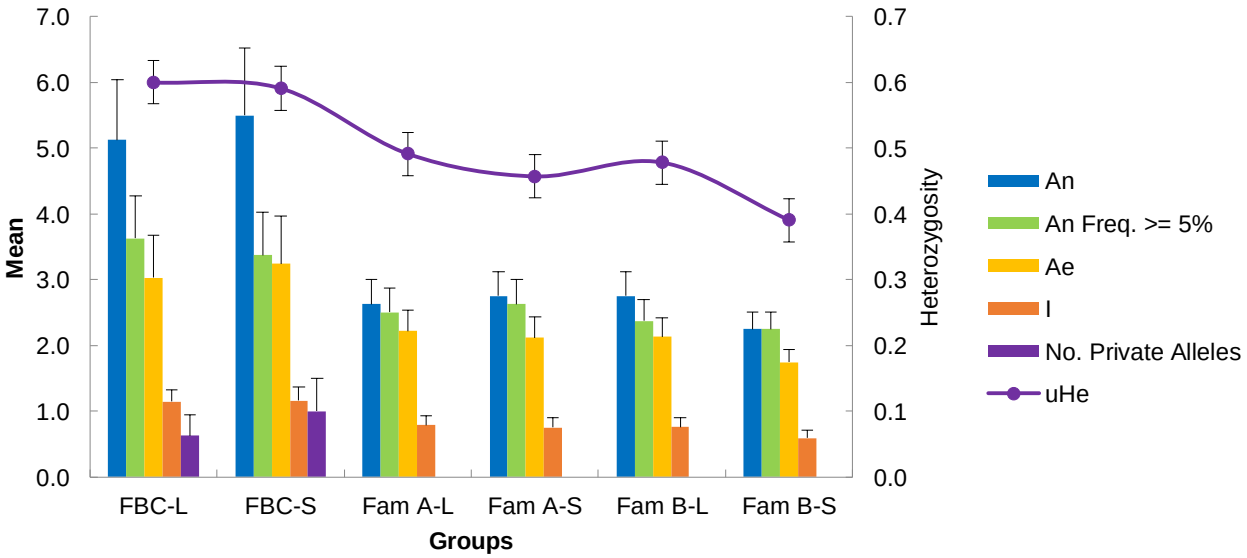


Figure 3.2: Summary of genetic diversity statistics across the large (L) and small (S) groups of the FBC cohort (FBC-L, FBC-S), and Families A (Fam A-L, Fam A-S) and B (Fam B-L, Fam B-S). These include the mean number of alleles (A_n), mean number of alleles with a frequency of above 5%, mean number of effective alleles (A_e), mean for Shannon's Information Index (I), mean number of private alleles and mean unbiased expected heterozygosities (uH_e). Error bars denote standard error.

The F_{ST} estimates indicated that most markers showed little evidence of differentiation (F_{ST} < 0.05; *P* > 0.05) between the large and small groups within the FBC cohort (Table 3.2) (Small *et al*. 1998; Mant *et al*. 2005), although the values for three markers, *HmNS18* (0.0117, *P* = 0.0291), *HmidILL.146360* (0.0232, *P* = 0.0045) and *HmidPS1.559* (0.0238, *P* = 0.0278), were significant at the 5% nominal level. The G''_{ST} estimates demonstrated slightly stronger evidence for differentiation than the F_{ST} estimates (*e.g. HmidILL.146360*), and as F_{ST} tends to underestimate differentiation, it is possible that the inflated G''_{ST} values represent a more accurate estimate of population differentiation (Bird *et al*. 2011; Meirmans & Hedrick 2011). Similarly, the locus-by-locus AMOVA (Table 3.2) indicated that significantly higher percentages of the variation between the large and small groups of the FBC cohort could be attributed to the following markers: *HmNS18* (1.45%), *HmidILL.146360* (3.25%) and *HmidPS1.559* (1.85%).

Table 3.2: Locus-by-locus Analysis of Molecular Variance (AMOVA) results for the variances among the large and small groups of the FBC cohort, as well as F_{st}- and G''_{st} estimates.

Marker	SSD[a]	Df	Variance	% Variation	F_{st}	G''_{st}
HmidILL2.87955	0.3091	1	0.0004	0.1179	0.0064	0.0051
HmLCS5	0.1293	1	0.0000	0.0000	0.0024	-0.0330
HmLCS48	0.1453	1	0.0000	0.0000	0.0036	-0.0200
Hm-NS18M	0.5574	1	0.0055	1.4459	**0.0117***	**0.0355***
HmidILL.146360	1.4740	1	0.0133	3.2478	**0.0232***	**0.1569***
HmidILL.64192	0.0264	1	0.0000	0.0000	0.0005	-0.0313
HmNR106	0.3132	1	0.0000	0.0000	0.0026	-0.0577
HmidPS1.559	0.2919	1	0.0020	1.8547	**0.0238***	**0.0395***

** Significant P-value (P < 0.05)*
[a] SSD = Sum of squared deviations; Df = Degrees of freedom

3.3.3) Association analyses

The results of the exact G-tests for allelic and genotypic differentiation revealed that a number of the markers demonstrated significant differentiation between the large and small groups within the FBC cohort and within Family B (Figure 3.3, Table S3.5). In particular, *HmidILL2.87955* and *HmidILL.146360* expressed both allelic and genotypic differentiation within the FBC cohort, with *HmidILL.146360* having the most significant values (*P* = 0.0017 and 0.0053 for allelic- and genotypic differentiation, respectively), whilst *HmNS18* was significant only for genotypic differentiation. Within Family B,

HmidILL.146360 had significant values for both allelic and genotypic differentiation, and *Hmid-064192P* was significant only for genotypic differentiation. No markers within Family A showed significant differentiation.

As with the exact tests, the genetic distance tests for allelic and genotypic association with size revealed that *HmidILL.146360* possessed a significant disparity in allelic- and genotypic frequencies between the large and small groups within the FBC cohort (Figure 3.3, Table S3.5). Significant results for this marker were also obtained for the multi-allelic trend test. A similar set of results was obtained for *HmNS18* within the FBC cohort. Within Family B, *Hmid-064192P* had significant values for both allelic and genotypic distance values. As with the exact tests for differentiation, no markers within Family A had significant distance values. The single locus F-test for associations with quantitative traits presented significant results ($P < 0.05$) for *HmidILL2.87955* with all three of the quantitative traits (shell width, shell length, and live weight) within the FBC cohort (Figure 3.3, Table S3.6). Within Family B, *HmidILL.64192* showed significant results for all three of the quantitative traits, and a significant association with shell width was observed for *HmidILL.146360* (Figure 3.3, Table S3.6). Within Family A, two markers, *HmLCS5* and *HmNS18*, had significant associations with growth traits: shell width and live weight for *HmLCS5*, and shell length and live weight for *HmNS18*.

Overall, the most significantly associated markers, with congruent results across multiple analyses, appeared to be *HmidILL2.87955*, *HmNS18* and *HmidILL.146360* within the FBC cohort, and *HmidILL.146360* and *HmidILL.64192* within Family B (Figure 3.3). No markers were found to be significant for more than one test within Family A. Examining the allelic distributions for the significant markers identified a number of alleles that could have been the cause of the respective associations with size (Figures S3.4A – C and Figures S3.5A – B); however the linear regression analyses revealed that only two alleles within the FBC cohort, 196 from *HmidILL2.87955* ($R^2 = 0.0926$; correlation coeff. = 0.3043; $P = 0.0064$) and 395 from *HmidILL.146360* ($R^2 = 0.0588$; correlation coeff. = 0.2426; $P = 0.0302$), were significantly associated with large size (Figure S3.6B and Figure S3.7A). Conversely, another two alleles from the same markers, 212 from *HmidILL2.87955* ($R^2 = 0.0856$; correlation coeff. = -0.2926; $P = 0.0089$) and 385 from *HmidILL.146360* ($R^2 = 0.0512$; correlation coeff. = -0.2263; $P = 0.0435$), were significantly associated with small size (Figure S3.6A and Figure S3.7B).

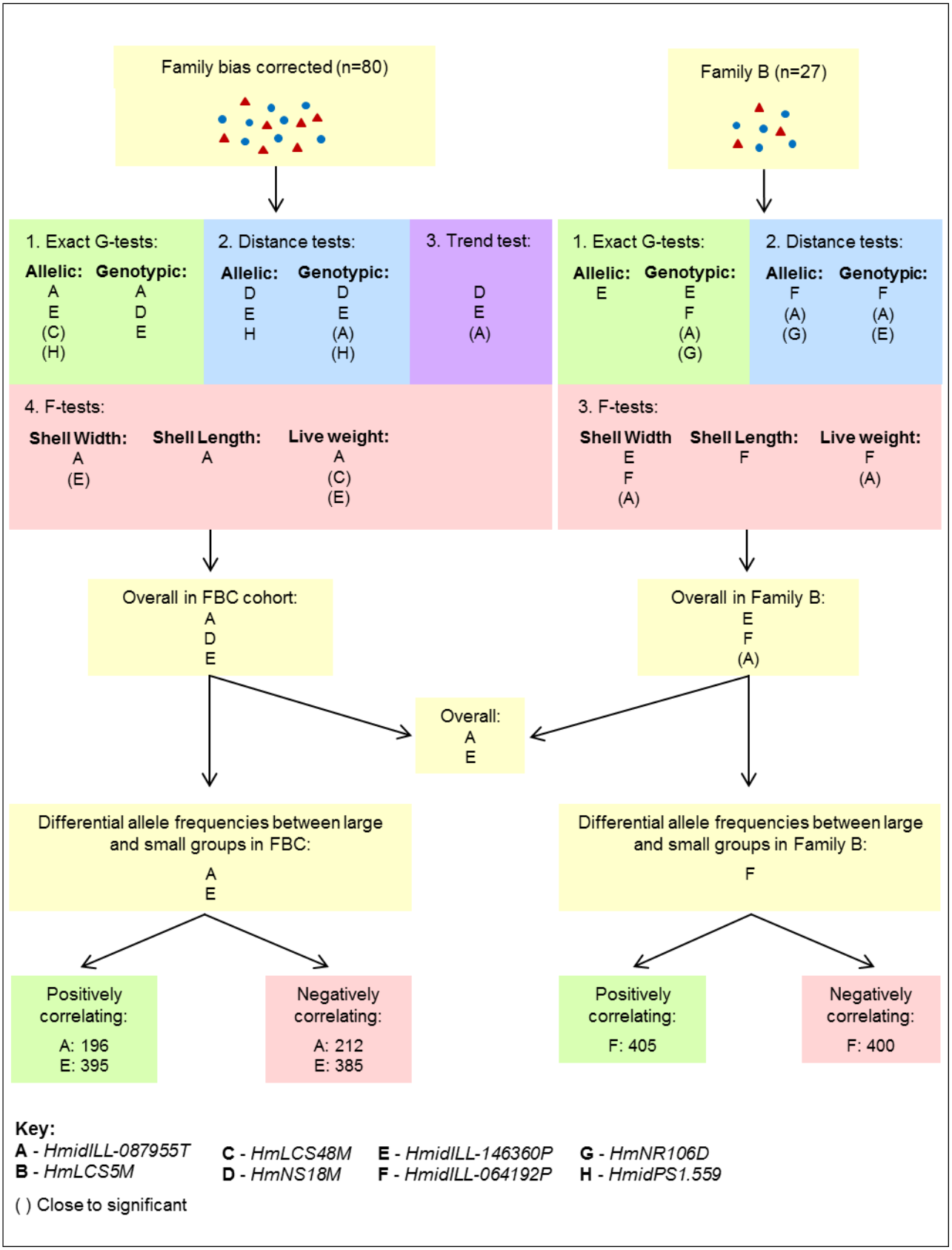


Figure 3.3: Graphical summary of the association analyses results for the FBC cohort and Family B, as well as the results of the assessment of allele-specific associations with size for significantly associated markers.

None of the alleles from *HmNS18* had significant associations with size. The associations for the allelic combinations, 196/395 (*HmidILL2.87955* / *HmidILL.146360*) (R^2 = 0.1041; correlation coeff. = 0.3227; *P* = 0.0035) and 212/385 (*HmidILL2.87955* / *HmidILL.146360*) (R^2 = 0.1092; correlation coeff. = -0.3304; *P* = 0.0028), were somewhat stronger than those for the individual alleles (Figures S3.8A – B). Within Family B, one allele, 405 from *HmidILL.64192* (R^2 = 0.3941; correlation coeff. = 0.6278; *P* = 0.0005), was positively associated with size, and one allele, 400 from *HmidILL.64192* (R^2 = 0.3513; correlation coeff. = -0.5927; *P* = 0.0011), was negatively associated (Figures S3.9A – B).

The BLAST searches conducted for the two most significantly associated markers, *HmidILL2.87955* and *HmidILL.146360*, as determined by the results of the association analyses and other population differentiation data, both produced multiple hits, the most significant of which were recorded (Table S3.7). For *HmidILL2.87955*, the most significant hit was to the 3' untranslated region (UTR) of the ATP-binding cassette (ABC) transporter 11 family A (*ABCA11*) gene in *Dictyostelium discoideum*, and for *HmidILL.146360*, the 5' UTR of the activin A receptor type 1 (*ACVR1*) gene in *Homo sapiens*. The majority of the remaining markers also produced significant hits to known or hypothetical gene sequences. For four of these markers (*HmNS18*, *HmNR106*, *HmLCS5* and *HmidILL.88398*) conserved regions within the query sequences had hits to multiple different genes (Table S3.8). These regions often included, but where never limited to, the microsatellite repeat regions. In contrast, markers *HmidILL.64192*, *HmidILL.37506* and *HmidILL1.2192*, each produced multiple hits to single genes within different organisms. The most significant for each was to the hypothetical homologue of the D-lactate dehydrogenase gene within *Lottia gigantea*, the Ras-related protein Rab-1A gene within *Haliotis discus*, and the 14-3-3 protein zeta gene within *Haliotis diversicolor*, respectively (Table S3.7).

3.4) Discussion

3.4.1. Marker evaluation

The majority of markers seemed not to conform to HW expectations; however this is not unexpected considering that the sample populations were not population representative samples, but a selection of individuals from phenotypic extremes. Furthermore, demographic and selective processes, accompanied by the recent founder event of the commercial populations from the wild population, also suggest that this population might

not conform to HW expectations (Rhode *et al*. 2012, 2014). This could have an impact on certain association analyses, as the allelic versions of the exact G-test and distance test assume loci to be in HW equilibrium. As a result, the genotypic exact and -distance tests are likely to provide more accurate results, as they assume independent sampling of genotypes, rather than alleles (Goudet *et al*. 1996). The presence of significant evidence for null allele frequencies for some of the markers is also not unexpected, but an excess of homozygotes at particular loci could also be attributed to directional selection (Rhode *et al*. 2013).

3.4.2) Phenotypic- and genetic diversity

There was a high and significant correlation between all size traits (live weight, shell length and shell width), which is consistent with the trends reported in previous studies for other abalone species (*e.g.* Rourke *et al*. 2003; You *et al*. 2010). Therefore, it may only be necessary to measure one of the traits as an estimate for growth rate in *H. midae*. Live weight is, however, the most frequently used size estimate for selection in abalone aquaculture practice, and was therefore also used in this study to make the primary differentiation between large and small animals. Live weight was also the trait that demonstrated the most variability, likely due to the compound nature of this trait that is dependent on a number of factors, including shell size (Luo *et al*. 2013). In general, the small phenotype seemed to be more variable than the large phenotype, suggesting possible large effect alleles are responsible for large size (Olson-Manning *et al*. 2012).

The levels of genetic diversity within the large and small groups of the FBC cohort were relatively high, and were comparable between the groups. Similarly, genetic diversity estimates within the family cohorts were also comparable between size classes, with the exception of the small group of Family B, which had a significantly lower heterozygosity relative to the large group. This can, however, be explained by the over-representation of large individuals within the family (19 large versus 8 small individuals). The presence of private alleles within the majority of size groups could be significant, as private alleles can represent the variants associated with differences in phenotype (Hinds *et al*. 2005). However, in the present study, all private alleles occurred at very low frequencies, which is more suggestive of the exclusivity being merely a consequence of the relatively small sample sizes. The regularity of genetic variability amongst the size classes within each cohort is indicative of the appropriateness of the genetic matches, in terms of case-control

association studies, as well as a fair representation of genetic variants that may underlie particular phenotypes.

3.4.3) Association with size

After considering congruence amongst the various association tests, the marker most strongly associated with size appeared to be *HmidILL.146360*, as it was highlighted in various tests within the FBC cohort, as well as in Family B. In addition, the AMOVA results showed that, out of all the markers, *HmidILL.146360* made the largest contribution (3.25%) to the genetic variance between the large and small groups within the FBC cohort. While slightly less supported, *HmidILL2.87955*, as well as *HmNS18*, could also be considered as fairly strong candidates for association with size, as both achieved significant results within multiple tests.

However, of these three candidates, only *HmidILL.146360* and *HmidILL2.87955* were derived from expressed sequence tags (ESTs) (Rhode *et al*. 2012), which, combined with their apparent association with size, increases the possibility that they are associated with genes involved in growth physiology. Furthermore, significant linkage disequilibrium was observed between these two markers, which may indicate a functional relationship, and further support the possibility of gene networks associated with growth traits (Rhode *et al*. 2013). Significant LD was also observed between numerous other markers within the FBC cohort; however, it is likely that most of these represent syntenic LD only, as the majority of these marker pairs were not found to be significant within the association tests, and non-syntenic LD could be considered more characteristic of a functional relationship between loci. Only four of the markers used for the association analysis (*HmidILL2.87955*, *HmidILL1.2192*, *HmLCS48*, and *HmidPS1.559*) are represented on the most recent *H. midae* linkage map (Vervalle *et al*. 2013), therefore, it is currently not possible to verify this. Lastly, *HmidILL.146360* and *HmidILL2.87955* were also the only markers within the FBC cohort with alleles that correlated significantly, albeit weakly, with size. While only one allele per locus seems to be associated with large size, the combination of both alleles (across loci) increased the strength of the correlation, which may indicate an additive, polygenic effect. A similar effect was observed for the alleles from these two loci that associated with small size. These results are not unexpected, as growth rate is a complex trait that is dependent on many factors, including genetic and environmental effects. The significantly associated alleles seemed to explain approximately 9 – 35% (R^2-values for

allelic correlations) of the phenotypic variation between large and small animals, which is consistent with other results obtained by similar studies in both abalone and other species (Rockman & Kruglyak 2006; Liu *et al*. 2007; Baranski *et al*. 2008; Slabbert 2010). However, it is possible that these values are an overestimation of the actual effect sizes of these alleles, as smaller sample sizes may fail to detect loci of small effect, *i.e.* the "Beavis-effect" (Beavis 1998). On the contrary, Rhode *et al*. (2013) proposed that the domestication event produced a selective sweep, facilitating divergence between wild and cultured populations, which would predominantly act on loci with large effects on the phenotype, creating molecular signatures of selection at these loci (Olson-Manning *et al*. 2012; Kemper *et al*. 2014).

While the above results are indeed significant, it is acknowledged that not all of the results were in complete agreement. In some cases, additional markers were highlighted within the families, as opposed to the FBC cohort, as was the case with *HmidILL.64192*. Alternatively, none of the markers were found to be significantly associated with size in Family A. While still of interest, it is important to note that family data is somewhat less informative in terms of identifying genuine population-wide candidate loci for association. Marker-phenotype associations drawn from segregating pedigrees are based on genetic linkage, and not necessarily LD, which limits their potential for direct extrapolation to the population level (Dekkers 2004). Therefore, rather than casting doubt on the validity of the candidate markers identified in the FBC cohort, these conflicting results may only be an indication that the microsatellites identified here are not the causal variants for the observed phenotypes, but may be linked to one or more unknown causal mutations (Hindorff *et al*. 2009). If this is indeed the case, the lack of significant associations within Family A could be explained by the occurrence of one or more recent recombination events between candidate loci and the true causal variants.

3.4.4) Artificial- versus *natural selection in generating signatures of selection*

All of the markers used within the present study demonstrated divergent selection between wild and F_1 cultured abalone populations (Rhode *et al*. 2013). However, the results of this study suggest that only two (±15%) of these loci (*HmidILL2.87955* and *HmidILL.146360*) are associated with size, and by extension, with artificial selection for increased growth rate. Both *HmidILL2.87955* and *HmidILL.146360* had BLAST hits to known genes; an ABC transporter and an activin A receptor, respectively. The microsatellite motifs were located

within either the 3' UTR or 5' UTR of the respective genes, which indicates possible causal variants within regulatory regions may be responsible for the observed association with size (Chen *et al*. 2006). These genes are also both involved in processes that could be associated with growth. The ABC transporters convey a wide variety of substances across cellular membranes (Peelman *et al*. 2003). The ABCA subfamily in particular is primarily associated with the transport of lipids, and is associated with a wide variety of biological processes, including cell signalling, membrane homeostasis, stem cell development, and waste disposal (Albrecht & Viturro 2007; Aye *et al*. 2009). The activin type 1 receptors are cell-surface serine/threonine kinases that play an integral role in signal transduction for the transforming growth factor β superfamily of ligands, and are conserved over a wide variety of taxa, including both vertebrates and invertebrates (Moustakas & Heldin 2009; Santibañez *et al*. 2011).

Many of the non-associated loci also had significant BLAST hits to known genes, which would suggest association with functional genomic regions. Among these were three out of the four loci found to be monomorphic within the study cohorts (*HmidILL.88398*, *HmidILL.37506*, and *HmidILL1.2192*), a result which is highly significant within the context of a selective sweep. Indeed, the degree of genetic polarisation caused by a selective sweep would likely be most extreme at loci with the greatest effect on fitness, although such an effect could also be explained by the reduction of diversity associated with factors such as founder events and unequal spawning contributions. Nevertheless, as *H. midae* has recently been introduced into the artificial aquaculture environment, it is therefore plausible that many of these markers, monomorphic or otherwise, are associated with traits involved in adaptation to the new environment, *i.e.* natural selection, where the production system produces novel selective pressures for survival (Johnsson *et al*. 2001; Robison & Rowland 2005; Wilcox & Martin 2006; Rhode *et al*. 2012). For example, the 14-3-3 protein zeta, to which *HmidILL1.2192* produced a significant hit, is known to play a key role in embryogenesis, and in particular, in neural differentiation and development (Kousteni *et al*. 1997; Acevedo *et al*. 2007). The differential selection at this marker may therefore represent a novel selective pressure for larval survival in the aquaculture environment. In addition to this, D-lactate dehydrogenase, to which *HmidILL.64192* produced a significant hit, forms an integral part of pyruvate metabolism during anaerobic glycolysis in molluscs (Baldwin *et al*. 1992). As sub-optimum dissolved oxygen levels have been observed within the aquaculture environment (Harris *et al*. 1999), and the build-up of D-lactate in muscle tissue has been associated with the response to anaerobic stress in

abalone (Baldwin *et al*. 1992; O'Omolo *et al*. 2003), a selective pressure for more efficient energy production in poorly oxygenated environments may be present in cultured populations. Alternatively, the observed signature of selection at this locus may represent a loss of selective pressure in the new environment. Increased levels of anaerobic metabolites were observed within the foot and abductor muscles of wild abalone exposed to wave action, as opposed to those inhabiting more sheltered environments, representing a potential for increased shell adhesion (Wells *et al*. 1998). In the wild, the need to remain attached to a rock despite being subject to heavy wave action represents an important survival pressure; however, as wave action is absent in the aquaculture environment, selection for stronger adhesive capabilities would be relaxed.

In light of these results, it is therefore suggested that the previously observed selective sweep caused by the domestication event is an interplay between all three proposed facets of selection, *i.e.* relaxed natural selection for wild survival traits, increased natural selection for captive survival traits, and increased artificial selection for traits of economic importance, although not necessarily within equal measure (O'Brian *et al*. 2005; Rhode *et al*. 2013). As the majority of loci within this subset of loci putatively under selection were not significantly associated with size, and therefore with artificial selection, it is therefore suggested that natural selection may be the predominant selective force in shaping genetic variation during the initial stages of abalone domestication within this population. Although by no means conclusive, such a result would not be unexpected; given the relatively low level of artificial selection currently in effect at this stage of domestication, it is unlikely that its effects would exceed the far more pervasive effects of natural selection.

3.5) Conclusion

The current study investigated the association of thirteen microsatellite loci, putatively identified as being under directional selection, with size-related traits in a first generation commercial abalone population. Size at age was used as a proxy for growth rate, the primary trait under artificial selection within abalone aquaculture. Significant associations were obtained for two loci, which were also putatively linked with genes that may have functional roles in growth and development. In particular, alleles were identified that showed significant correlation with either large or small size. As faster growth rate is currently the primary trait of economic importance in abalone aquaculture, these loci might therefore find application in marker-assisted selection programmes for *H. midae*.

Regarding the non-associated loci, bioinformatics analyses revealed that many of these loci could also be linked with known genes, which suggests they may be associated with traits involved in adaptation to the aquaculture environment. As the number of loci not associated with size represents the majority of the candidate loci under selection, these results would suggest that natural selection is the primary force driving the divergence of cultured abalone populations from their wild progenitors.

References

Acevedo S.F., Tsigkari K.K., Grammenoudi S. & Skoulakis E.M.C. (2007) *In vivo* functional specificity and homeostasis of *Drosophila* 14-3-3 proteins. *Genetics* 177, 239-253.

Albrecht C. & Viturro E. (2007) ABCA subfamily - gene and protein structures, functions and associated hereditary diseases. *European Journal of Physiology* 453, 581-589.

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

Aye I.L.M.H., Singh A.T. & Keelan J.A. (2009) Transport of lipids by ABC proteins: Interactions and implications for cellular toxicity, viability and function. *Chemico-Biological Interactions* 180, 327-339.

Baldwin J., Wells R.M.G., Low M. & Ryder J.M. (1992) Tauropine and D-lactate as metabolic stress indicators during transport and storage of live paua, (New Zealand abalone) (*Haliotis iris*). *Journal of Food Science* 57, 280-282.

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.

Beavis W.D. (1998) QTL analyses: Power, precision, and accuracy. In: Paterson A.H. (ed.), *Molecular Dissection of Complex Traits*. Boca Raton: CRC Press. pp 145-162.

Bird C.E., Karl S.A., Smouse P.E. & Toonen R.J. (2011) Detecting and measuring genetic differentiation. In: Held C., Koenemann S., Schubart C.D. (eds), *Phylogeography and Population Genetics in Crustacea*. Boca Raton: CRC Press. pp 31-55.

Carneiro M., Piorno V., Rubin C.-J., Alves J.M., Ferrand N., Alves P.C. & Andersson L. (2015) Candidate genes underlying heritable differences in reproductive seasonality between wild and domestic rabbits. *Animal Genetics* 46, 418-425.

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

Chen J.M., Férec C. & Cooper D.N. (2006) A systematic analysis of disease-associated variants in the 3' regulatory regions of human protein-coding genes I: general principles and overview. *Human Genetics* 120, 1-21.

Coombs J.A., Letcher B.H. & Nislow K.H. (2008) CREATE: a software to create input files from diploid genotypic data for 52 genetic software programs. *Molecular Ecology Resources* 8, 578-580.

Dekkers J.C.M. (2004) Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. *Journal of Animal Science* 82, 313-328.

Dixon T.J., Coman G.J., Arnold S.J., Sellars M.J., Lyons R.E., Dierens L., Preston N.P. & Li Y. (2008) Shifts in genetic diversity during domestication of Black Tiger shrimp, *Penaeus monodon*, monitored using two multiplexed microsatellite systems. *Aquaculture* 283, 1-6.

Dunn O.J. (1961) Multiple comparisons among means. *Journal of the American Statistical Association* 56, 52-64.

Evin A., Dobney K., Schafberg R., Owen J., Vidarsdottir U.S., Larson G. & Cucchi T. (2015) Phenotype and animal domestication: A study of dental variation between domestic, wild, captive, hybrid and insular *Sus scrofa*. *BMC Evolutionary Biology* 15, 6.

Feder M.E. & Mitchell-Olds T. (2003) Evolutionary and ecological functional genomics. *Nature Reviews Genetics* 4, 649-699.

Gholami M., Reimer C., Erbe M., Preisinger R., Weigend A., Weigend S., Servin B. & Simianer H. (2015) Genome scan for selection in structured layer chicken populations exploiting linkage disequilibrium information. *PLoS ONE 10*, e0130497.

Goudet J., Raymond M., de Meeüs T. & Rousset F. (1996) Testing differentiation in diploid populations. *Genetics* 144, 1933-1940.

Hara M. & Sekino M. (2007) Genetic differences between hatchery stocks and natural populations in Pacific Abalone (*Haliotis discus*) estimated using microsatellite DNA markers. *Marine Biotechnology* 9, 74-81.

Harris J.O., Maguire G.B., Edwards S.J. & Johns D.R. (1999) Low dissolved oxygen reduces growth rate and oxygen consumption rate of juvenile greenlip abalone, *Haliotis laevigata* Donovan. *Aquaculture* 174, 265-278.

Hedgecock D. & Pudovkin A.I. (2011) Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. *Bulletin of Marine Science* 87, 971-1002.

Hindorff L.A., Sethupathy P., Junkins H.A., Ramos E.M., Mehta J.P., Collins F.S. & Manolio T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* 106, 9362-9367.

Hinds D.A., Stuve L.L., Nilsen G.B., Halperin E., Eskin E., Ballinger D.G., Frazer K.A. & Cox D.R. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* 307, 1072-1079.

Johnsson J.I., Höjesjö J. & Fleming I.A. (2001) Behavioural and heart rate responses to predation risk in wild and domesticated Atlantic salmon. *Canadian Journal of Fisheries and Aquatic Sciences* 58, 788-794.

Kemper K.E., Saxton S.J., Bolormaa S., Hayes B.J. & Goddard M.E. (2014) Selection for complex traits leaves little or no classic signatures of selection. *BMC Genomics* 15, 246.

Konovalov D.A. (2006) Accuracy of four heuristics for the full sibship reconstruction problem in the presence of genotype errors. In: Jiang T., Yang U.C., Chen Y.P.P. & Wong L. (eds), *APBC*. London: Imperial College Press. pp 7-16.

Konovalov D.A., Manning C. & Henshaw M.T. (2004) KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Molecular Ecology Notes* 4, 779-782.

Kousteni S., Tura F., Sweeney G.E. & Ramji D.P. (1997) Sequence and expression analysis of a *Xenopus laevis* cDNA which encodes a homologue of mammalian 14-3-3 zeta protein. *Gene* 190, 279-285.

Li Q., Park Q., Endo T. & Kijima A. (2004) Loss of genetic variation at microsatellite loci in hatchery strains of the Pacific abalone (*Haliotis discus hannai*). *Aquaculture* 235, 207-222.

Lind C.E., Evans B.S., Knauer J., Taylor J.J.U. & Jerry D.R. (2009) Decreased genetic diversity and a reduced effective population size in cultured silver-lipped pearl oysters (*Pinctada maxima*). *Aquaculture* 286, 12-19.

Liu K. & Muse S.V. (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21, 2128-2129.

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.

Luo X., Ke C. & You W. (2013) Estimates of correlations for shell morphological traits on body weight of interspecific hybrid abalone (*Haliotis discus hannai* and *Haliotis gigantea*). *Journal of Shellfish Research* 32, 115-118.

Mant J., Peakall R. & Schiestl F.P. (2005) Does selection on floral odor promote differentiation among populations and species of the sexually deceptive orchid genus *Ophrys*? *Evolution* 59, 1449-1463.

Meirmans P. & Hedrick P.W. (2011) Assessing population structure: F_{ST} and related measures. *Molecular Ecology Resources* 11, 5-18.

Mignon-Grasteau S., Boissy A., Bouix J., Faure J.-M., Fisher A., Hinch G.N., Jensen P., Le Neindre P., Mormède P., Prunet P., Vandeputte M. & Beaumont C. (2005) Genetics of adaptation and domestication in livestock. *Livestock Production Science* 93, 3-14.

Moioli B., Scatà M.C., Steri R., Napolitano F. & Catillo G. (2013) Signatures of selection identify loci associated with milk yield in sheep. *BMC Genetics* 14, 76.

Moustakas A. & Heldin C.-H. (2009) The regulation of TGFβ signal transduction. *Development* 136, 3699-3714.

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

Nielsen D.M. & Weir B.S. (1999) A classical setting for associations between markers and loci affecting quantitative traits. *Genetical Research* 74, 271-277.

O'Brian E., Bartlett J., Crump P., Dixon B. & Duncan P. (2005) Enhancement of saucer scallops (*Amusium balloti*) in Queensland and Western Australia – Genetic considerations. Report No. 2003/033. Department of Primary Industries and Fisheries, Australia.

Olson-Manning C.F., Wagner M.R. & Mitchell-Olds T. (2012) Adaptive evolution: evaluating empirical support for theoretical predictions. *Nature Reviews Genetics* 13, 867-877.

O'Omolo S., Gäde G., Cook P.A. & Brown A.C. (2003) Can the end products of anaerobic metabolism, tauropine and D-lactase, be used as metabolic stress indicators during transport of live South African abalone *Haliotis midae*? *African Journal of Marine Science* 25, 301-309.

Park S.D.E. (2001) Trypanotolerance in West African Cattle and the population genetic effects of selection. Unpublished PhD thesis, University of Dublin, Dublin.

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

Peakall R. & Smouse P.E. (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics* 28, 2537-2539.

Peelman F., Labeur C., Vanloo B., Roosbeek S., Devaud C., Duverger N., Denèfle P., Rosier M., Vandekerckhove J. & Rosseneu M. (2003) Characterization of the ABCA transporter subfamily: Identification of prokaryotic and eukaryotic members, phylogeny and topology. *Journal of Molecular Biology* 325, 259-274.

Praipue P., Klinbunga S. & Jarayabhand P. (2010) Genetic diversity of wild and domesticated stocks of Thai abalone, *Haliotis asinina* (Haliotidae), analyzed by single-strand conformational polymorphism of AFLP-derived markers. *Genetics and Molecular Research* 9, 1136-1152.

Prevosti A., Ocaña J. & Alonso G. (1975) Distances between populations for *Drosophila subobscura* based on chromosome arrangement frequencies. *Theoretical and Applied Genetics* 45, 231-241.

Rhode C., Hepple J., Jansen S., Davis T., Vervalle J., Bester-van der Merwe A.E. & Roodt-Wilding R. (2012) A population genetics analysis of abalone domestication events in South Africa: Implications for the management of the abalone resource. *Aquaculture* 356–357, 235-242.

Rhode C., Vervalle J., Bester-van der Merwe A.E. & Roodt-Wilding R. (2013) Detection of molecular signatures of selection at microsatellite loci in the South African abalone (*Haliotis midae*) using a population genomic approach. *Marine Genomics* 10, 27-36.

Rhode C., Maduna S.N., Roodt-Wilding R. & Bester-van der Merwe A.E. (2014) Comparison of population genetic estimates amongst wild, F1 and F2 cultured abalone (*Haliotis midae*). *Animal Genetics* 45, 456-459.

Robison B.D. & Rowland W. (2005) A potential model system for studying the genetics of domestication: behavioral variation among wild and domesticated strains of zebra danio (*Danio rerio*). *Canadian Journal of Fisheries and Aquatic Sciences* 62, 2046-2054.

Rockman M.V. & Kruglyak L. (2006) Genetics of global gene expression. *Nature Reviews Genetics* 7, 862-872.

Rourke M., Baranski M., McKinnon L., Hayes B. & Robinson N. (2003) Preliminary investigation of growth and meat quality traits in farmed abalone. In: *Proceedings of the 15th conference of the Association for the Advancement of Animal Breeding and Genetics*, 7-11 July, Melbourne, Australia. pp 413-416.

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

Rubin C.-J., Zody M.C., Eriksson J., Meadows J.R., Sherwood E., Webster M.T., Jiang L., Ingman M., Sharpe T., Ka S., Hallböök F., Besnier F., Carlborg Ö., Bed'hom B., Tixier-Boichard M., Jensen P., Siegel P., Lindblad-Toh K. & Andersson L. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464, 587-591.

Saghai-Maroof M.A., Soliman K.M., Jorgenson R.A. & Allard R.W. (1984) Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proceedings of the National Academy of Sciences of the United States of America* 81, 8014-8018.

Salin F. (2010) AutoBin v0.9. Available at http://www4.bordeaux-aquitaine.inra.fr/biogeco/Ressources/Logiciels/ Autobin [accessed 26 August 2013].

Santibañez J.F., Quintanilla M. & Bernabeu C. (2011) TGF-β/ TGF-β receptor system and its role in physiological and pathological conditions. *Clinical Science* 121, 233-251.

Skaala Ø., Høyheim B., Glover K. & Dahle K. (2004) Microsatellite analysis in domesticated and wild Atlantic salmon (*Salmo salar* L.): allelic diversity and identification of individuals. *Aquaculture* 240, 131-143.

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

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

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

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

Slager S.L. & Schaid D.J. (2001) Evaluation of candidate genes in case-control studies: A statistical method to account for related subjects. *The American Journal of Human Genetics* 68, 1457-1462.

Small M.P., Beacham T.D., Withler R.E. & Nelson R.J. (1998) Discriminating coho salmon (*Oncorhynchus kisutch*) populations within the Fraser River, British Columbia, using microsatellite DNA markers. *Molecular Ecology* 7, 141-155.

Stella A., Ajmone-Marsan P., Lazzari B. & Boettcher P. (2010) Identification of selection signatures in cattle breeds selected for dairy production. *Genetics* 185, 1451-1461.

Storz J.F. (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* 14, 671-688.

Vasemägi A., Nilsson J., McGinnity P., Cross T., O'Reilly P., Glebe B., Peng B., Berg P.R. & Primmer C.R. (2012) Screen for footprints of selection during domestication/captive breeding of Atlantic salmon. *Comparative and Functional Genomics* 2012, 1-14.

Vervalle J., Hepple J., Jansen S., du Plessis J., Wang P., Rhode C. & Roodt-Wilding R. (2013) Integrated linkage map of *Haliotis midae* Linnaeus based on microsatellite and SNP markers. *Journal of Shellfish Research* 32, 89-103.

Wells R.M.G., McShane P.E., Ling N., Wong R.J., Lee T.O.C. & Baldwin J. (1998) Effect of wave action on muscle composition, metabolites and growth indices in the New Zealand abalone, paua (*Haliotis iris*), with implications for harvesting and aquaculture. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* 119, 129-136.

Wilcox J.L. & Martin A.P. (2006) The devil's in the details: genetic and phenotypic divergence between artificial and native populations of the endangered pupfish (*Cyprinodon diabolis*). *Animal Conservation* 9, 316-321.

Wilkinson S., Lu Z.H., Megens H.-J., Archibald A.L., Haley C., Jackson I.J., Groenen M.A.M., Crooijmans R.P.M.A., Ogden R. & Wiener P. (2013) Signatures of diversifying selection in European pig breeds. *PLoS Genetics* 9, e1003453.

You W., Ke C., Luo X. & Wang D. (2010) Genetic correlations to morphological traits of small abalone *Haliotis diversicolor*. *Journal of Shellfish Research* 29, 683-686.

Chapter 4
Study Conclusions

4.1) Overview

Population genetic theory defines linkage disequilibrium (LD) as the non-random association of alleles at two or more loci within a population (Slatkin 2008). However, although much of the LD observed within populations is the result of genetic linkage, the independence of alleles at different loci is also sensitive to a variety of population-specific biological phenomena. These include locus-specific factors, such as recombination rate, selection, mutation, and epistasis, as well as demographic factors, such as effective population size, genetic drift, population subdivision, mating system, migration and admixture (Flint-Garcia *et al*. 2003; Gaut & Long 2003; Mueller 2004; Slatkin 2008). Given that LD is sensitive to such a wide variety of population-specific phenomena, the characterisation of LD within species of interest can provide much insight into the demographic and evolutionary factors that have shaped them genetically (Slatkin 2008).

In addition to its utility in investigating population genetic processes, LD data has also been recognised as an effective means of identifying and locating the genomic regions responsible for the expression of economically important traits, or quantitative trait loci (QTLs) (Goddard & Hayes 2009). Linkage disequilibrium- or association mapping is based on the detection of genotype-phenotype associations resulting from marker-QTL co-segregation, and can be implemented as a genome-wide association study, or targeted towards a selection of candidate regions/loci (Stinchcombe & Hoekstra 2007; Goddard & Hayes 2009). Quantitative trait loci identified in this manner can then be applied as performance evaluation markers in marker-assisted selection (MAS) programmes, which offset many of the shortfalls of conventional phenotypic selection (Knapp 1998; Bentsen & Olesen 2002; Dekkers & Hospital 2002; Liu & Cordes 2004; Sonesson & Meuwissen 2009; Yue 2014). The South African abalone industry, which centres around a single abalone species, *Haliotis midae*, is an emerging, but already highly successful aquaculture production industry (Raemaekers *et al*. 2011). While the development of molecular genetic resources for this species has progressed considerably in recent years (Roodt-Wilding & Brink 2011), no large scale LD data has yet been generated, the addition of which would provide a number of opportunities for both population genetic research and the development of molecular genetic tools for future enhancement strategies in cultured

abalone. As such, the primary aim of the current study was to quantify and characterise levels of genome-wide LD within the South African abalone, and also to demonstrate the various ways in which this genetic resource can be applied.

4.2) Summary and Synthesis of Results

4.2.1) Linkage disequilibrium in Haliotis midae

The second chapter of the current study investigated the levels of LD across the genome of wild and cultured cohorts of the commercially important mollusc species, *Haliotis midae*. This represented the first of such studies in a member of the *Haliotis* genus, and also provided a unique opportunity to investigate how the demographic and micro-evolutionary factors associated with natural life history characteristics and domestication events have affected LD across the abalone genome. In general, small effective population size appears to have had a large impact on LD within both wild and cultured populations of *H. midae*. Levels of LD were found to be unusually high, particularly as compared with other prominent aquaculture species, such as Atlantic salmon (Moen *et al*. 2008) and the silver-lipped pearl oyster (Jones *et al*. 2013). However, given the reproductive strategy of abalone, these inflated levels of LD were not entirely unexpected. Unequal parental contributions during spawning are a common occurrence in both wild and cultured populations of broadcast spawning species, such as abalone (Lee & Boulding 2009; Lind *et al*. 2009; Slabbert *et al*. 2009; Hedgecock & Pudovkin 2011). Subsequently, individual spawning events are characterised by small short-term effective population sizes, exacerbating the sampling effect of genetic drift, and thus inflating LD through reduced genetic diversity (Rhode 2013). Furthermore, as skewed parental contributions are known to produce both spatial and temporal population structure, the occurrence of gene flow between otherwise isolated populations in the wild, as well as the contribution of overlapping generations during spawning, would result in continual “admixture-like” events, and thus further inflate levels of LD (Rhode 2013). However, in addition to the life history characteristics of abalone, anthropogenic factors are also likely to have had a significant impact on effective population size and LD in wild and cultured populations of abalone. The continued removal of large numbers of breeding individuals from wild populations for the purposes of both legal (fisheries and aquaculture) and illegal (poaching) exploitation would have caused severe reductions in effective population size (Raemaekers *et al*. 2011; Rhode 2013). A similar effect would have been perpetuated in the cultured

populations by the collection of a finite number of broodstock individuals from wild populations for the establishment of cultured stock.

In terms of locus-specific effects on LD, evidence of various micro-evolutionary processes was also observed in both wild and cultured cohorts of abalone. In particular, levels of LD across individual linkage groups were observed to be highly heterogeneous; while some linkage groups were dominated by one or two large blocks of significant marker-marker associations (*e.g.* LG1), others had much smaller blocks spread through-out the entire length of the chromosome (*e.g.* LG5). Although differing inter-marker distances could account for some of the heterogeneity, regions of consistently low LD could also indicate the locations of so-called recombination "hot-spots", while regions of particularly extensive LD might represent corresponding recombination "cold-spots" (Slate & Pemberton 2007; Li & Merilä 2010a). Alternatively, the latter could also represent the effects of selection on one or more of the affected loci by means of genetic “hitch-hiking” [*i.e.* the co-segregation of closely linked loci with an advantageous variant (Maynard Smith & Haigh 1974; McVean 2007)], as many of the markers involved were also highlighted as candidate loci under selection. While it is difficult to determine which factor is primarily responsible in each case (local recombination rates across the abalone genome have not been extensively characterised), the explanation of selection would be consistent with the observation of persistent patterns of long-range LD linking blocks of local LD (*e.g.* LG5 and -8), as this could represent the effects of selection on a group of functionally linked genes. Such "gene networks" have been observed in an increasing number of species, and would be consistent with the complex genetic architecture of most biologically important traits (Mackay 2001; Goddard & Hayes 2009; Yue 2014).

4.2.2) Contributions of natural- and artificial selection during domestication

In addition to furthering our understanding of the processes that shape diversity within biological populations, another key endeavour of genetics is the identification and characterisation of the genetic elements that make up traits of interest (Feder & Mitchell-Olds 2003). In this regard, understanding the role of selection in generating phenotypic differences between populations can be of great importance. In particular, the identification of genetic regions that are involved in adaptation *via* natural selection, or that could be targeted by artificial selection, can provide important clues in the search for genes determining biologically and economically important traits (Feder & Mitchell-Olds 2003;

Storz 2005; Gholami *et al*. 2015). These candidate regions can then be further investigated using association analyses to discover genotype-phenotype correlations (Long & Langley 1999; Mackay 2001; Stinchcombe & Hoekstra 2007).

To this effect, Rhode *et al*. (2013) identified a selection of microsatellite markers within the South African abalone believed to be under divergent selection between wild and cultured populations. Although domestication of this species is still within the early stages as compared with other cultured halotids (Li *et al*. 2004; Praipue *et al*. 2010), significant population differentiation has already been observed between wild and cultured (F_1 and F_2) populations (*e.g.* Slabbert *et al*. 2009; Rhode *et al*. 2012, 2014). While such differentiation has often been attributed to the demographic factors associated with the domestication event, it has been suggested that a significant proportion of the differentiation observed between these populations could also be due to the locus-specific effects of selection (Vasemägi *et al*. 2012; Rhode *et al*., 2013). As crude phenotypic selection for faster growth rate is practiced in these populations, regions under selection could represent loci associated with growth rate, and thus be of interest for future applications in marker-assisted breeding programmes. However, in addition to artificial selection pressures, two forms of natural selection could also be at play during domestication, namely, relaxed natural selection for survival in the wild, and increased natural selection for adaptation to the novel aquaculture environment.

Using a candidate locus association analysis, the third chapter of the current study investigated the relative contributions of natural and artificial selection during the domestication of abalone by determining whether the loci identified by Rhode *et al*. (2013) are associated with larger size (a proxy for growth rate) within a first generation commercial population. For this purpose, a selection of statistical tests was used to detect significant genotype-phenotype associations within a family-bias corrected population cohort, and two family cohorts. Significant associations were observed for two of the 13 loci, both of which could also be putatively linked with known genes potentially involved in growth and development. However, as significant associations for these markers were not observed across all cohorts, it is possible that these loci do not represent the causative variants for larger size (Hindorff *et al*. 2009). Indeed, the extent of significant LD over relatively far distances in cultured abalone, as determined by the previous chapter, would suggest that significant genotype-phenotype associations could well arise as a result of LD between a marker locus and the true causal variant.

Regarding the loci that were not significantly associated with size, bioinformatics analyses revealed that many of these loci could also be linked with known genes. As all of the markers used within the current study demonstrated divergent selection between wild and cultured abalone populations, this could indicate possible associations with traits involved in adaptation to the aquaculture environment. Furthermore, as the vast majority of candidate loci (85%) were found not to be significantly associated with size within the study cohorts, it was suggested that natural selection, rather than artificial selection, is the predominant force shaping genetic variation during domestication, and thus driving the divergence of cultured abalone populations from their wild progenitors. In particular, possible examples of both positive natural selection in response to novel selective pressures, as well as relaxed natural selection for traits no longer required for survival within the new environment, were provided. Interestingly, possible evidence of the latter was also observed based on differences in patterns of LD between the wild and cultured cohorts within the previous chapter. Distinctive patterns of both local and long-range LD between putative loci under selection on LG8 (as determined by the current study) suggested that multiple loci on this linkage group could be involved with traits important for survival in the wild, as these patterns were present to a much lesser extent, or absent, within the cultured cohorts.

4.2.3) Association studies in Haliotis midae

Linkage disequilibrium- or association mapping represents a faster and more economical alternative to linkage mapping as a means of identifying and locating QTLs for traits of interest (Meadows *et al*. 2008). However, as this approach is based on the likelihood that one or more study markers will be located close enough to a causal variant to be in LD with it, the success of such a study is heavily dependent on taking existing levels of LD within the study cohorts into account (Gaut & Long 2003; Stinchcombe & Hoekstra 2007; Goddard & Hayes 2009). While less pronounced than in many other species (Moen *et al.* 2008; Rexroad & Vallejo 2009; Jones *et al.* 2013), the current study found that the decay of LD within cultured abalone cohorts was nonetheless evident. The LD mapping approach is based on the assumption that LD decays with genetic distance, and it is therefore important to first confirm that the target population/s adhere to this expectation (Slate & Pemberton 2007). As a similar trend was not also observed within the wild cohorts, in

which generations overlap, it was suggested that association studies should primarily be focused towards generationally discrete cohorts obtained from aquaculture facilities.

In terms of the extent of LD, it was observed that significant associations were prevalent over considerable distances within both cultured cohorts (up to 20 cM), suggesting that relatively few markers would be required for implementing a genome-wide association (GWA) study (Slate & Pemberton 2007; Goddard & Hayes 2009; Li & Merilä 2010b). However, as the presence of large blocks of LD would likely result in similarly large confidence intervals for identified QTLs, regardless of marker density, the GWA approach would prove much less effective (Goddard & Hayes 2009; Li & Merilä 2010b). In contrast, more extensive LD might prove beneficial in applying the candidate locus approach, as QTLs of interest would be more likely to co-segregate with candidate markers (Stinchcombe & Hoekstra 2007; Li & Merilä 2010b). In terms of the current study, this factor may have played a significant role in the successful detection of genotype-phenotype associations within the candidate locus association analysis, particularly as the associated markers appeared more likely to be adjacent to the respective causal variants, rather than the functional variants themselves. As this approach relies more on the quality of candidate loci than the number of markers used to ensure successful detection of QTLs, less extensive LD may have resulted in fewer, if any, associations being detected. However, regardless of which approach is used to map QTLs to a particular genomic region, such extensive LD would nonetheless prove prohibitive in efforts to fine-map QTLs and locate true causal variants (Li & Merilä 2010b).

4.3) Shortcomings and Future Research

The current study reported on the levels of LD within wild and cultured abalone cohorts obtained from two of the three primary geographical regions associated with the natural range of *H. midae* (*i.e.* the east and west coasts of South Africa). While this provides reasonable predictions for levels of LD within other populations of abalone with similar demographic histories, the characterisation of LD in additional wild and cultured populations could be highly informative in terms of providing corroborative evidence for many of the current study's findings. Furthermore, as the south coast represents a natural transition zone between the east- and west coasts, characterising levels of LD within south coast populations would provide an opportunity to further investigate the effects of continued admixture between genetically distinct east- and west coast populations (Bester-

van der Merwe *et al*. 2011). As admixture tends to inflate LD, it would be expected that higher and more extensive levels would be maintained within these “transitional” populations relative to the more isolated populations. The continued effects of domestication on the abalone genome could also be investigated by characterising LD in future generations of cultured abalone. As some aquaculture facilities have begun to close the reproductive cycle by replacing wild broodstock with superior F_1-generation individuals, second generation cohorts of cultured *H. midae* have recently become available (Rhode *et al*. 2014). In particular, it would be of interest to determine whether levels of genome-wide LD continue to increase, remain stable, or decrease in successive generations. As levels of existing LD are expected to be broken down by recombination events in the following generations, it is possible that both the magnitude and extent of LD might be decreased relative to the current study results (Mueller 2004; Slate & Pemberton 2007). However, as closing the reproductive cycle can result in higher levels of inbreeding, exacerbated by skewed parental contributions, the resulting decrease in effective population size might also serve to increase LD (Slatkin 2008). Based on an assessment of genetic diversity in an F_2 cohort by Rhode *et al*. (2014), it seems likely that the latter will have had a significant impact, as estimates of effective population size were noticeably decreased as compared with the progenitor F_1 cohort. Furthermore, as both natural and artificial selective pressures would have had additional opportunity to influence the genome, locus-specific patterns of LD might reveal additional signatures of selection surrounding loci determining biologically and economically important traits. In this regard, it would be of particular interest to include the candidate loci under selection that were investigated within the third chapter, although this would first require their integration into the *H. midae* linkage map (only four of the thirteen loci are currently mapped) (Vervalle *et al*. 2013).

In terms of marker saturation, the current study was somewhat limited by the use of only microsatellite markers, despite a large number of SNP markers also being mapped. As a result, many of the linkage groups were poorly saturated (*e.g.* LG11, -13 and -16), which severely limited, or even precluded, more in-depth investigations of LD on these linkage groups. As such, future studies on LD in abalone would benefit greatly from the incorporation of all available mapped markers, as well as the integration of additional markers to fill large inter-marker distances (*e.g.* the ~40 cM gap on LG2). Sample size is also acknowledged to have been a limiting factor within the current study, particularly given the use of microsatellite markers. Ruzzante (1998) demonstrated that a sample size lower than 50 is more likely to result in sampling errors, and consequently bias population

genetic estimates of diversity and differentiation, as well as LD. The highly polymorphic nature of microsatellite markers can further exacerbate this effect, as large sample sizes are required to accurately reflect true levels of diversity, particularly in the event of rare alleles. However, in cases where samples are limiting, the severity of such biases can be reduced by standardising sample sizes and using a large number of markers (Rhode 2013), both of which were implemented within the current study.

Regarding the association analysis, rather than performing a genome-wide association study, the current study utilised a candidate locus approach to determine whether natural- or artificial selection is predominant during the early stages of abalone domestication. In contrast with a genome-wide approach, the candidate locus approach focuses on only a few loci that have already shown potential in terms of functional significance, either as known genes, or through statistical analyses (*e.g.* F_{ST}-outlier tests) (Stinchcombe & Hoekstra 2007). As both approaches have advantages and disadvantages, the choice of which is more appropriate is largely dependent on the resources available (Stinchcombe & Hoekstra 2007; Goddard & Hayes 2009). Studies in species for which fewer resources are available, such as *H. midae*, would tend to benefit more from the candidate locus approach, as the limitation of fewer markers can be overcome by focusing studies on only the most promising loci. While this may not have provided a comprehensive view of the effects of selection during domestication, the targeted approach nevertheless proved highly successful in that promising candidates for both natural- and artificial selection were identified. As faster growth rate is currently the primary trait of economic importance in abalone aquaculture, the loci found to be significantly associated with size might therefore find application in marker-assisted selection programmes for *H. midae*. While these loci are not confirmed to represent true causal variants for larger size, markers in significant LD with causal variants could still be utilised effectively as evaluation markers for traits of interest (Dekkers 2004; Liu & Cordes 2004). However, with such an application in mind, it would be of great importance to not only validate these associations in independent populations and subsequent generations, but also to account for the manner and extent to which non-genetic factors within the aquaculture environment, such as dissolved oxygen levels and nutrient availability, are contributing towards variation in growth rate. In the most extreme case, individuals with poor genetic potential may still perform as well as, or better than, those with supposedly higher potential if their environmental conditions are more favourable, which could potentially confound efforts to identify causal variants and determine their true effect sizes. Lastly, as a more comprehensive understanding of how

adaptation is taking place within the aquaculture environment might create additional opportunities for the selective manipulation of cultured stock *via* improved husbandry practices, future studies using the markers putatively associated with natural selection might therefore focus on determining which stage/s of abalone development are the most influential in terms of adaptation potential.

4.4) Final Remarks

The current study aimed to characterise LD within the commercially important marine mollusc, *Haliotis midae*, and further, to demonstrate its utility within the context of population genetic investigations and characterising the genetic architecture of complex traits. While some conclusions were speculative, a number of findings represent intriguing directions for possible future research in abalone. The first experimental chapter aimed to investigate, and therefore demonstrate, the various ways in which LD data can be used to study the demographic and evolutionary properties of a population within the context of abalone, and in this regard it was highly successful. Because it is sensitive to both demographic and locus-specific processes, the addition of LD data provides a unique perspective on the interpretation of many of the more commonly used population genetic estimates, such as genetic diversity, effective population size and neutrality of loci, and it is therefore a worthwhile estimate to include in any population genetic investigation. While the first experimental chapter investigated LD directly, the second experimental chapter demonstrated how the phenomenon of LD can be exploited indirectly to detect genotype-phenotype associations between candidate markers under selection and a primary trait of economic importance (growth rate). Using *H. midae* as an opportunistic model, the results demonstrated, putatively, the effects of both natural- and artificial selection on an animal population that is in the initial phases of domestication. In addition, this chapter also served to highlight the efficacy of the candidate locus approach in a species with both limited molecular resources and extensive LD.

References

Bentsen H.B. & Olesen I. (2002) Designing aquaculture mass selection programs to avoid high inbreeding rates. *Aquaculture* 204, 349-359.

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

Dekkers J.C.M. (2004) Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. *Journal of Animal Science* 82, E-Suppl:E313-328.

Dekkers J.C.M. & Hospital F. (2002) The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics* 3, 22-32.

Feder M.E. & Mitchell-Olds T. (2003) Evolutionary and ecological functional genomics. *Nature Reviews Genetics* 4, 649-699.

Flint-Garcia S.A., Thornsberry J.M. & Buckler IV E.S. (2003) Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* 54, 357-374.

Gaut B.S. & Long A.D. (2003) The lowdown on linkage disequilibrium. *The Plant Cell* 15, 1502-1506.

Gholami M., Reimer C., Erbe M., Preisinger R., Weigend A., Weigend S., Servin B. & Simianer H. (2015) Genome scan for selection in structured layer chicken populations exploiting linkage disequilibrium information. *PLoS ONE* 10, e0130497.

Goddard M.E. & Hayes B.J. (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics* 10, 381-391.

Hedgecock D. & Pudovkin A.I. (2011) Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. *Bulletin of Marine Science* 87, 971-1002.

Hindorff L.A., Sethupathy P., Junkins H.A., Ramos E.M., Mehta J.P., Collins F.S. & Manolio T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* 106, 9362-9367.

Jones D.B., Jerry D.R., Khatkar M.S., Raadsma H.W. & Zenger K.R. (2013) A high-density SNP genetic linkage map for the silver-lipped pearl oyster, *Pinctada maxima*: a valuable resource for gene localisation and marker-assisted selection. *BMC Genomics* 14, 810.

Knapp S.J. (1998) Marker-assisted selection as a strategy for increasing the probability of selecting superior genotypes. *Crop Science* 38, 1164-1174.

Lee H.J.E. & Boulding E.G. (2009) Spatial and temporal population genetic structure of four northeastern Pacific littorinid gastropods: the effect of mode of larval development on variation at one mitochondrial and two nuclear DNA markers. *Molecular Ecology* 18, 2165-2184.

Li M.-H. & Merilä J. (2010a) Sex-specific population structure, natural selection, and linkage disequilibrium in a wild bird population as revealed by genome-wide microsatellite analyses. *BMC Evolutionary Biology* 10, 66.

Li M.-H. & Merilä J. (2010b) Extensive linkage disequilibrium in a wild bird population. *Heredity* 104, 600-610.

Li Q., Park Q., Endo T. & Kijima A. (2004) Loss of genetic variation at microsatellite loci in hatchery strains of the Pacific abalone (*Haliotis discus hannai*). *Aquaculture* 235, 207-222.

Lind C.E., Evans B.S., Knauer J., Taylor J.J.U. & Jerry D.R. (2009) Decreased genetic diversity and a reduced effective population size in cultured silver-lipped pearl oysters (*Pinctada maxima*). *Aquaculture* 286, 12-19.

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

Long A.D. & Langley C.H. (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research* 9, 720-731.

Mackay T.F.C. (2001) The genetic architecture of quantitative traits. *Annual Review of Genetics* 35, 303-339.

Maynard Smith J. & Haigh J. (1974) The hitch-hiking effect of a favourable gene. *Genetical Research* 23, 23-35.

McVean G. (2007) The structure of linkage disequilibrium around a selective sweep. *Genetics* 175, 1395-1406.

Moen T., Hayes B., Baranski M., Berg P.R., Kjøglum S., Koop B.F., Davidson W.S., Omholt S.W. & Lien S. (2008) A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers. *BMC Genomics* 9, 223.

Mueller J.C. (2004) Linkage disequilibrium for different scales and applications. *Briefings in Bioinformatics* 5, 355-364.

Praipue P., Klinbunga S. & Jarayabhand P. (2010) Genetic diversity of wild and domesticated stocks of Thai abalone, *Haliotis asinina* (Haliotidae), analyzed by single-strand conformational polymorphism of AFLP-derived markers. *Genetics and Molecular Research* 9, 1136-1152.

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.

Rexroad C.E. & Vallejo R.L. (2009) Estimates of linkage disequilibrium and effective population size in rainbow trout. *BMC Genetics* 10, 83.

Rhode C. (2013) Signatures of selection in natural and cultured abalone (*Haliotis midae*): A population genomics study. Unpublished PhD thesis, Stellenbosch University, Stellenbosch.

Rhode C., Hepple J., Jansen S., Davis T., Vervalle J., Bester-van der Merwe A.E. & Roodt-Wilding R. (2012) A population genetics analysis of abalone domestication events in South Africa: Implications for the management of the abalone resource. *Aquaculture* 356-357, 235-242.

Rhode C., Maduna S.N., Roodt-Wilding R. & Bester-van der Merwe A.E. (2014) Comparison of population genetic estimates amongst wild, F1 and F2 cultured abalone (*Haliotis midae*). *Animal Genetics* 45, 456-459.

Rhode C., Vervalle J., Bester-van der Merwe A.E. & Roodt-Wilding R. (2013) Detection of molecular signatures of selection at microsatellite loci in the South African abalone (*Haliotis midae*) using a population genomic approach. *Marine Genomics* 10, 27-36.

Roodt-Wilding R. & Brink D. (2011) Selection and sea snails: the South African story. *Philosophical Transactions in Genetics* 1, 1-40.

Ruzzante D.E. (1998) A comparison of several measures of genetic distance and population structure with microsatellite data: bias and sampling variance. *Canadian Journal of Fisheries and Aquatic Sciences* 55, 1-14.

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

Slate J. & Pemberton J.M. (2007) Admixture and patterns of linkage disequilibrium in a free-living vertebrate population. *Journal of Evolutionary Biology* 20, 1415-1427.

Slatkin M. (2008) Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* 9, 477-485.

Sonesson A.K. & Meuwissen T.H. (2009) Testing strategies for genomic selection in aquaculture breeding programs. *Genetics Selection Evolution* 41, 37.

Stinchcombe J.R. & Hoekstra H.E. (2007) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* 100, 158-170.

Storz J.F. (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* 14, 671-688.

Vasemägi A., Nilsson J., McGinnity P., Cross T., O'Reilly P., Glebe B., Peng B., Berg P.R. & Primmer C.R. (2012) Screen for footprints of selection during domestication/captive breeding of Atlantic salmon. *Comparative and Functional Genomics*, 1-14.

Vervalle J., Hepple J., Jansen S., du Plessis J., Wang P., Rhode C. & Roodt-Wilding R. (2013) Integrated linkage map of *Haliotis midae* Linnaeus based on microsatellite and SNP markers. *Journal of Shellfish Research* 32, 89-103.

Yue G.H. (2014) Recent advances of genome mapping and marker-assisted selection in aquaculture. *Fish and Fisheries* 15, 376-396.

Appendix A

Supplementary Information for Chapter 2

Figure S2.1: Integrated linkage map of the South African abalone, *Haliotis midae* (Vervalle *et al*. 2013), with the 112 microsatellite markers used in the current study highlighted.

Table S2.1: Basic marker information for the 112 microsatellite markers used in the current study. Marker name, linkage group, repeat motif, accession number and source are indicated.

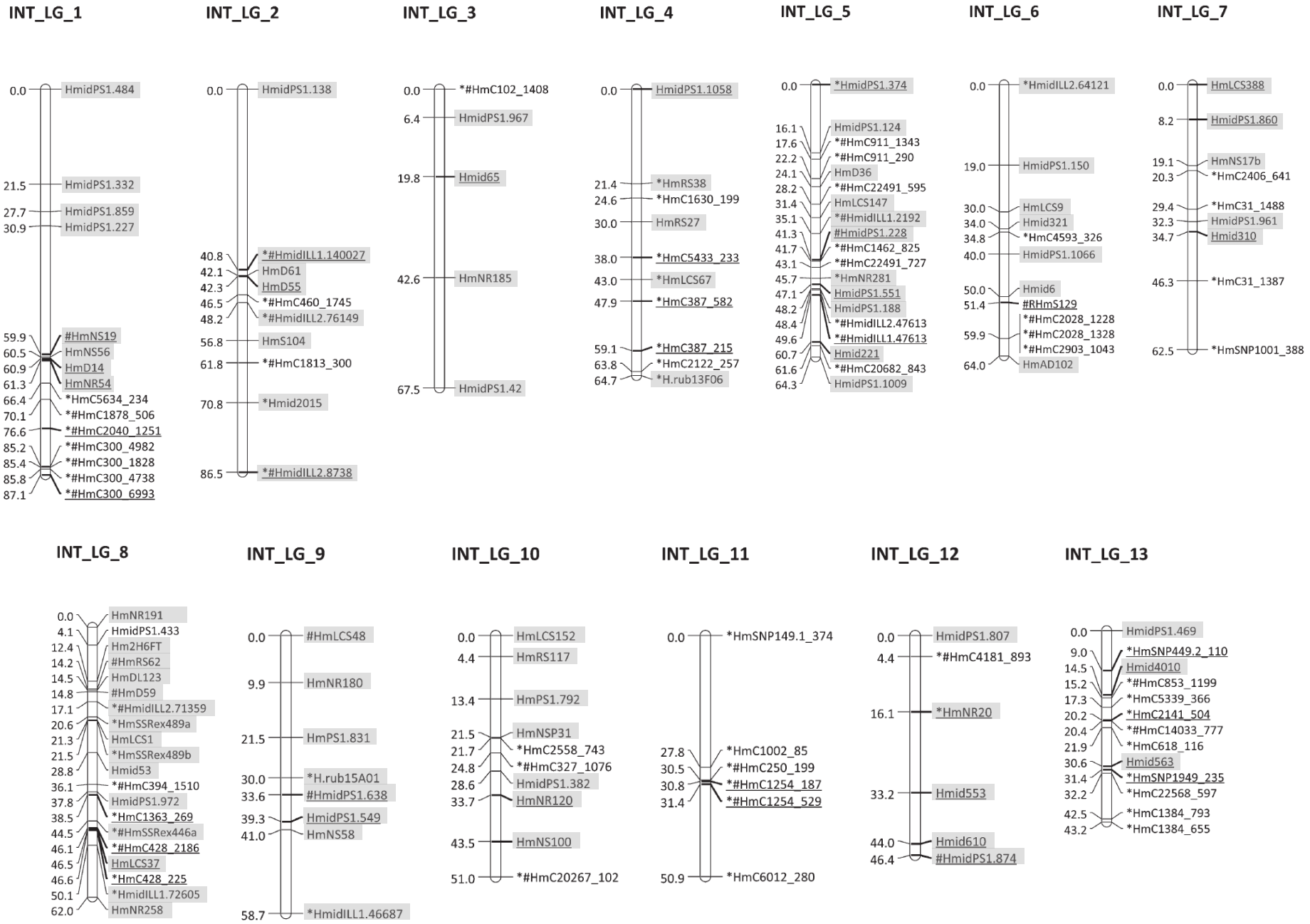
Table S2.2: The subset of unlinked markers used to estimate effective population size for each cohort. Markers were chosen based on adherence to Hardy-Weinberg equilibrium and level of diversity.

Table S2.3: The subset of markers used to calculate the two baseline levels of LD within each cohort. Markers were chosen based on adherence to Hardy-Weinberg equilibrium and level of diversity. Additional markers from some linkage groups were included to provide the most accurate estimate possible. Pairwise comparisons between markers on the same linkage group were eliminated before calculating the baseline values.

Table S2.4: Basic diversity statistics per marker for each cohort. These include: sample size (N), number of alleles per marker (A_n), effective number of alleles per marker (A_e), Shannon's Information Index values (I), observed heterozygosity (H_o), unbiased expected heterozygosity (uH_e), F_{IS} values, and Hardy-Weinberg *P*-values.

Table S2.5: Estimated null allele frequencies, and 0.025- and 0.975 confidence bounds for each marker within each cohort.

Table S2.6: Significant *P*-values for candidate loci under selection, as indicated by the Ewens-Watterson and F_{ST}-outlier tests. The Ewens-Watterson test was conducted for each cohort separately, and interpreted as a two-tailed test against the alternative hypotheses of balancing ($P < 0.025$) or directional ($P > 0.975$) selection (significant at the 5% level). For the F_{ST}-outlier test, cohorts were grouped according to region to facilitate comparisons between wild and cultured cohorts within the respective groups (F_{SC}), as well as between all cohorts (F_{ST}). Significant F_{ST}/F_{CT} values ($P < 0.05$) that exceeded the 0.95 percentile expected value indicated directional selection, while values lower than the 0.05 percentile expected value indicated balancing selection. Candidates for balancing selection are underlined, while candidates for directional selection are in bold. Dashes indicate non-significant values.



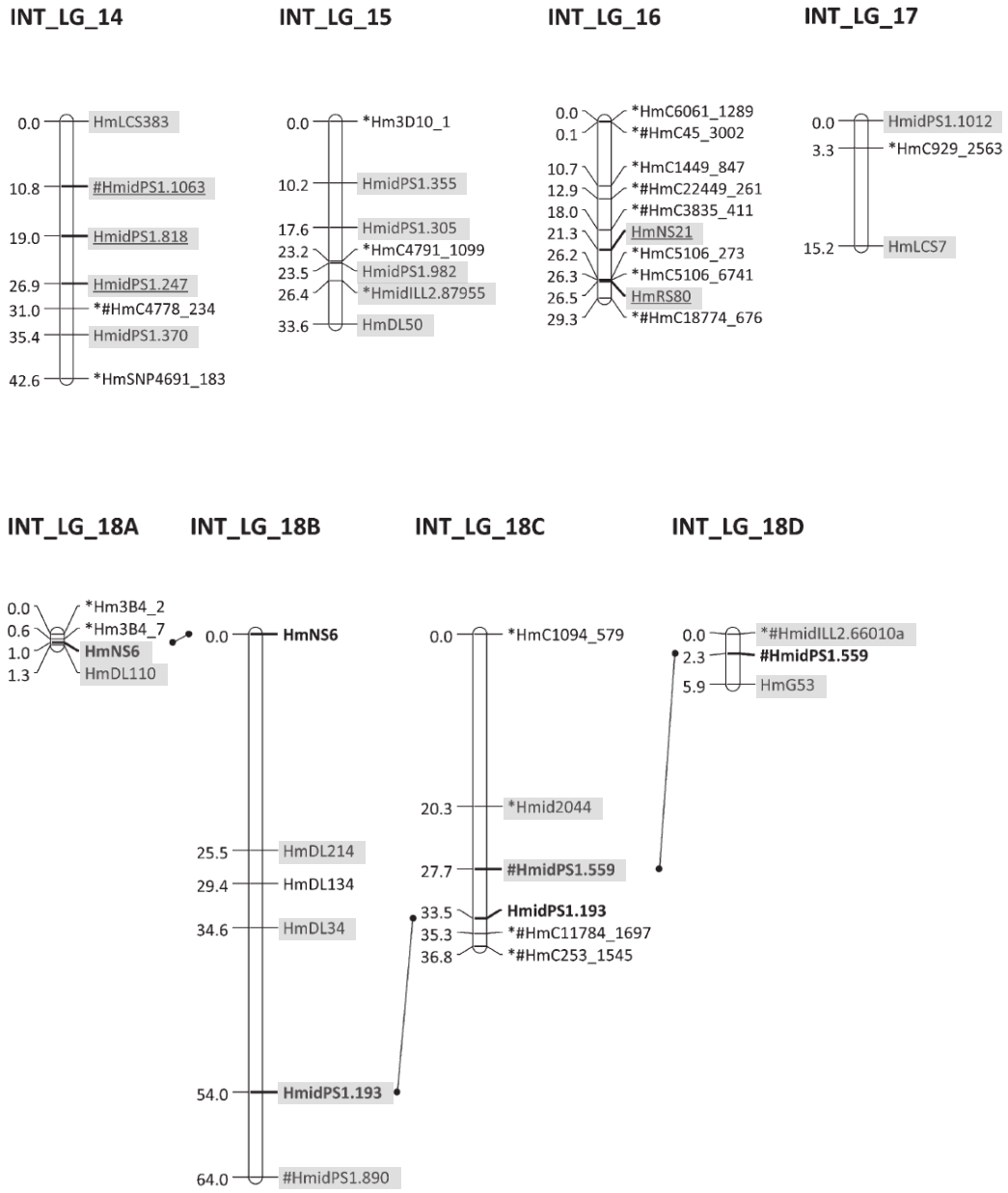


Figure S2.1: Integrated linkage map of the South African abalone, *Haliotis midae* (Vervalle *et al*. 2013), with the 112 microsatellite markers used in the current study highlighted.

Table S2.1: Basic marker information for the 112 microsatellite markers used in the current study. Marker name, linkage group, repeat motif, accession number and source are indicated.

Linkage Group	Marker	Repeat motif	Accession	Source
1	*HmidPS1.484*	$(GAGT)_n...(GTGA)_n...(GTGA)_n$	GU256693	Slabbert *et al*. 2012
	HmidPS1.332	$(AC)_n$	GU256680	Slabbert *et al*. 2012
	HmidPS1.859	$(CTCA)_n$	GU256715	Slabbert *et al*. 2012
	HmidPS1.227	$(ATGT)_n$	GU256676	Slabbert *et al*. 2012
	HmNS19L	$(AACACCC)_9$	EF033330	Slabbert *et al*. 2008
	HmNS56	$(CA)_{20}$	EF455619	Slabbert *et al*. 2008
	HmD14	$(CA)_{10}$	AY303333	Bester *et al*. 2004
	HmNR54	$(TTAGGG)_4$	EF063103	Slabbert *et al*. 2008
2	*HmPS1.138*	$(CA)_n$	GU256660	Slabbert *et al*. 2012
	HmidILL1.140027	$(AGGGC)_3$	JN793422	Rhode *et al*. 2012
	HmD61	$(CA)_{24}$	AY303340	Bester *et al*. 2004
	HmD55	$(GTGA)_{12}$	AY303337	Bester *et al*. 2004
	HmidILL2.76149	AT	JN793427	Rhode *et al*. 2012
	HmS104	$(GAGT)_n$	GQ927137	Slabbert *et al*. 2010
	Hmid2015	$(TG)_n(TGTC)_n$	GQ927124	Slabbert *et al*. 2010
	HmidILL2.8738	$(TGT)_n(TGC)$	JN793445	Rhode *et al*. 2012
3	*HmidPS1.967*	$(TGTC)_n(TG)_n$	GU256725	Slabbert *et al*. 2012
	Hmid65	$(CT)_n...(AC)_n...(AC)_n...(AC)_n$	GQ927111	Slabbert *et al*. 2010
	HmNR185	$(GT)_{13}$	EF121750	Slabbert *et al*. 2008
	HmidPS1.42	$(ATCC)_n…(ATCC)_n$	GU256657	Slabbert *et al*. 2012
4	*HmidPS1.1058*	$(TGAG)_n…(AGTG)_n…(AGTG)_n$	GU256735	Slabbert *et al*. 2012
	HmRS38	$(GT)_{14}(GA)_9$	DQ785755	Slabbert *et al*. 2008
	HmRS27	$(TCAC)_{30}$	DQ785751	Slabbert *et al*. 2008
	HmLCS67	$(GAGT)_3(GT)_5(GC)_4$	DQ993222	Slabbert *et al*. 2008
	H.rub13F06	$(GT)_n$	DQ278037	Rhode 2010
5	*HmidPS1.374*	$(GAGT)_n$	GU256684	Slabbert *et al*. 2012
	HmidPS1.124	$(AC)_n$	GU256659	Slabbert *et al*. 2012
	HmD36	$(GTGA)_{14}$	AY303335	Bester *et al*. 2004
	HmLCS147	$(GAGT)_n$	GQ927134	Slabbert *et al*. 2010
	HmidILL1.2192	$(ATAC)_4$	JN793413	Rhode *et al*. 2012
	HmidPS1.228	$(ACTC)_n$	GU256677	Slabbert *et al*. 2012
	HmNR281P	$(CTCAA)_{24}$	EF512274	Slabbert *et al*. 2008
	HmidPS1.551	$(TATG)_n…(TGTA)_n$	GU256697	Slabbert *et al*. 2012
	HmidPS1.188	$(GTGC)_n…(GT)_n$	GU256668	Slabbert *et al*. 2012
	HmidILL1.47613	$(ACAG)_5$	JN793418	Rhode *et al*. 2012
	Hmid221	$(ACAG)_n$	GQ927115	Slabbert *et al*. 2010
	HmidPS1.1009	$(GTGGGT)_n$	GU256730	Slabbert *et al*. 2012
6	*HmidILL2.64121*	CCT	JN793435	Rhode *et al*. 2012
	HmidPS1.150	$(CA)_n…(CA)_n…(CACT)_n…(CA)_n…(CA)_n$	GU256662	Slabbert *et al*. 2012
	HmLCS9M	$(GC)_2(GT)_2(GCGTGT)_2(GCGT)_2(GC)$	DQ993214	Slabbert *et al*. 2008
	Hmid321	$(GT)_n...(GT)_n$	GQ927121	Slabbert *et al*. 2010

	HmidPS1.1066	$(GT)_n(TG)_n(TGTT)_n$	GU256737	Slabbert *et al*. 2012
	Hmid6	$(ACAT)_n(AC)_n(ACAT)_n$	GQ927108	Slabbert *et al*. 2010
	HmRS129	$(GT)_{15}$	DQ785766	Slabbert *et al*. 2008
	HmAD102	$(ACTC)_{15}$	DQ785747	Slabbert *et al*. 2008
	HmLCS388	$(GCGT)_n(GTT)_n(GT)_n$	GQ927140	Slabbert *et al*. 2010
	HmidPS1.860	$(GT)_n$	GU256716	Slabbert *et al*. 2012
7	*HmNS17b*	$(CACT)_{31}$	EF367116	Van den Berg 2008
	HmidPS1.961	$(GTAG)_n$	GU256724	Slabbert *et al*. 2012
	Hmid310	$(GT)_n$	GQ927119	Slabbert *et al*. 2010
	HmNR191	$(GAGT)_6$	EF121752	Slabbert *et al*. 2008
	Hm2H6FT	$(CACT)_n$	GQ927136	Slabbert *et al*. 2010
	HmRS62	$(GT)_{12}$	DQ785777	Slabbert *et al*. 2008
	HmDL123	$(CT)_{20}$	EF054865	Slabbert *et al*. 2008
	HmD59	$(CA)_{15}$	AY303338	Bester *et al*. 2004
	HmidILL2.71359	GCAT	JN793441	Rhode *et al*. 2012
	HmSSRex489a	CACT	ex534489	Rhode 2010
8	*HmLCS1*	$(CGTG)_6$	DQ825701	Slabbert *et al*. 2008
	HmSSRex489b	ACTC	ex534489	Rhode 2010
	Hmid53	$(GT)_n$	GQ927110	Slabbert *et al*. 2010
	HmidPS1.972	$(TCAC)_n$	GU256726	Slabbert *et al*. 2012
	HmSSRex446a	GTGA	ex534446	Rhode 2010
	HmLCS37	$(GA)_{13}(CA)(GA)_8(CA)(GA)_4$	DQ993229	Slabbert *et al*. 2008
	HmidILL1.72605	$(AGGTG)_4$		Hepple 2010
	HmNR258	$(CAA)_{11}$	EF512272	Slabbert *et al*. 2008
	HmLCS48	$(CT)_{14}(CA)_9$	DQ993227	Slabbert *et al*. 2008
	HmNR180	$(GT)_{24}$	EF121748	Slabbert *et al*. 2008
	HmidPS1.831	$(CACC)_n(CACT)_n(CACC)_n(CACT)_n(CAC)_n$ $(CACT)_n$	GU256712	Slabbert *et al*. 2012
9	*H.rub15A01*	CAGA	DQ278045	Rhode 2010
	HmidPS1.638	$(GTGA)_n$	GU256703	Slabbert *et al*. 2012
	HmidPS1.549	$(TG)_n$	GU256696	Slabbert *et al*. 2012
	HmNS58	$(GTT)_8$	EF367119	Slabbert *et al*. 2008
	HmidILL1.46687	$(TGAG)_4$		Hepple 2010
	HmLCS152	$(CAA)_n(CCA)_n(CTA)_n(CCA)_n$	GQ927139	Slabbert *et al*. 2010
	HmRS117	$(GAGT)_{33}(GCGT)_3$	DQ785765	Slabbert *et al*. 2008
	HmPS1.792	$(CA)_n$	JX853745	Vervalle *et al*. 2013
10	*HmNSP31*	$(CAA)_n(CAG)_n(CAA)_n$	EU126856	Rhode 2010
	HmidPS1.382	$(TG)_n$	GU256687	Slabbert *et al*. 2012
	HmNR120	$(TGAG)_{23}$	EF121745	Slabbert *et al*. 2008
	HmNS100	$(GAGT)_{16}$	EF367114	Slabbert *et al*. 2008
	HmidPS1.807	$(GAGT)_n$	GU256709	Slabbert *et al*. 2012
	HmNR20	$(TCC)_5(TAC)_7$	EF063097	Slabbert *et al*. 2008
12	*Hmid553*	$(GT)_n$	GQ927122	Slabbert *et al*. 2010
	Hmid610	$(GT)_n$	GQ927118	Slabbert *et al*. 2010

	HmidPS1.874	$(CACG)_n…(AC)_n$	GU256720	Slabbert *et al*. 2012
13	*HmidPS1.469*	$(ATC)_n$	GU256692	Slabbert *et al*. 2012
	Hmid4010	$(AC)_n$	GQ927131	Slabbert *et al*. 2010
	Hmid563	$(AG)_n(ACAT)(AG)_n(ACAGAG)_n(ACAG)_n$ $(ACTG)(ACAG)_n$	GQ927117	Slabbert *et al*. 2010
14	*HmLCS383*	$(GTGA)_n$	GQ927141	Slabbert *et al*. 2010
	HmidPS1.1063	$(TC)_n…(CGTG)_n$	GU256736	Slabbert *et al*. 2012
	HmidPS1.818	$(ATGG)_n…(TGGA)_n…(AC)_n$	GU256711	Slabbert *et al*. 2012
	HmidPS1.247	$(GA)_n(GAAT)_n$	GU256678	Slabbert *et al*. 2012
	HmidPS1.370	$(CAACC)_n…(CACT)_n$	GU256683	Slabbert *et al*. 2012
15	*HmidPS1.355*	$(TCA)_n$	GU256682	Slabbert *et al*. 2012
	HmidPS1.305	$(GCAC)_n$	GU256679	Slabbert *et al*. 2012
	HmidPS1.982	$(TGTA)_n(TG)_n$	GU256728	Slabbert *et al*. 2012
	HimidILL2.87955	GTGA	JN793442	Rhode *et al*. 2012
	HmDL50	$(TGTC)_{11}(GGTC)_6$	EF054861	Slabbert *et al*. 2008
16	*HmNS21*	$(CT)_n(C)_n(CT)_n(T)_n(CT)_n$	GQ927143	Hepple 2010
	HmRS80	$(GAGT)_{17}(GA)_3(GAGT)$	DQ785756	Slabbert *et al*. 2008
17	*HmidPS1.1012*	$(CAT)_n$	GU256731	Slabbert *et al*. 2012
	HmLCS7	$(GT)_7(GCGT)_6(GT)_7$	DQ825707	Slabbert *et al*. 2008
18A	*HmNS6*	$(ACGC)_6$	EF367117	Slabbert *et al*. 2008
	HmDL110	$(TCAC)_{23}$	EF054864	Slabbert *et al*. 2008
18B	*HmNS6*	$(ACGC)_6$	EF367117	Slabbert *et al*. 2008
	HmDL214	$(TGAG)_{15}$	EF054871	Slabbert *et al*. 2008
	HmDL34b	$(CAGA)_{16}$	EF054860	Slabbert *et al*. 2008
	HmidPS1.193	$(ACTC)_n…(TCAC)_n…(CACT)_n…(CACT)_n…$ $(TCAC)_n$	GU256669	Slabbert *et al*. 2012
	HmidPS1.890	$(CACT)_n(CT)_n$	GU256721	Slabbert *et al*. 2012
18C	*Hmid2044*	$(GAGT)_n$	GQ927126	Slabbert *et al*. 2010
	HmPS1.559	$(CA)_n(G)(TCAC)_n$	GU256698	Slabbert *et al*. 2012
	HmidPS1.193	$(ACTC)_n…(TCAC)_n…(CACT)_n…(CACT)_n…$ $(TCAC)_n$	GU256669	Slabbert *et al*. 2012
18D	*HmidILL2.66010a*	ATT	JN793436	Rhode *et al*. 2012
	HmPS1.559	$(CA)_n(G)(TCAC)_n$	GU256698	Slabbert *et al*. 2012
	HmG53T	$(CACT)_{31}$	DQ785746	Slabbert *et al*. 2008

Table S2.2: The subset of unlinked markers used to estimate effective population size for each cohort. Markers were chosen based on adherence to Hardy-Weinberg equilibrium and level of diversity.

Linkage Group	Marker	Cohort	A_e	H_o	uH_e	HW *P*-value
1	*HmNR54*	ASF	6.020	0.867	0.848	0.280
		SAL	6.006	0.875	0.847	0.494
		WCA	5.721	0.862	0.840	0.624
		RP	7.394	0.875	0.878	0.433
2	*HmidILL1.140027*	ASF	2.038	0.484	0.518	0.710
		SAL	2.000	0.355	0.508	0.148
		WCA	1.923	0.533	0.488	0.712
		RP	1.969	0.500	0.500	1.000
4	*HmidPS1.1058*	ASF	5.128	0.933	0.819	0.886
		SAL	6.161	0.828	0.852	0.240
		WCA	6.621	0.875	0.867	0.475
		RP	6.755	0.828	0.867	0.241
5	*HmidPS1.551*	ASF	2.410	0.667	0.595	0.723
		SAL	2.234	0.552	0.562	0.696
		WCA	3.003	0.613	0.678	0.845
		RP	3.436	0.625	0.720	0.052
6	*HmidILL2.64121*	ASF	1.697	0.407	0.419	1.000
		SAL	1.394	0.233	0.288	0.067
		WCA	1.471	0.267	0.325	0.304
		RP	1.685	0.433	0.414	0.619
7	*HmidPS1.860*	ASF	3.003	0.742	0.678	0.960
		SAL	2.839	0.677	0.658	0.363
		WCA	2.681	0.594	0.637	0.594
		RP	2.594	0.677	0.625	0.984
8	*HmidILL2.71359*	ASF	2.181	0.594	0.550	0.658
		SAL	1.958	0.563	0.497	0.656
		WCA	3.136	0.781	0.692	0.195
		RP	2.253	0.533	0.566	0.521
9	*HmidPS1.638*	ASF	1.381	0.281	0.280	0.576
		SAL	1.334	0.281	0.254	1.000
		WCA	1.833	0.469	0.462	0.324
		RP	1.466	0.375	0.323	1.000
10	*HmLCS152*	ASF	5.389	0.750	0.827	0.349
		SAL	4.921	0.870	0.814	0.774
		WCA	11.571	0.938	0.928	0.695
		RP	12.013	1.000	0.932	0.326
13	*Hmid563*	ASF	8.430	0.821	0.897	0.019
		SAL	8.048	0.885	0.893	0.536
		WCA	6.961	0.826	0.875	0.753
		RP	8.491	0.867	0.897	0.114
14	*HmidPS1.818*	ASF	6.828	0.808	0.870	0.227
		SAL	6.461	0.813	0.859	0.493
		WCA	6.453	0.636	0.865	0.069

		RP	6.870	0.767	0.869	0.285
17	*HmidPS1.1012*	ASF	6.541	0.818	0.887	0.129
		SAL	7.688	0.742	0.884	0.090
		WCA	4.455	0.714	0.835	0.659
		RP	8.219	0.867	0.893	0.948
18A	*HmNS6*	ASF	2.576	0.581	0.622	0.623
		SAL	4.171	0.688	0.772	0.184
		WCA	3.852	0.613	0.753	0.068
		RP	4.582	0.656	0.794	0.092

Table S2.3: The subset of markers used to calculate the two baseline levels of LD within each cohort. Markers were chosen based on adherence to Hardy-Weinberg equilibrium and level of diversity. Additional markers from some linkage groups were included to provide the most accurate estimate possible. Pairwise comparisons between markers on the same linkage group were eliminated before calculating the baseline values.

Linkage Group	Marker	Cohort	A_e	H_o	uH_e	HW *P*-value
1	*HmD14*	ASF	2.538	0.656	0.616	0.902
		SAL	2.573	0.548	0.621	0.171
		WCA	2.485	0.594	0.607	0.837
		RP	4.146	0.875	0.771	0.950
	HmNR54	ASF	6.020	0.867	0.848	0.280
		SAL	6.006	0.875	0.847	0.494
		WCA	5.721	0.862	0.840	0.624
		RP	7.394	0.875	0.878	0.433
	HmNS56	ASF	2.195	0.531	0.553	0.333
		SAL	2.488	0.563	0.608	0.156
		WCA	2.476	0.531	0.606	0.416
		RP	3.765	0.781	0.746	0.901
2	*HmidILL1.140027*	ASF	2.038	0.484	0.518	0.710
		SAL	2.000	0.355	0.508	0.148
		WCA	1.923	0.533	0.488	0.712
		RP	1.969	0.500	0.500	1.000
4	*HmRS38*	ASF	1.275	0.200	0.219	0.216
		SAL	1.215	0.188	0.180	1.000
		WCA	1.550	0.281	0.361	0.145
		RP	1.141	0.129	0.125	1.000
	HmRS27	ASF	15.515	0.906	0.950	0.170
		SAL	23.814	0.969	0.973	0.802
		WCA	16.490	0.966	0.956	0.093
		RP	20.480	0.969	0.966	0.633
	HmidPS1.1058	ASF	5.128	0.933	0.819	0.886
		SAL	6.161	0.828	0.852	0.240
		WCA	6.621	0.875	0.867	0.475
		RP	6.755	0.828	0.867	0.241
5	*HmidPS1.124*	ASF	1.074	0.071	0.070	1.000
		SAL	1.242	0.219	0.198	1.000
		WCA	1.492	0.344	0.335	1.000
		RP	1.591	0.375	0.377	0.191
	HmidPS1.551	ASF	2.410	0.667	0.595	0.723
		SAL	2.234	0.552	0.562	0.696
		WCA	3.003	0.613	0.678	0.845
		RP	3.436	0.625	0.720	0.052
	HmidPS1.228	ASF	2.266	0.643	0.579	1.000
		SAL	3.075	0.548	0.686	0.480
		WCA	3.789	0.750	0.768	0.232
		RP	3.156	0.645	0.694	0.131
6	*HmidILL2.64121*	ASF	1.697	0.407	0.419	1.000
		SAL	1.394	0.233	0.288	0.067

		WCA	1.471	0.267	0.325	0.304
		RP	1.685	0.433	0.414	0.619
7	*HmidPS1.860*	ASF	3.003	0.742	0.678	0.960
		SAL	2.839	0.677	0.658	0.363
		WCA	2.681	0.594	0.637	0.594
		RP	2.594	0.677	0.625	0.984
	HmLCS388	ASF	1.582	0.455	0.385	1.000
		SAL	2.076	0.533	0.527	1.000
		WCA	2.174	0.400	0.600	0.616
		RP	1.929	0.500	0.490	1.000
8	*HmSSRex446a*	ASF	2.190	0.531	0.552	0.126
		SAL	2.267	0.677	0.568	0.808
		WCA	2.495	0.625	0.609	0.751
		RP	3.821	0.750	0.750	0.493
	Hm2H6F	ASF	2.430	0.677	0.598	0.578
		SAL	2.635	0.571	0.632	0.756
		WCA	3.094	0.656	0.688	1.000
		RP	2.462	0.625	0.603	0.725
	HmD59	ASF	11.636	0.875	0.929	0.079
		SAL	10.503	0.813	0.919	0.057
		WCA	8.866	0.813	0.901	0.181
		RP	9.225	0.844	0.906	0.371
	HmNR258	ASF	4.046	0.839	0.765	0.210
		SAL	4.104	0.719	0.768	0.192
		WCA	3.640	0.742	0.737	0.798
		RP	4.214	0.781	0.775	0.472
	HmidILL2.71359	ASF	2.181	0.594	0.550	0.658
		SAL	1.958	0.563	0.497	0.656
		WCA	3.136	0.781	0.692	0.195
		RP	2.253	0.533	0.566	0.521
	HmSSRex489b	ASF	2.571	0.667	0.626	0.939
		SAL	2.563	0.563	0.620	0.769
		WCA	2.692	0.739	0.643	0.829
		RP	2.421	0.645	0.597	0.868
	HmSSRex489a	ASF	2.276	0.536	0.571	0.830
		SAL	2.519	0.531	0.613	0.758
		WCA	2.660	0.656	0.634	1.000
		RP	2.376	0.621	0.589	0.962
9	*HmidPS1.638*	ASF	1.381	0.281	0.280	0.576
		SAL	1.334	0.281	0.254	1.000
		WCA	1.833	0.469	0.462	0.324
		RP	1.466	0.375	0.323	1.000
10	*HmLCS152*	ASF	5.389	0.750	0.827	0.349
		SAL	4.921	0.870	0.814	0.774
		WCA	11.571	0.938	0.928	0.695
		RP	12.013	1.000	0.932	0.326

	HmPS1.792	ASF	1.247	0.214	0.202	1.000
		SAL	1.032	0.031	0.031	-
		WCA	1.065	0.063	0.062	1.000
		RP	1.100	0.094	0.092	1.000
13	*Hmid4010*	ASF	3.638	0.563	0.737	0.323
		SAL	3.729	0.593	0.746	0.192
		WCA	4.878	0.700	0.808	0.334
		RP	3.838	0.733	0.752	0.117
14	*HmLCS383*	ASF	2.482	0.594	0.607	0.583
		SAL	3.013	0.742	0.679	0.264
		WCA	3.867	0.667	0.755	0.363
		RP	2.977	0.625	0.675	0.844
	HmidPS1.818	ASF	6.828	0.808	0.870	0.227
		SAL	6.461	0.813	0.859	0.493
		WCA	6.453	0.636	0.865	0.069
		RP	6.870	0.767	0.869	0.285
	HmidPS1.247	ASF	2.922	0.594	0.668	0.623
		SAL	3.325	0.839	0.711	0.177
		WCA	3.314	0.719	0.709	0.589
		RP	3.187	0.645	0.698	0.098
17	*HmidPS1.1012*	ASF	6.541	0.818	0.887	0.129
		SAL	7.688	0.742	0.884	0.090
		WCA	4.455	0.714	0.835	0.659
		RP	8.219	0.867	0.893	0.948
18A	*HmNS6*	ASF	2.576	0.581	0.622	0.623
		SAL	4.171	0.688	0.772	0.184
		WCA	3.852	0.613	0.753	0.068
		RP	4.582	0.656	0.794	0.092

Table S2.4: Basic diversity statistics per marker for each cohort, per linkage group. These include: sample size (N), number of alleles per marker (A_n), effective number of alleles per marker (A_e), Shannon's Information Index values (I), observed heterozygosity (H_o), unbiased expected heterozygosity (uH_e), F_{IS} values, and Hardy-Weinberg *P*-values.

Cohort	Linkage Group	Marker	N	A_n	A_e	I	H_o	uH_e	F_{IS}	HW *P*-value
ASF	1	*HmD14*	32	13	2.538	1.523	0.656	0.616	-0.083	0.902
		HmidPS1.484	25	4	1.799	0.843	0.080	0.453	0.820	0.000
		HmNR54	30	12	6.020	2.025	0.867	0.848	-0.039	0.280
		HmNS56	32	9	2.195	1.252	0.531	0.553	0.024	0.333
		HmNS19	27	21	15.848	2.888	0.741	0.955	0.209	0.000
		HmidPS1.332	27	16	8.055	2.348	0.778	0.892	0.112	0.303
		HmidPS1.859	32	9	3.094	1.487	0.688	0.688	-0.016	0.295
		HmidPS1.227	30	4	2.151	0.919	0.767	0.544	-0.433	0.033
	2	*Hmid2015*	32	10	6.302	2.009	0.750	0.855	0.109	0.145
		HmD55	32	13	8.258	2.305	0.594	0.893	0.324	0.000
		HmidILL1.140027	31	3	2.038	0.758	0.484	0.518	0.050	0.710
		HmidILL2.76149	29	4	3.150	1.221	0.759	0.694	-0.111	0.944
		HmidILL2.8738	30	5	3.340	1.341	0.833	0.712	-0.190	0.038
		HmidPS1.138	26	7	3.449	1.472	0.308	0.724	0.567	0.000
		HmD61	29	14	5.128	2.015	0.828	0.819	-0.028	0.855
	3	*HmidPS1.42*	28	2	1.461	0.495	0.179	0.321	0.434	0.040
		HmidPS1.967	32	6	4.039	1.515	0.969	0.764	-0.287	0.030
		HmNR185	31	14	8.076	2.318	0.935	0.891	-0.068	0.980
		Hmid65	30	18	10.778	2.606	0.767	0.923	0.155	0.000
	4	*HmLCS67*	32	5	2.065	0.972	0.500	0.524	0.030	0.881
		HmRS38	30	5	1.275	0.509	0.200	0.219	0.072	0.216
		HmRS27	32	28	15.515	3.013	0.906	0.950	0.031	0.170
		HmidPS1.1058	30	14	5.128	2.066	0.933	0.819	-0.159	0.886
	5	*HmidPS1.1009*	28	9	5.640	1.876	0.643	0.838	0.219	0.020
		HmLCS147	30	10	4.800	1.827	0.267	0.805	0.663	0.000
		HmD36	32	16	11.838	2.606	0.719	0.930	0.215	0.024
		HmidPS1.374	32	9	3.977	1.663	0.531	0.760	0.290	0.001
		HmidPS1.124	28	2	1.074	0.154	0.071	0.070	-0.037	1.000
		HmidPS1.188	28	3	1.244	0.409	0.214	0.200	-0.091	1.000
		HmNR281	19	15	9.256	2.469	0.474	0.916	0.469	0.000
		HmidILL1.2192	32	2	1.519	0.525	0.063	0.347	0.817	0.000
		HmidILL1.47613	32	17	6.206	2.261	0.875	0.852	-0.043	0.740
		HmidPS1.551	30	4	2.410	1.027	0.667	0.595	-0.140	0.723
		HmidPS1.228	14	3	2.266	0.924	0.643	0.579	-0.151	1.000
		Hmid221	31	11	6.792	2.110	0.677	0.867	0.206	0.001
	6	*Hmid6*	29	23	17.340	2.984	0.345	0.959	0.634	0.000
		HmidPS1.150	32	28	12.190	2.980	0.781	0.933	0.149	0.017
		HmidPS1.1066	18	10	6.894	2.114	0.111	0.879	0.870	0.000
		HmLCS9	32	9	2.216	1.292	0.438	0.558	0.203	0.020
		HmAD102	32	27	18.789	3.108	0.688	0.962	0.274	0.000
		HmRS129	32	23	12.337	2.799	0.719	0.934	0.218	0.000
		HmidILL2.64121	27	3	1.697	0.659	0.407	0.419	0.008	1.000
		Hmid321	30	9	3.905	1.642	0.533	0.756	0.283	0.000
	7	*HmidPS1.961*	24	5	1.629	0.806	0.125	0.395	0.676	0.000

	HmNS17b	16	17	12.488	2.697	0.688	0.950	0.253	0.006
	Hmid310	28	15	9.679	2.437	0.714	0.913	0.203	0.001
	HmidPS1.860	31	5	3.003	1.218	0.742	0.678	-0.112	0.960
	HmLCS388	11	3	1.582	0.650	0.455	0.385	-0.236	1.000
8	*HmidPS1.972*	32	8	3.039	1.393	0.594	0.682	0.115	0.000
	HmSSRex446a	32	4	2.190	0.973	0.531	0.552	0.022	0.126
	HmDL123	26	17	14.383	2.743	0.500	0.949	0.463	0.000
	Hmid53	32	6	1.936	1.045	0.500	0.491	-0.034	0.772
	Hm2H6F	31	4	2.430	1.113	0.677	0.598	-0.151	0.578
	HmLCS37	0	0	0.000	0.000	0.000	0.000	N/A	-
	HmLCS1	32	7	3.136	1.398	0.688	0.692	-0.009	0.224
	HmD59	32	16	11.636	2.581	0.875	0.929	0.043	0.079
	HmRS62	31	17	10.010	2.544	0.742	0.915	0.176	0.001
	HmNR191	23	7	5.482	1.807	0.565	0.836	0.309	0.000
	HmNR258	31	5	4.046	1.458	0.839	0.765	-0.114	0.210
	HmidILL2.71359	32	4	2.181	1.024	0.594	0.550	-0.096	0.658
	HmidILL1.72605	31	1	1.000	0.000	0.000	0.000	N/A	-
	HmSSRex489b	21	4	2.571	1.150	0.667	0.626	-0.091	0.939
	HmSSRex489a	28	4	2.276	1.065	0.536	0.571	0.044	0.830
9	*HmidPS1.638*	32	3	1.381	0.542	0.281	0.280	-0.019	0.576
	H.rub15A01	29	4	2.132	0.841	0.517	0.540	0.026	0.042
	HmidPS1.831	28	5	3.045	1.211	0.607	0.684	0.096	0.012
	HmLCS48	32	5	2.317	1.152	0.656	0.577	-0.155	0.595
	HmNR180	23	15	9.121	2.432	0.478	0.910	0.463	0.000
	HmNS58	28	14	8.209	2.299	0.607	0.894	0.309	0.000
	HmidPS1.549	31	10	4.699	1.802	0.710	0.800	0.098	0.449
10	*HmidPS1.382*	32	10	5.107	1.874	0.750	0.817	0.067	0.197
	HmNSP31	30	7	3.930	1.583	0.433	0.758	0.419	0.000
	HmNS100	27	10	6.284	2.017	0.481	0.857	0.427	0.000
	HmLCS152	32	13	5.389	2.093	0.750	0.827	0.079	0.349
	HmPS1.792	28	4	1.247	0.449	0.214	0.202	-0.080	1.000
	HmRS117	21	18	10.889	2.654	0.857	0.930	0.056	0.206
	HmNR120	30	12	6.870	2.123	0.767	0.869	0.103	0.025
12	*HmidPS1.807*	32	6	2.322	1.181	0.531	0.578	0.067	0.182
	HmNR20	25	17	10.776	2.575	0.760	0.926	0.162	0.004
	Hmid553	25	4	2.006	0.914	0.480	0.512	0.043	0.001
	Hmid610	22	11	4.610	1.915	0.591	0.801	0.245	0.106
	HmidPS1.874	30	21	12.676	2.770	0.833	0.937	0.095	0.222
13	*HmidPS1.469*	32	3	2.314	0.912	0.938	0.577	-0.651	0.000
	Hmid563	28	16	8.430	2.396	0.821	0.897	0.068	0.019
	Hmid4010	32	4	3.638	1.336	0.563	0.737	0.224	0.323
14	*HmLCS383*	32	3	2.482	0.978	0.594	0.607	0.006	0.583
	HmidPS1.818	26	12	6.828	2.114	0.808	0.870	0.054	0.227
	HmidPS1.1063	23	12	7.896	2.251	0.609	0.893	0.303	0.000
	HmidPS1.370	30	5	2.113	0.946	0.767	0.536	-0.456	0.028
	HmidPS1.247	32	4	2.922	1.160	0.594	0.668	0.097	0.623
15	*HmidPS1.355*	32	7	4.395	1.618	0.750	0.785	0.029	0.397
	HmidPS1.982	31	6	1.458	0.714	0.161	0.319	0.487	0.001

		HmidPS1.305	26	10	7.116	2.080	0.692	0.876	0.194	0.003
		HmDL50	31	18	9.707	2.538	0.774	0.912	0.137	0.081
		HimidILL2.87955	30	7	4.000	1.590	0.767	0.763	-0.022	0.893
	16	*HmNS21*	31	18	11.373	2.620	0.903	0.927	0.010	0.638
		HmRS80	32	16	10.779	2.541	0.875	0.922	0.036	0.079
	17	*HmLCS7*	31	8	2.164	1.230	0.258	0.547	0.520	0.000
		HmidPS1.1012	11	9	6.541	2.008	0.818	0.887	0.034	0.129
	18A	*HmNS6*	31	6	2.576	1.190	0.581	0.622	0.051	0.623
		HmDL110	9	3	2.160	0.901	0.778	0.569	-0.448	0.339
	18B	*HmidPS1.890*	26	4	2.384	1.078	0.615	0.592	-0.060	0.925
		HmDL34b	30	13	8.257	2.270	0.633	0.894	0.279	0.003
		HmDL214	31	13	2.034	1.331	0.452	0.517	0.112	0.110
		HmidPS1.193	32	8	4.321	1.674	0.750	0.781	0.024	0.849
	18C	*HmidPS1.559*	27	7	2.008	1.094	0.333	0.512	0.336	0.006
		Hmid2044	32	14	6.302	2.149	0.344	0.855	0.591	0.000
	18D	*HmG53*	31	29	21.596	3.214	0.871	0.969	0.087	0.000
		HmidILL2.66010	29	2	1.665	0.589	0.483	0.407	-0.208	0.389
SAL	1	*HmD14*	31	15	2.573	1.649	0.548	0.621	0.103	0.171
		HmidPS1.484	29	5	1.884	0.939	0.207	0.477	0.559	0.000
		HmNR54	32	13	6.006	2.030	0.875	0.847	-0.050	0.494
		HmNS56	32	17	2.488	1.656	0.563	0.608	0.060	0.156
		HmNS19	32	28	18.124	3.120	0.969	0.960	-0.025	0.663
		HmidPS1.332	31	18	7.813	2.378	0.935	0.886	-0.073	0.724
		HmidPS1.859	26	8	1.980	1.102	0.423	0.505	0.145	0.427
		HmidPS1.227	29	5	2.113	0.981	0.655	0.536	-0.244	0.395
	2	*Hmid2015*	23	11	7.197	2.124	0.870	0.880	-0.010	0.946
		HmD55	30	17	9.326	2.478	0.700	0.908	0.216	0.030
		HmidILL1.140027	31	2	2.000	0.693	0.355	0.508	0.290	0.148
		HmidILL2.76149	32	5	2.813	1.162	0.469	0.655	0.273	0.004
		HmidILL2.8738	32	8	5.198	1.779	0.625	0.820	0.226	0.000
		HmidPS1.138	31	11	5.686	1.981	0.742	0.838	0.100	0.646
		HmD61	27	13	5.380	2.043	0.630	0.829	0.227	0.000
	3	*HmidPS1.42*	32	3	1.373	0.512	0.313	0.276	-0.151	1.000
		HmidPS1.967	32	7	3.765	1.500	0.750	0.746	-0.021	0.110
		HmNR185	31	13	7.813	2.278	0.677	0.886	0.223	0.026
		Hmid65	29	22	15.431	2.892	0.724	0.952	0.226	0.000
	4	*HmLCS67*	29	5	1.738	0.779	0.241	0.432	0.431	0.001
		HmRS38	32	6	1.215	0.457	0.188	0.180	-0.061	1.000
		HmRS27	32	28	23.814	3.241	0.969	0.973	-0.011	0.802
		HmidPS1.1058	29	14	6.161	2.197	0.828	0.852	0.012	0.240
	5	*HmidPS1.1009*	31	7	3.546	1.443	0.484	0.730	0.326	0.013
		HmLCS147	21	11	5.478	1.987	0.381	0.837	0.534	0.000
		HmD36	25	15	5.682	2.202	0.640	0.841	0.223	0.020
		HmidPS1.374	26	6	1.815	0.990	0.269	0.458	0.400	0.002
		HmidPS1.124	32	2	1.242	0.345	0.219	0.198	-0.123	1.000
		HmidPS1.188	32	5	1.477	0.667	0.250	0.328	0.225	0.265
		HmNR281	32	23	14.027	2.848	0.563	0.943	0.394	0.000
		HmidILL1.2192	32	4	1.860	0.813	0.219	0.470	0.527	0.002

	HmidILL1.47613	32	14	8.866	2.391	0.906	0.901	-0.021	0.940
	HmidPS1.551	29	6	2.234	1.135	0.552	0.562	0.001	0.696
	HmidPS1.228	31	6	3.075	1.330	0.548	0.686	0.187	0.480
	Hmid221	29	9	4.558	1.740	0.862	0.794	-0.104	0.562
6	*Hmid6*	25	26	20.492	3.128	0.680	0.971	0.285	0.000
	HmidPS1.150	25	23	12.376	2.836	0.720	0.938	0.217	0.043
	HmidPS1.1066	20	11	6.400	2.102	0.050	0.865	0.941	0.000
	HmLCS9	29	5	2.393	1.064	0.310	0.592	0.467	0.001
	HmAD102	30	31	23.684	3.288	0.633	0.974	0.339	0.000
	HmRS129	31	19	12.242	2.701	0.548	0.933	0.403	0.000
	HmidILL2.64121	30	3	1.394	0.505	0.233	0.288	0.175	0.067
	Hmid321	31	9	2.464	1.383	0.613	0.604	-0.032	0.837
7	*HmidPS1.961*	18	4	1.815	0.877	0.222	0.462	0.505	0.007
	HmNS17b	31	22	14.672	2.882	0.903	0.947	0.031	0.654
	Hmid310	27	14	8.055	2.293	0.741	0.892	0.154	0.130
	HmidPS1.860	31	7	2.839	1.310	0.677	0.658	-0.046	0.363
	HmLCS388	30	4	2.076	0.825	0.533	0.527	-0.029	1.000
8	*HmidPS1.972*	30	5	2.161	0.966	0.533	0.546	0.007	1.000
	HmSSRex446a	31	5	2.267	1.041	0.677	0.568	-0.212	0.808
	HmDL123	22	17	12.410	2.648	0.545	0.941	0.407	0.000
	Hmid53	22	8	2.645	1.426	0.591	0.636	0.050	0.469
	Hm2H6F	28	5	2.635	1.194	0.571	0.632	0.079	0.756
	HmLCS37	31	31	19.220	3.190	0.839	0.964	0.115	0.000
	HmLCS1	30	8	2.247	1.274	0.500	0.564	0.099	0.005
	HmD59	32	18	10.503	2.558	0.813	0.919	0.102	0.057
	HmRS62	32	15	7.613	2.316	0.656	0.882	0.245	0.000
	HmNR191	32	16	6.649	2.277	0.719	0.863	0.154	0.000
	HmNR258	32	9	4.104	1.618	0.719	0.768	0.050	0.192
	HmidILL2.71359	32	5	1.958	0.968	0.563	0.497	-0.150	0.656
	HmidILL1.72605	32	4	1.136	0.299	0.125	0.122	-0.045	1.000
	HmSSRex489b	32	6	2.563	1.184	0.563	0.620	0.078	0.769
	HmSSRex489a	32	5	2.519	1.127	0.531	0.613	0.119	0.758
9	*HmidPS1.638*	32	3	1.334	0.496	0.281	0.254	-0.123	1.000
	H.rub15A01	21	3	2.172	0.847	0.476	0.553	0.118	0.041
	HmidPS1.831	32	6	4.223	1.569	0.500	0.775	0.345	0.000
	HmLCS48	32	10	2.727	1.441	0.531	0.643	0.161	0.123
	HmNR180	25	12	5.319	2.073	0.320	0.829	0.606	0.000
	HmNS58	32	13	8.000	2.271	0.750	0.889	0.143	0.055
	HmidPS1.549	28	13	6.125	2.135	0.643	0.852	0.232	0.000
10	*HmidPS1.382*	30	9	4.615	1.717	0.600	0.797	0.234	0.067
	HmNSP31	26	4	3.494	1.310	0.385	0.728	0.461	0.000
	HmNS100	30	11	7.500	2.163	0.600	0.881	0.308	0.001
	HmLCS152	23	12	4.921	1.998	0.870	0.814	-0.091	0.774
	HmPS1.792	32	2	1.032	0.080	0.031	0.031	-0.016	-
	HmRS117	32	22	13.299	2.833	0.906	0.939	0.020	0.559
	HmNR120	32	16	6.759	2.218	0.781	0.866	0.083	0.136
12	*HmidPS1.807*	32	4	2.557	1.081	0.688	0.619	-0.129	0.982
	HmNR20	32	17	8.866	2.437	0.906	0.901	-0.021	0.595

		Hmid553	26	2	1.210	0.317	0.192	0.177	-0.106	1.000
		Hmid610	26	13	7.953	2.276	0.731	0.891	0.164	0.032
		HmidPS1.874	24	18	12.800	2.684	0.708	0.941	0.232	0.000
	13	*HmidPS1.469*	32	3	2.404	0.963	0.938	0.593	-0.605	0.000
		Hmid563	26	14	8.048	2.359	0.885	0.893	-0.010	0.536
		Hmid4010	27	7	3.729	1.543	0.593	0.746	0.190	0.192
	14	*HmLCS383*	31	5	3.013	1.196	0.742	0.679	-0.111	0.264
		HmidPS1.818	32	11	6.461	2.080	0.813	0.859	0.039	0.493
		HmidPS1.1063	29	15	7.972	2.330	0.897	0.890	-0.025	0.614
		HmidPS1.370	29	8	2.704	1.317	0.931	0.641	-0.477	0.030
		HmidPS1.247	31	5	3.325	1.293	0.839	0.711	-0.199	0.177
	15	*HmidPS1.355*	28	6	3.431	1.444	0.536	0.721	0.244	0.022
		HmidPS1.982	31	4	1.394	0.571	0.129	0.287	0.543	0.002
		HmidPS1.305	29	10	5.551	1.964	0.793	0.834	0.033	0.157
		HmDL50	32	17	12.488	2.645	0.906	0.935	0.015	0.486
		HimidILL2.87955	31	10	4.631	1.780	0.613	0.797	0.218	0.000
	16	*HmNS21*	28	16	10.316	2.524	0.714	0.919	0.209	0.000
		HmRS80	32	21	13.563	2.806	0.969	0.941	-0.046	0.543
	17	*HmLCS7*	32	9	1.614	0.934	0.250	0.386	0.343	0.007
		HmidPS1.1012	31	10	7.688	2.133	0.742	0.884	0.147	0.090
	18A	*HmNS6*	32	8	4.171	1.626	0.688	0.772	0.096	0.184
		HmDL110	31	6	3.230	1.383	0.548	0.702	0.206	0.214
	18B	*HmidPS1.890*	20	4	2.319	1.026	0.700	0.583	-0.231	0.400
		HmDL34b	32	15	9.894	2.464	0.563	0.913	0.374	0.000
		HmDL214	32	7	1.491	0.759	0.344	0.334	-0.045	0.267
		HmidPS1.193	29	7	4.509	1.703	0.345	0.792	0.557	0.000
	18C	*HmidPS1.559*	31	8	1.640	0.933	0.290	0.397	0.256	0.082
		Hmid2044	30	13	5.070	2.034	0.467	0.816	0.419	0.000
	18D	*HmG53*	32	33	23.814	3.334	1.000	0.973	-0.044	0.937
		HmidILL2.66010	31	3	1.338	0.465	0.290	0.256	-0.151	1.000
WCA	1	*HmD14*	32	12	2.485	1.450	0.594	0.607	0.007	0.837
		HmidPS1.484	28	3	1.524	0.581	0.179	0.350	0.481	0.007
		HmNR54	29	9	5.721	1.916	0.862	0.840	-0.045	0.624
		HmNS56	32	11	2.476	1.415	0.531	0.606	0.109	0.416
		HmNS19	27	22	17.566	2.971	0.667	0.961	0.293	0.000
		HmidPS1.332	22	13	8.800	2.359	0.682	0.907	0.231	0.004
		HmidPS1.859	31	7	2.818	1.382	0.419	0.656	0.350	0.003
		HmidPS1.227	24	3	2.072	0.860	0.625	0.528	-0.208	0.067
	2	*Hmid2015*	32	10	6.827	2.056	0.688	0.867	0.195	0.002
		HmD55	29	15	6.184	2.130	0.690	0.853	0.177	0.097
		HmidILL1.140027	30	2	1.923	0.673	0.533	0.488	-0.111	0.712
		HmidILL2.76149	29	5	3.689	1.392	0.517	0.742	0.290	0.000
		HmidILL2.8738	29	6	4.258	1.573	0.759	0.779	0.009	0.408
		HmidPS1.138	21	8	3.645	1.589	0.429	0.743	0.409	0.001
		HmD61	28	17	7.649	2.359	0.643	0.885	0.260	0.001
	3	*HmidPS1.42*	32	2	1.280	0.377	0.188	0.222	0.143	0.395
		HmidPS1.967	31	5	3.606	1.390	0.806	0.735	-0.116	0.203
		HmNR185	32	17	9.394	2.516	0.750	0.908	0.161	0.007

	Hmid65	24	18	14.049	2.754	0.833	0.949	0.103	0.004
4	*HmLCS67*	31	3	1.981	0.819	0.355	0.503	0.284	0.056
	HmRS38	32	7	1.550	0.827	0.281	0.361	0.208	0.145
	HmRS27	29	23	16.490	2.940	0.966	0.956	-0.028	0.093
	HmidPS1.1058	24	11	6.621	2.119	0.875	0.867	-0.031	0.475
5	*HmidPS1.1009*	30	5	4.000	1.477	0.433	0.763	0.422	0.001
	HmLCS147	28	12	9.862	2.358	0.179	0.915	0.801	0.000
	HmD36	31	20	12.088	2.715	0.871	0.932	0.050	0.681
	HmidPS1.374	32	5	1.855	0.963	0.250	0.468	0.458	0.000
	HmidPS1.124	32	3	1.492	0.560	0.344	0.335	-0.043	1.000
	HmidPS1.188	32	5	1.493	0.732	0.313	0.335	0.053	0.319
	HmNR281	30	25	15.385	2.969	0.567	0.951	0.394	0.000
	HmidILL1.2192	31	4	1.661	0.691	0.065	0.405	0.838	0.000
	HmidILL1.47613	31	14	9.756	2.411	0.935	0.912	-0.042	0.317
	HmidPS1.551	31	5	3.003	1.278	0.613	0.678	0.081	0.845
	HmidPS1.228	12	5	3.789	1.434	0.750	0.768	-0.019	0.232
	Hmid221	30	9	6.061	1.927	0.900	0.849	-0.078	0.485
6	*Hmid6*	27	21	16.022	2.893	0.556	0.955	0.407	0.000
	HmidPS1.150	32	30	15.634	3.089	0.844	0.951	0.099	0.000
	HmidPS1.1066	14	7	5.765	1.847	0.000	0.857	1.000	0.000
	HmLCS9	31	7	1.918	1.080	0.323	0.487	0.326	0.004
	HmAD102	32	26	18.124	3.055	0.625	0.960	0.339	0.000
	HmRS129	26	20	13.386	2.778	0.577	0.943	0.376	0.000
	HmidILL2.64121	30	2	1.471	0.500	0.267	0.325	0.167	0.304
	Hmid321	32	8	3.108	1.385	0.656	0.689	0.032	0.458
7	*HmidPS1.961*	19	3	1.465	0.604	0.053	0.326	0.834	0.000
	HmNS17b	25	18	7.862	2.459	0.520	0.891	0.404	0.000
	Hmid310	18	14	10.623	2.486	0.889	0.932	0.019	0.192
	HmidPS1.860	32	5	2.681	1.220	0.594	0.637	0.053	0.594
	HmLCS388	5	3	2.174	0.898	0.400	0.600	0.259	0.616
8	*HmidPS1.972*	32	5	2.151	0.891	0.531	0.544	0.007	0.664
	HmSSRex446a	32	5	2.495	1.202	0.625	0.609	-0.043	0.751
	HmDL123	32	22	14.222	2.855	0.469	0.944	0.496	0.000
	Hmid53	32	7	2.229	1.178	0.500	0.560	0.093	0.004
	Hm2H6F	32	5	3.094	1.296	0.656	0.688	0.030	1.000
	HmLCS37	0	0	0.000	0.000	0.000	0.000	N/A	-
	HmLCS1	32	8	3.631	1.555	0.625	0.736	0.137	0.015
	HmD59	32	15	8.866	2.413	0.813	0.901	0.084	0.181
	HmRS62	30	16	6.522	2.278	0.667	0.861	0.213	0.000
	HmNR191	31	11	6.560	2.059	0.645	0.861	0.239	0.000
	HmNR258	31	5	3.640	1.421	0.742	0.737	-0.023	0.798
	HmidILL2.71359	32	5	3.136	1.289	0.781	0.692	-0.147	0.195
	HmidILL1.72605	32	2	1.064	0.139	0.000	0.062	1.000	0.016
	HmSSRex489b	23	4	2.692	1.150	0.739	0.643	-0.176	0.829
	HmSSRex489a	32	5	2.660	1.194	0.656	0.634	-0.052	1.000
9	*HmidPS1.638*	32	3	1.833	0.753	0.469	0.462	-0.031	0.324
	H.rub15A01	29	3	2.019	0.756	0.414	0.514	0.180	1.000
	HmidPS1.831	31	5	3.100	1.213	0.581	0.689	0.143	0.582

		HmLCS48	31	7	2.529	1.301	0.613	0.614	-0.014	0.019
		HmNR180	24	14	9.681	2.403	0.375	0.916	0.582	0.000
		HmNS58	30	11	9.091	2.278	0.833	0.905	0.064	0.086
		HmidPS1.549	31	12	5.555	1.960	0.710	0.833	0.135	0.118
	10	*HmidPS1.382*	32	8	4.481	1.705	0.719	0.789	0.075	0.435
		HmNSP31	31	6	3.263	1.339	0.323	0.705	0.535	0.000
		HmNS100	28	14	8.340	2.353	0.464	0.896	0.472	0.000
		HmLCS152	32	17	11.571	2.604	0.938	0.928	-0.026	0.695
		HmPS1.792	32	3	1.065	0.161	0.063	0.062	-0.024	1.000
		HmRS117	27	17	9.656	2.559	0.778	0.913	0.132	0.020
		HmNR120	28	12	5.279	1.949	0.536	0.825	0.339	0.000
	12	*HmidPS1.807*	32	4	2.384	1.070	0.469	0.590	0.193	0.004
		HmNR20	23	15	8.744	2.405	0.696	0.905	0.215	0.001
		Hmid553	29	4	1.715	0.717	0.483	0.424	-0.158	0.810
		Hmid610	28	12	5.744	2.024	0.643	0.841	0.222	0.013
		HmidPS1.874	30	21	12.500	2.774	0.967	0.936	-0.051	0.175
	13	*HmidPS1.469*	32	3	2.268	0.902	0.938	0.568	-0.677	0.000
		Hmid563	23	11	6.961	2.105	0.826	0.875	0.035	0.753
		Hmid4010	30	7	4.878	1.693	0.700	0.808	0.119	0.334
	14	*HmLCS383*	27	6	3.867	1.477	0.667	0.755	0.101	0.363
		HmidPS1.818	22	10	6.453	2.057	0.636	0.865	0.247	0.069
		HmidPS1.1063	15	12	6.000	2.078	0.667	0.862	0.200	0.299
		HmidPS1.370	31	6	1.903	1.026	0.581	0.482	-0.224	1.000
		HmidPS1.247	32	5	3.314	1.311	0.719	0.709	-0.029	0.589
	15	*HmidPS1.355*	32	6	3.612	1.433	0.625	0.735	0.136	0.159
		HmidPS1.982	32	3	1.503	0.592	0.219	0.340	0.346	0.073
		HmidPS1.305	21	8	5.618	1.874	0.667	0.842	0.189	0.087
		HmDL50	29	21	13.565	2.805	0.862	0.943	0.069	0.047
		HimidILL2.87955	29	7	3.689	1.567	0.690	0.742	0.054	0.415
	16	*HmNS21*	32	22	11.907	2.779	0.813	0.931	0.113	0.009
		HmRS80	30	17	10.345	2.532	0.800	0.919	0.114	0.017
	17	*HmLCS7*	32	9	1.681	1.003	0.313	0.412	0.229	0.103
		HmidPS1.1012	7	6	4.455	1.631	0.714	0.835	0.079	0.659
	18A	*HmNS6*	31	7	3.852	1.513	0.613	0.753	0.172	0.068
		HmDL110	5	6	5.000	1.696	0.400	0.889	0.500	0.012
	18B	*HmidPS1.890*	32	4	2.809	1.140	0.719	0.654	-0.116	0.012
		HmDL34b	31	14	9.376	2.396	0.516	0.908	0.422	0.000
		HmDL214	30	9	1.921	1.071	0.533	0.488	-0.112	0.093
		HmidPS1.193	31	10	5.195	1.866	0.581	0.821	0.281	0.004
	18C	*HmidPS1.559*	22	10	2.602	1.443	0.455	0.630	0.262	0.065
		Hmid2044	31	11	5.636	2.002	0.516	0.836	0.373	0.000
	18D	*HmG53*	31	28	18.132	3.115	1.000	0.960	-0.058	0.784
		HmidILL2.66010	22	2	1.658	0.586	0.455	0.406	-0.146	1.000
RP	1	*HmD14*	32	18	4.146	2.062	0.875	0.771	-0.153	0.950
		HmidPS1.484	31	4	1.693	0.752	0.194	0.416	0.527	0.001
		HmNR54	32	15	7.394	2.259	0.875	0.878	-0.012	0.433
		HmNS56	32	15	3.765	1.887	0.781	0.746	-0.064	0.901
		HmNS19	32	27	15.398	3.014	0.781	0.950	0.164	0.016

	HmidPS1.332	32	13	5.565	2.056	0.688	0.833	0.162	0.145
	HmidPS1.859	30	9	2.786	1.459	0.500	0.652	0.220	0.046
	HmidPS1.227	29	3	1.711	0.746	0.517	0.423	-0.245	0.507
2	*Hmid2015*	32	9	6.263	1.941	0.781	0.854	0.070	0.161
	HmD55	32	8	4.995	1.783	0.656	0.813	0.179	0.049
	HmidILL1.140027	32	2	1.969	0.685	0.500	0.500	-0.016	1.000
	HmidILL2.76149	31	6	3.280	1.397	0.742	0.707	-0.067	0.180
	HmidILL2.8738	32	6	3.779	1.505	0.625	0.747	0.150	0.278
	HmidPS1.138	31	16	7.119	2.310	0.774	0.874	0.099	0.088
	HmD61	26	14	6.500	2.257	0.538	0.863	0.364	0.004
3	*HmidPS1.42*	32	4	1.468	0.618	0.313	0.324	0.020	0.043
	HmidPS1.967	32	6	3.850	1.444	0.719	0.752	0.029	0.220
	HmNR185	31	12	7.877	2.209	0.774	0.887	0.113	0.239
	Hmid65	29	20	13.901	2.783	0.897	0.944	0.034	0.021
4	*HmLCS67*	32	5	1.886	0.871	0.375	0.477	0.202	0.144
	HmRS38	31	4	1.141	0.306	0.129	0.125	-0.046	1.000
	HmRS27	32	32	20.480	3.239	0.969	0.966	-0.018	0.633
	HmidPS1.1058	29	11	6.755	2.123	0.828	0.867	0.029	0.241
5	*HmidPS1.1009*	31	5	1.967	1.018	0.548	0.500	-0.115	0.517
	HmLCS147	31	13	7.392	2.274	0.258	0.879	0.702	0.000
	HmD36	32	20	10.723	2.624	0.781	0.921	0.138	0.028
	HmidPS1.374	32	5	2.745	1.256	0.375	0.646	0.410	0.001
	HmidPS1.124	32	4	1.591	0.752	0.375	0.377	-0.009	0.191
	HmidPS1.188	32	3	1.171	0.313	0.031	0.148	0.786	0.001
	HmNR281	29	30	21.564	3.237	0.621	0.970	0.349	0.000
	HmidILL1.2192	31	2	1.250	0.353	0.226	0.204	-0.127	1.000
	HmidILL1.47613	31	16	8.042	2.412	0.839	0.890	0.042	0.008
	HmidPS1.551	32	8	3.436	1.482	0.625	0.720	0.118	0.052
	HmidPS1.228	31	6	3.156	1.355	0.645	0.694	0.056	0.131
	Hmid221	32	9	6.169	1.946	0.938	0.851	-0.119	0.840
6	*Hmid6*	32	31	21.787	3.245	0.781	0.969	0.181	0.000
	HmidPS1.150	32	23	9.799	2.706	0.781	0.912	0.130	0.213
	HmidPS1.1066	15	7	3.600	1.569	0.200	0.747	0.723	0.000
	HmLCS9	32	6	2.351	1.116	0.250	0.584	0.565	0.000
	HmAD102	31	28	18.660	3.119	0.742	0.962	0.216	0.000
	HmRS129	29	20	12.647	2.747	0.586	0.937	0.363	0.000
	HmidILL2.64121	30	3	1.685	0.680	0.433	0.414	-0.066	0.619
	Hmid321	32	9	2.775	1.312	0.719	0.650	-0.124	0.045
7	*HmidPS1.961*	25	4	2.073	0.933	0.160	0.528	0.691	0.000
	HmNS17b	32	23	13.932	2.883	0.719	0.943	0.226	0.000
	Hmid310	29	12	6.893	2.117	0.724	0.870	0.153	0.185
	HmidPS1.860	31	6	2.594	1.220	0.677	0.625	-0.102	0.984
	HmLCS388	30	4	1.929	0.841	0.500	0.490	-0.038	1.000
8	*HmidPS1.972*	32	5	2.286	0.977	0.594	0.571	-0.056	0.540
	HmSSRex446a	32	9	3.821	1.571	0.750	0.750	-0.016	0.493
	HmDL123	31	25	18.843	3.060	0.613	0.962	0.353	0.000
	Hmid53	31	9	2.746	1.432	0.645	0.646	-0.015	0.372
	Hm2H6F	32	5	2.462	1.161	0.625	0.603	-0.053	0.725

	HmLCS37	32	31	20.687	3.220	0.906	0.967	0.048	0.000
	HmLCS1	32	8	3.205	1.539	0.688	0.699	0.001	0.458
	HmD59	32	19	9.225	2.568	0.844	0.906	0.054	0.371
	HmRS62	27	14	6.658	2.248	0.630	0.866	0.259	0.002
	HmNR191	23	11	6.046	2.090	0.478	0.853	0.427	0.000
	HmNR258	32	6	4.214	1.499	0.781	0.775	-0.024	0.472
	HmidILL2.71359	30	6	2.253	1.072	0.533	0.566	0.041	0.521
	HmidILL1.72605	30	1	1.000	0.000	0.000	0.000	N/A	-
	HmSSRex489b	31	5	2.421	1.138	0.645	0.597	-0.099	0.868
	HmSSRex489a	29	5	2.376	1.120	0.621	0.589	-0.072	0.962
9	*HmidPS1.638*	32	3	1.466	0.588	0.375	0.323	-0.180	1.000
	H.rub15A01	30	3	2.067	0.766	0.367	0.525	0.290	0.092
	HmidPS1.831	31	4	3.075	1.160	0.613	0.686	0.092	0.779
	HmLCS48	30	5	2.707	1.203	0.467	0.641	0.260	0.018
	HmNR180	28	13	8.522	2.320	0.286	0.899	0.676	0.000
	HmNS58	31	12	6.840	2.112	0.613	0.868	0.282	0.000
	HmidPS1.549	28	11	5.939	2.062	0.643	0.847	0.227	0.106
10	*HmidPS1.382*	32	9	4.830	1.794	0.625	0.806	0.212	0.010
	HmNSP31	32	7	3.657	1.512	0.281	0.738	0.613	0.000
	HmNS100	30	12	5.696	2.018	0.467	0.838	0.434	0.000
	HmLCS152	31	18	12.013	2.631	1.000	0.932	-0.091	0.326
	HmPS1.792	32	4	1.100	0.241	0.094	0.092	-0.032	1.000
	HmRS117	32	24	12.047	2.822	0.906	0.932	0.012	0.084
	HmNR120	30	13	7.031	2.155	0.667	0.872	0.223	0.002
12	*HmidPS1.807*	32	4	2.241	0.983	0.625	0.563	-0.129	0.739
	HmNR20	32	15	10.089	2.496	0.813	0.915	0.098	0.045
	Hmid553	25	5	2.049	0.896	0.440	0.522	0.141	0.341
	Hmid610	24	13	6.128	2.119	0.792	0.855	0.054	0.540
	HmidPS1.874	26	17	11.757	2.612	0.962	0.933	-0.051	0.245
13	*HmidPS1.469*	32	3	2.062	0.763	0.906	0.523	-0.759	0.000
	Hmid563	30	15	8.491	2.355	0.867	0.897	0.018	0.114
	Hmid4010	30	5	3.838	1.455	0.733	0.752	0.008	0.117
14	*HmLCS383*	32	5	2.977	1.210	0.625	0.675	0.059	0.844
	HmidPS1.818	30	10	6.870	2.074	0.767	0.869	0.103	0.285
	HmidPS1.1063	28	16	9.800	2.480	0.821	0.914	0.085	0.010
	HmidPS1.370	32	6	2.560	1.155	0.938	0.619	-0.538	0.001
	HmidPS1.247	31	4	3.187	1.256	0.645	0.698	0.060	0.098
15	*HmidPS1.355*	32	7	4.808	1.687	0.844	0.805	-0.065	0.824
	HmidPS1.982	30	3	1.265	0.408	0.100	0.213	0.523	0.014
	HmidPS1.305	30	10	6.767	2.065	0.767	0.867	0.100	0.284
	HmDL50	30	18	11.250	2.638	0.800	0.927	0.122	0.034
	HimidILL2.87955	32	8	3.651	1.557	0.625	0.738	0.139	0.493
16	*HmNS21*	32	24	16.384	2.959	0.844	0.954	0.101	0.050
	HmRS80	32	22	11.130	2.688	0.906	0.925	0.004	0.772
17	*HmLCS7*	31	7	1.570	0.833	0.290	0.369	0.201	0.165
	HmidPS1.1012	30	12	8.219	2.276	0.867	0.893	0.013	0.948
18A	*HmNS6*	32	8	4.582	1.706	0.656	0.794	0.161	0.092
	HmDL110	30	5	2.651	1.149	0.500	0.633	0.197	0.067

18B	*HmidPS1.890*	32	3	2.705	1.040	0.781	0.640	-0.239	0.302
	HmDL34b	26	12	8.723	2.292	0.769	0.903	0.131	0.084
	HmDL214	32	7	1.715	0.909	0.344	0.424	0.176	0.001
	HmidPS1.193	29	8	4.725	1.721	0.621	0.802	0.213	0.053
18C	*HmidPS1.559*	32	9	1.680	0.998	0.375	0.411	0.074	0.278
	Hmid2044	31	16	8.076	2.376	0.581	0.891	0.337	0.000
18D	*HmG53*	26	26	20.485	3.129	1.000	0.970	-0.051	0.760
	HmidILL2.66010	32	3	1.415	0.540	0.281	0.298	0.042	0.027

Table S2.5: Estimated null allele frequencies, and 0.025- and 0.975 confidence bounds for each marker within each cohort.

				Confidence Intervals	
Linkage Group	**Marker**	**Cohort**	**Null Frequency**	**0.025 bound**	**0.975 bound**
1	*HmidPS1.484*	ASF	0.422	0.263	0.597
		SAL	0.322	0.174	0.500
		WCA	0.167	0.047	0.270
		RP	0.187	0.078	0.293
	HmidPS1.332	ASF	0.039	0.000	0.151
		SAL	0.000	-	-
		WCA	0.110	0.038	0.232
		RP	0.071	0.000	0.176
	HmidPS1.859	ASF	0.000	-	-
		SAL	0.063	0.000	0.206
		WCA	0.141	0.057	0.242
		RP	0.100	0.022	0.219
	HmidPS1.227	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.000	-	-
	HmNS19	ASF	0.104	0.043	0.210
		SAL	0.006	-	-
		WCA	0.145	0.074	0.252
		RP	0.078	0.028	0.171
	HmNS56	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.096	0.019	0.214
		RP	0.000	-	-
	HmD14	ASF	0.000	-	-
		SAL	0.034	-	-
		WCA	0.044	0.000	0.159
		RP	0.000	-	-
	HmNR54	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.018	0.001	0.108
		RP	0.041	0.009	0.131
2	*HmidPS1.138*	ASF	0.237	0.133	0.341
		SAL	0.047	0.000	0.147
		WCA	0.165	0.057	0.290
		RP	0.062	0.009	0.166
	HmidILL1.140027	ASF	0.555	0.400	0.702
		SAL	0.568	0.416	0.711
		WCA	0.365	0.201	0.548
		RP	0.433	0.277	0.597
	HmD61	ASF	0.000	-	-
		SAL	0.079	0.000	0.191
		WCA	0.125	0.051	0.234
		RP	0.163	0.077	0.277

	HmD55	ASF	0.153	0.079	0.253
		SAL	0.109	0.039	0.214
		WCA	0.074	0.008	0.178
		RP	0.078	0.007	0.186
	HmidILL2.76149	ASF	0.000	-	-
		SAL	0.114	0.021	0.222
		WCA	0.122	0.036	0.227
		RP	0.000	-	-
	Hmid2015	ASF	0.028	0.000	0.120
		SAL	0.012	0.000	0.115
		WCA	0.098	0.034	0.195
		RP	0.035	0.000	0.138
	HmidILL2.8738	ASF	0.000	-	-
		SAL	0.119	0.058	0.213
		WCA	0.032	0.000	0.137
		RP	0.072	0.000	0.180
3	*HmidPS1.967*	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.039	-	-
		RP	0.013	0.000	0.120
	Hmid65	ASF	0.060	0.005	0.156
		SAL	0.111	0.050	0.213
		WCA	0.058	0.014	0.163
		RP	0.030	0.000	0.121
	HmNR185	ASF	0.000	-	-
		SAL	0.104	0.037	0.207
		WCA	0.053	0.000	0.149
		RP	0.049	0.000	0.149
	HmidPS1.42	ASF	0.126	0.003	0.244
		SAL	0.000	-	-
		WCA	0.042	-	-
		RP	0.102	-	-
4	*HmidPS1.1058*	ASF	0.000	-	-
		SAL	0.011	-	-
		WCA	0.000	-	-
		RP	0.033	0.000	0.127
	HmRS38	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.093	0.005	0.209
		RP	0.000	-	-
	HmRS27	ASF	0.036	0.008	0.117
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.000	-	-
	HmLCS67	ASF	0.000	-	-
		SAL	0.144	0.027	0.250
		WCA	0.108	0.000	0.231
		RP	0.090	0.000	0.216

5	*HmidPS1.374*	ASF	0.170	0.087	0.269
		SAL	0.149	0.050	0.273
		WCA	0.160	0.068	0.265
		RP	0.171	0.077	0.279
	HmidPS1.124	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.037	-	-
	HmD36	ASF	0.099	0.037	0.198
		SAL	0.099	0.025	0.220
		WCA	0.040	0.000	0.132
		RP	0.071	0.022	0.165
	HmLCS147	ASF	0.310	0.219	0.392
		SAL	0.240	0.135	0.353
		WCA	0.380	0.299	0.444
		RP	0.327	0.242	0.406
	HmidILL1.2192	ASF	0.866	0.758	0.938
		SAL	0.766	0.640	0.865
		WCA	0.452	0.290	0.617
		RP	0.880	0.773	0.948
	HmidPS1.228	ASF	0.000	-	-
		SAL	0.080	0.000	0.200
		WCA	0.000	-	-
		RP	0.090	0.004	0.211
	HmNR281	ASF	0.245	0.145	0.361
		SAL	0.187	0.111	0.285
		WCA	0.189	0.115	0.287
		RP	0.171	0.099	0.273
	HmidPS1.551	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.049	0.000	0.170
		RP	0.047	-	-
	HmidPS1.188	ASF	0.000	-	-
		SAL	0.084	0.000	0.208
		WCA	0.000	-	-
		RP	0.156	0.048	0.266
	HmidILL1.47613	ASF	0.026	0.000	0.117
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.000	-	-
	Hmid221	ASF	0.083	0.025	0.172
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.000	-	-
	HmidPS1.1009	ASF	0.113	0.035	0.219
		SAL	0.125	0.032	0.230
		WCA	0.180	0.088	0.280
		RP	0.009	-	-

6	*HmidILL2.64121*	ASF	0.001	0.000	0.144
		SAL	0.082	0.000	0.213
		WCA	0.052	0.000	0.187
		RP	0.000	-	-
	HmidPS1.150	ASF	0.071	0.020	0.165
		SAL	0.104	0.039	0.217
		WCA	0.058	0.019	0.145
		RP	0.067	0.018	0.160
	HmLCS9	ASF	0.070	0.000	0.197
		SAL	0.183	0.074	0.294
		WCA	0.120	0.027	0.246
		RP	0.214	0.114	0.313
	Hmid321	ASF	0.117	0.047	0.213
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.029	-	-
	HmidPS1.1066	ASF	0.401	-	-
		SAL	0.431	-	-
		WCA	0.453	-	-
		RP	0.398	0.230	0.608
	Hmid6	ASF	0.308	0.224	0.394
		SAL	0.140	0.070	0.252
		WCA	0.199	0.118	0.305
		RP	0.120	0.059	0.217
	HmRS129	ASF	0.138	0.069	0.238
		SAL	0.197	0.121	0.295
		WCA	0.180	0.100	0.288
		RP	0.174	0.097	0.278
	HmAD102	ASF	0.132	0.068	0.228
		SAL	0.167	0.096	0.268
		WCA	0.167	0.098	0.262
		RP	0.127	0.057	0.261
7	*HmLCS388*	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.091	-	-
		RP	0.000	-	-
	HmidPS1.860	ASF	0.000	-	-
		SAL	0.065	-	-
		WCA	0.000	-	-
		RP	0.000	-	-
	HmNS17b	ASF	0.129	0.053	0.271
		SAL	0.042	0.007	0.127
		WCA	0.180	0.095	0.290
		RP	0.114	0.056	0.209
	HmidPS1.961	ASF	0.849	0.734	0.934
		SAL	0.795	0.653	0.909
		WCA	0.892	0.761	0.968
		RP	0.762	0.634	0.869

	Hmid310	ASF	0.096	0.032	0.202
		SAL	0.071	0.008	0.179
		WCA	0.017	0.001	0.120
		RP	0.091	0.023	0.201
8	*HmNR191*	ASF	0.170	-	-
		SAL	0.137	0.042	0.297
		WCA	0.100	0.027	0.200
		RP	0.198	0.103	0.316
	Hm2H6F	ASF	0.000	-	-
		SAL	0.008	0.000	0.118
		WCA	0.000	-	-
		RP	0.000	-	-
	HmRS62	ASF	0.086	0.028	0.185
		SAL	0.128	0.058	0.226
		WCA	0.090	0.022	0.198
		RP	0.118	0.042	0.232
	HmDL123	ASF	0.225	0.140	0.329
		SAL	0.192	0.102	0.311
		WCA	0.255	0.176	0.344
		RP	0.174	0.102	0.273
	HmD59	ASF	0.001	0.000	0.032
		SAL	0.043	0.000	0.136
		WCA	0.041	0.000	0.135
		RP	0.041	0.004	0.133
	HmidILL2.71359	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.007	-	-
		RP	0.001	0.000	0.100
	HmSSRex489a	ASF	0.020	0.000	0.155
		SAL	0.040	0.000	0.159
		WCA	0.000	-	-
		RP	0.000	-	-
	HmLCS1	ASF	0.024	-	-
		SAL	0.063	0.000	0.190
		WCA	0.009	0.000	0.107
		RP	0.000	-	-
	HmSSRex489b	ASF	0.000	-	-
		SAL	0.040	0.000	0.159
		WCA	0.000	-	-
		RP	0.000	-	-
	Hmid53	ASF	0.000	-	-
		SAL	0.054	0.000	0.191
		WCA	0.036	-	-
		RP	0.005	-	-
	HmidPS1.972	ASF	0.081	0.007	0.198
		SAL	0.015	0.000	0.137
		WCA	0.000	-	-
		RP	0.000	-	-

	HmSSRex446a	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.013	0.000	0.123
	HmLCS37	ASF			
		SAL	0.051	0.009	0.138
		WCA			
		RP	0.045	0.012	0.127
	HmidILL1.72605	ASF			
		SAL	0.000	-	-
		WCA	0.177	-	-
		RP			
	HmNR258	ASF	0.000	-	-
		SAL	0.013	0.000	0.120
		WCA	0.000	-	-
		RP	0.000	-	-
9	*HmLCS48*	ASF	0.000	-	-
		SAL	0.082	0.000	0.195
		WCA	0.041	0.000	0.154
		RP	0.090	0.000	0.183
	HmNR180	ASF	0.239	0.144	0.348
		SAL	0.272	0.177	0.369
		WCA	0.277	0.183	0.376
		RP	0.318	0.230	0.401
	HmidPS1.831	ASF	0.039	0.000	0.151
		SAL	0.220	0.121	0.352
		WCA	0.055	0.000	0.169
		RP	0.033	0.000	0.150
	H.rub15A01	ASF	0.447	0.285	0.617
		SAL	0.457	0.278	0.648
		WCA	0.469	0.315	0.630
		RP	0.564	0.406	0.712
	HmidPS1.638	ASF	0.068	-	-
		SAL	0.000	-	-
		WCA	0.024	0.000	0.150
		RP	0.000	-	-
	HmidPS1.549	ASF	0.009	0.000	0.116
		SAL	0.090	0.017	0.201
		WCA	0.067	0.008	0.173
		RP	0.144	-	-
	HmNS58	ASF	0.138	0.058	0.246
		SAL	0.075	0.018	0.171
		WCA	0.069	0.010	0.173
		RP	0.125	0.053	0.225
10	*HmLCS152*	ASF	0.034	0.001	0.132
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.000	-	-

	HmRS117	ASF	0.028	0.002	0.138
		SAL	0.026	0.002	0.107
		WCA	0.078	0.025	0.180
		RP	0.000	-	-
	HmPS1.792	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.000	-	-
	HmNSP31	ASF	0.199	0.105	0.303
		SAL	0.195	0.090	0.306
		WCA	0.221	0.124	0.316
		RP	0.263	0.170	0.351
	HmidPS1.382	ASF	0.020	0.000	0.124
		SAL	0.120	0.035	0.233
		WCA	0.037	0.000	0.145
		RP	0.102	0.025	0.207
	HmNR120	ASF	0.063	0.010	0.163
		SAL	0.046	0.001	0.147
		WCA	0.156	0.070	0.267
		RP	0.111	0.043	0.214
	HmNS100	ASF	0.195	0.104	0.302
		SAL	0.145	0.068	0.248
		WCA	0.242	0.153	0.342
		RP	0.198	0.112	0.299
12	*HmidPS1.807*	ASF	0.099	0.011	0.212
		SAL	0.000	-	-
		WCA	0.053	0.000	0.156
		RP	0.000	-	-
	HmNR20	ASF	0.110	-	-
		SAL	0.000	-	-
		WCA	0.104	0.036	0.220
		RP	0.052	0.009	0.142
	Hmid553	ASF	0.120	-	-
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.027	0.000	0.161
	Hmid610	ASF	0.149	-	-
		SAL	0.094	0.032	0.205
		WCA	0.085	0.006	0.197
		RP	0.031	0.000	0.143
	HmidPS1.874	ASF	0.047	0.012	0.134
		SAL	0.112	0.043	0.226
		WCA	0.022	0.006	0.090
		RP	0.000	-	-
13	*HmidPS1.469*	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.000	-	-

	Hmid4010	ASF	0.090	0.002	0.200
		SAL	0.075	0.000	0.194
		WCA	0.075	0.013	0.186
		RP	0.022	0.000	0.128
	Hmid563	ASF	0.040	0.004	0.143
		SAL	0.000	-	-
		WCA	0.018	0.000	0.130
		RP	0.006	0.000	0.083
14	*HmLCS383*	ASF	0.000	0.000	0.092
		SAL	0.000	-	-
		WCA	0.136	-	-
		RP	0.025	0.000	0.144
	HmidPS1.1063	ASF	0.150	0.075	0.264
		SAL	0.000	-	-
		WCA	0.092	0.011	0.250
		RP	0.160	0.088	0.262
	HmidPS1.818	ASF	0.011	0.000	0.112
		SAL	0.000	-	-
		WCA	0.122	0.040	0.250
		RP	0.070	0.010	0.172
	HmidPS1.247	ASF	0.024	0.000	0.140
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.049	0.000	0.163
	HmidPS1.370	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.000	-	-
15	*HmidPS1.355*	ASF	0.018	0.000	0.128
		SAL	0.108	0.023	0.231
		WCA	0.088	0.001	0.200
		RP	0.000	-	-
	HmidPS1.305	ASF	0.094	0.023	0.206
		SAL	0.000	-	-
		WCA	0.100	0.021	0.235
		RP	0.070	0.010	0.172
	HmidPS1.982	ASF	0.165	0.062	0.282
		SAL	0.160	0.049	0.280
		WCA	0.107	0.000	0.220
		RP	0.131	0.013	0.252
	HimidILL2.87955	ASF	0.000	-	-
		SAL	0.105	0.033	0.205
		WCA	0.000	-	-
		RP	0.038	0.000	0.148
	HmDL50	ASF	0.061	0.011	0.160
		SAL	0.007	0.000	0.076
		WCA	0.031	0.003	0.124
		RP	0.092	-	-

16	*HmNS21*	ASF	0.000	-	-
		SAL	0.095	0.031	0.199
		WCA	0.060	0.012	0.152
		RP	0.053	0.013	0.142
	HmRS80	ASF	0.000	-	-
		SAL	0.000	-	-
		WCA	0.034	0.000	0.122
		RP	0.000	-	-
17	*HmidPS1.1012*	ASF	0.073	0.007	0.253
		SAL	0.092	0.030	0.194
		WCA	0.051	0.003	0.268
		RP	0.046	0.004	0.146
	HmLCS7	ASF	0.201	0.104	0.308
		SAL	0.107	0.011	0.237
		WCA	0.090	0.008	0.220
		RP	0.103	0.005	0.240
18A	*HmNS6*	ASF	0.030	0.000	0.153
		SAL	0.046	0.000	0.156
		WCA	0.134	-	-
		RP	0.094	0.024	0.200
	HmDL110	ASF	0.000	-	-
		SAL	0.099	0.017	0.216
		WCA	0.363	-	-
		RP	0.098	0.009	0.221
18B	*HmDL214*	ASF	0.039	-	-
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.044	-	-
	HmDL34b	ASF	0.142	0.065	0.245
		SAL	0.176	0.102	0.272
		WCA	0.216	0.133	0.313
		RP	0.053	0.000	0.165
	HmidPS1.193	ASF	0.010	0.000	0.116
		SAL	0.252	0.156	0.351
		WCA	0.146	0.069	0.249
		RP	0.105	0.029	0.218
	HmidPS1.890	ASF	0.598	0.474	0.727
		SAL	0.572	0.452	0.726
		WCA	0.464	0.354	0.594
		RP	0.388	0.260	0.535
18C	*Hmid2044*	ASF	0.272	0.186	0.360
		SAL	0.202	0.114	0.304
		WCA	0.169	0.085	0.274
		RP	0.158	0.083	0.259
	HmPS1.559	ASF	0.111	0.016	0.220
		SAL	0.095	0.007	0.228
		WCA	0.097	0.012	0.207
		RP	0.051	-	-

18D	*HmidILL2.66010*	ASF	0.695	0.550	0.816
		SAL	0.845	0.734	0.924
		WCA	0.707	0.540	0.842
		RP	0.832	0.721	0.913
	HmG53	ASF	0.047	0.012	0.131
		SAL	0.000	-	-
		WCA	0.000	-	-
		RP	0.000	-	-

Table S2.6: Significant *P*-values for candidate loci under selection, as indicated by the Ewens-Watterson and F_{ST}-outlier tests. The Ewens-Watterson test was conducted for each cohort separately, and interpreted as a two-tailed test against the alternative hypotheses of balancing ($P < 0.025$) or directional ($P > 0.975$) selection (significant at the 5% level). For the F_{ST}-outlier test, cohorts were grouped according to region to facilitate comparisons between wild and cultured cohorts within the respective groups (F_{SC}), as well as between all cohorts (F_{ST}). Significant F_{ST}/F_{CT} values ($P < 0.05$) that exceeded the 0.95 percentile expected value indicated directional selection; while values lower than the 0.05 percentile expected value indicated balancing selection. Candidates for balancing selection are underlined, while candidates for directional selection are in bold. Dashes indicate non-significant values.

		EW Test *P*-values				**F_{ST}-outlier *P*-values**	
LG	**Marker**	**ASF**	**SAL**	**WCA**	**RP**	**F_{ST}**	**F_{CT}**
1	*HmNS19*	0.009	-	0.002	-	-	-
	HmNS56	-	**1.000**	**0.977**	-	-	-
	HmD14	**0.994**	**0.998**	**0.987**	**0.992**	-	-
	HmNR54	-	-	-	-	0.013	-
2	*HmidILL1.140027*	-	0.006	-	-	-	-
	HmD61	-	-	-	-	0.038	-
	Hmid2015	-	-	-	-	0.034	-
3	*Hmid65*	-	-	0.005	-	-	-
	HmNR185	-	-	-	-	0.020	-
4	*HmRS38*	-	**0.999**	**0.977**	-	-	-
	HmRS27	-	0.000	-	-	-	-
5	*HmD36*	0.006	-	-	-	-	-
	HmLCS147	-	-	0.001	-	-	-
	HmidILL1.2192	-	-	-	-	**0.000**	0.045
	HmidILL1.47613	-	-	0.016	-	-	-
	HmidPS1.1009	-	-	0.021	-	**0.031**	-
6	*HmidPS1.150*	-	-	-	-	0.022	-
	HmidPS1.1066	-	-	0.016	-	**0.000**	0.026
	Hmid6	0.010	0.018	0.009	-	-	-
	HmAD102	-	-	-	-	0.033	-
7	*HmLCS388*	-	-	-	-	-	0.025
8	*HmNR191*	0.015	-	-	-	-	0.011
	Hm2H6F	-	-	-	-	0.039	-
	HmRS62	-	-	-	-	0.045	-
	HmDL123	0.000	-	-	0.008	-	-
	HmD59	0.010	-	-	-	-	-
	HmSSRex489a	-	-	-	-	0.023	0.050
	HmSSRex446a	-	-	-	-	-	0.001
	HmidILL1.72605	-	**0.988**	-	-	-	-
	HmNR258	0.018	-	-	-	-	0.034
9	*HmLCS48*	-	-	-	-	**0.000**	**0.000**
	HmidPS1.831	-	-	-	-	0.046	-
	H.rub15A01	-	-	-	-	-	0.009
	HmidPS1.549	-	-	-	-	0.039	-
	HmNS58	-	-	0.001	-	-	0.048
10	*HmLCS152*	-	-	-	-	-	0.005
	HmPS1.792	-	**1.000**	**1.000**	**1.000**	-	-
	HmNSP31	-	0.013	-	-	-	-
12	*HmidPS1.807*	-	-	-	-	0.042	-

	Hmid610	-	-	-	-	0.004	-
13	*Hmid4010*	0.006	-	-	-	-	-
	Hmid563	-	-	-	-	0.036	-
14	*HmidPS1.1063*	-	-	-	-	-	0.047
15	*HmidILL2.87955*	-	-	-	-	-	0.040
	HmDL50	-	0.007	-	-	-	-
17	*HmidPS1.1012*	-	0.004	-	-	-	-
	HmLCS7	-	**0.997**	**0.993**	**0.976**	-	0.009
18B	*HmDL214*	**0.999**	-	**0.982**	-	0.047	-
	HmDL34b	-	-	-	0.018	-	-
18C	*HmidPS1.559*	-	**0.988**	-	**0.994**	**0.000**	**0.000**
18D	*HmG53*	0.018	-	-	0.007	0.021	-

Appendix B

Supplementary Information for Chapter 3

Figures S3.1A – C: Scatterplots depicting correlation analysis for shell length *versus* live weight (A), shell width *versus* –length (B), and shell width *versus* live weight (C). Trend line equations, correlation coefficients and corresponding significance values, and R^2-values are also indicated.

Figures S3.2A – C: Summary of genetic diversity estimates across the various loci within the large and small groups of the FBC cohort (A), and Families A (B) and B (C), individually. These include the mean number of alleles (A_n), mean number of alleles with a frequency of above 5%, mean effective number alleles (A_e), Shannon's Information Index mean (I), mean number of private alleles and mean unbiased expected heterozygosities (uH_e). Error bars denote standard error.

Figures S3.3A – D: Allele frequencies for the markers with private alleles between the large and small groups of the FBC cohort. They include: *HmidILL2.87955* (A), *HmLCS48* (B), *HmidILL.146360* (C), and *HmidPS1.559* (D).

Figure S3.4A – C: Allelic distributions for the most significantly associated loci within the FBC cohort, *HmidILL2.87955* (A), *HmNS18* (B), and *HmidILL.146360* (C).

Figure S3.5A – B: Allelic distributions for the most significantly associated loci within Family B, *HmidILL.146360* (A) and *HmidILL.64192* (B).

Figure S3.6A – B: Summary results for the linear regressions for the divergent alleles, 212 (A) and 196 (B), from the *HmidILL2.87955* locus and mean live weight within the FBC cohort. Correlation coefficients, corresponding significance values, and R^2-values are indicated. The graphs, constructed using mean live weight, provide a summary of the regressions. Error bars represent standard deviations. Individuals were defined as either homozygous (A : A) for the allele in question, heterozygous (A : -), or not having the allele at all (- : -).

Figure S3.7A – B: Summary results for the linear regressions for the divergent alleles, 395 (A) and 385 (B), from the *HmidILL.146360* locus and mean live weight within the FBC cohort. Correlation coefficients, corresponding significance values, and R^2-values are indicated. The graphs, constructed using mean live weight, provide a summary of the regressions. Error bars represent standard deviations. Individuals were defined as either homozygous (A : A) for the allele in question, heterozygous (A : -), or not having the allele at all (- : -).

Figure S3.8A – B: Summary results for the linear regressions for pairs of similarly correlating alleles from different loci and live weight within the FBC cohort. Correlation coefficients, corresponding significance values, and R^2-values are indicated. The graphs, constructed using mean live weight, provide a summary of the regressions. Error bars represent standard deviations. Individuals were grouped according to the number of alleles of interest present; i.e. individuals that were homozygous at only one of the loci (A : A / - : -) and those that were heterozygous at both loci (A : - / B : -) were grouped together. Negatively correlating alleles included 212 (*HmidILL2.87955*) and 385 (*HmidILL.146360*) (A), while positively correlating alleles included 196 (*HmidILL2.87955*) and 395 (*HmidILL.146360*) (B).

Figure S3.9A – B: Summary results for the linear regressions for the divergent alleles, 400 (A) and 405 (B), from the *HmidILL.64192* locus and live weight within Family B. Correlation coefficients, corresponding significance values, and R^2-values are indicated. The graphs, constructed using mean live weight, provide a summary of the regressions. Error bars represent standard deviations. Individuals were defined as either homozygous (A : A) for the allele in question, heterozygous (A : -), or not having the allele at all (- : -).

Table S3.1: Multiplex reactions and marker information.

[a] MP = Multiplex

[b] T_a = Annealing temperature (°C)

Table S3.2: Basic diversity statistics for the large and small groups of the FBC cohort, Family A, and Family B, and for each cohort as a whole. These include: sample size (N), number of alleles per marker (A_n), effective number of alleles per marker (A_e), Shannon's Information Index values (I), observed heterozygosity (H_o), unbiased expected heterozygosity (uH_e), F_{IS} values, and Hardy-Weinberg *P*-values.

Table S3.3: Estimated null allele frequencies, and 0.025- and 0.975 confidence bounds for the FBC cohort.

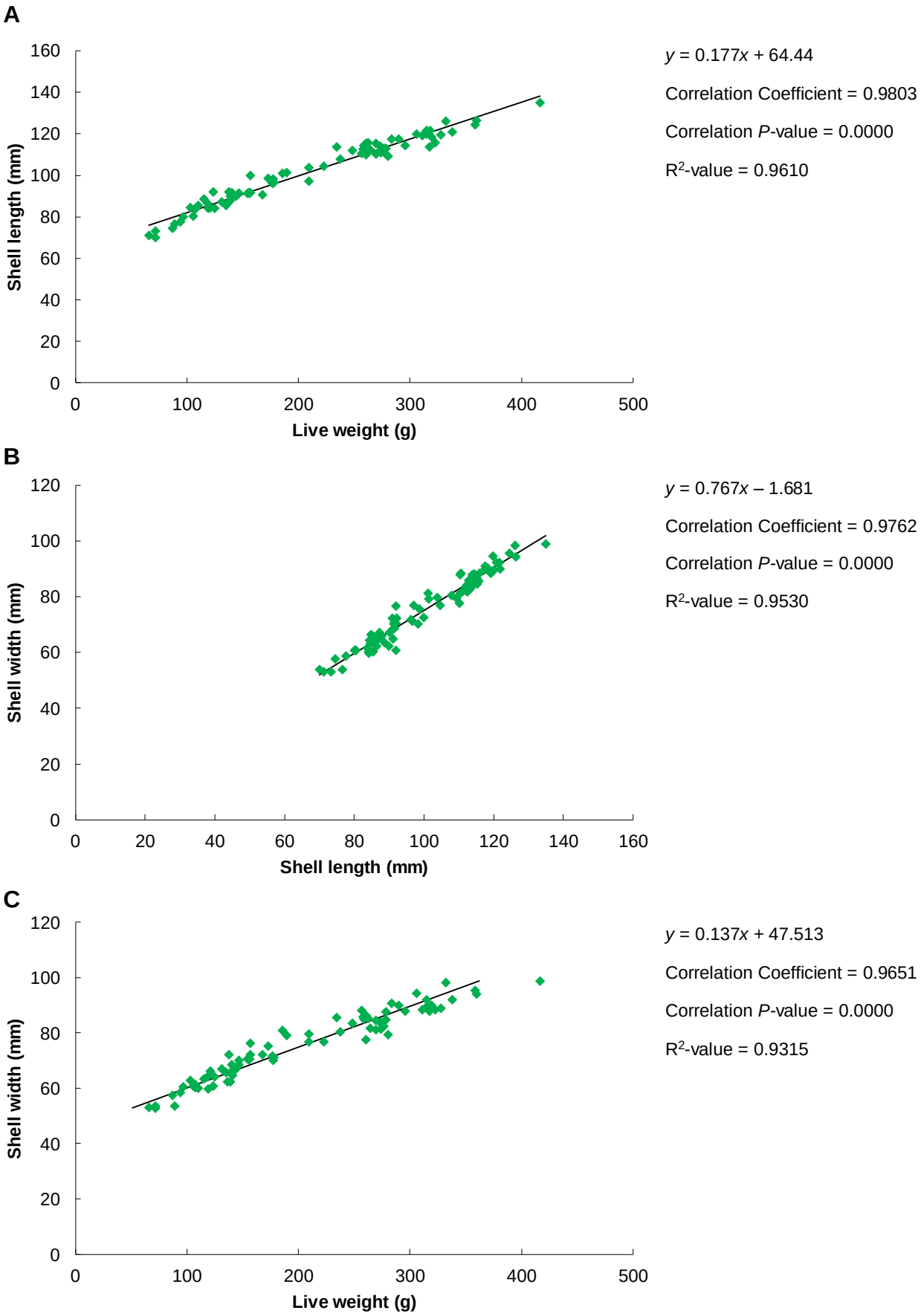
Table S3.4: Marker pairs that showed significant linkage disequilibrium ($P < 0.05$) within the FBC cohort.

Table S3.5: Significance values for the exact G-tests for allelic and genotypic differentiation, and distance values for the allelic and genotypic distance tests, for the large and small groups of the FBC cohort, Family A and Family B, as well as the significance values for the multi-allelic trend test performed for the FBC cohort.

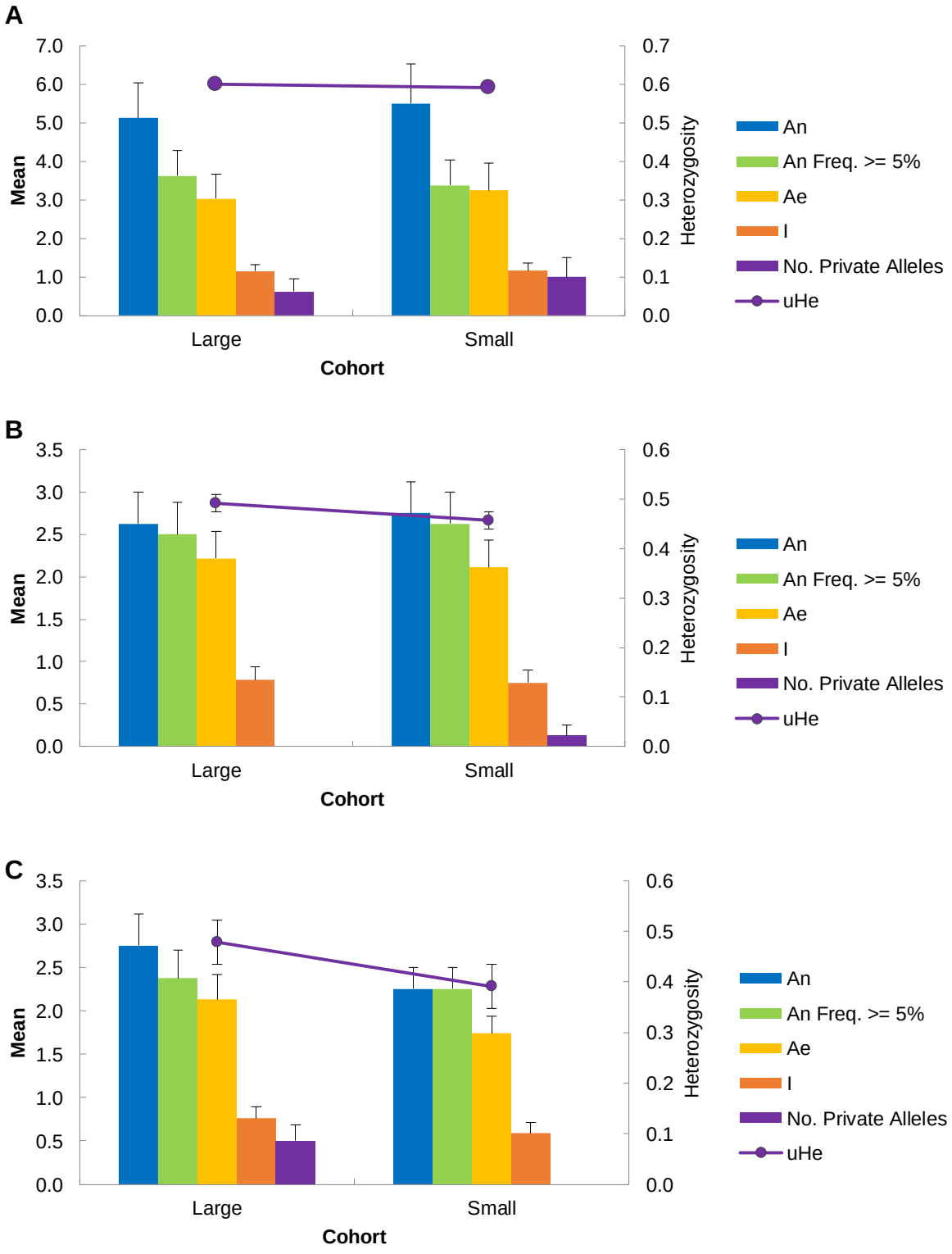
Table S3.6: F-statistics for the single-locus F-tests performed for shell width and –length and live weight in the FBC cohort, Family A and Family B.

Table S3.7: BLAST results for candidate loci under selection with significant similarity to known genes.

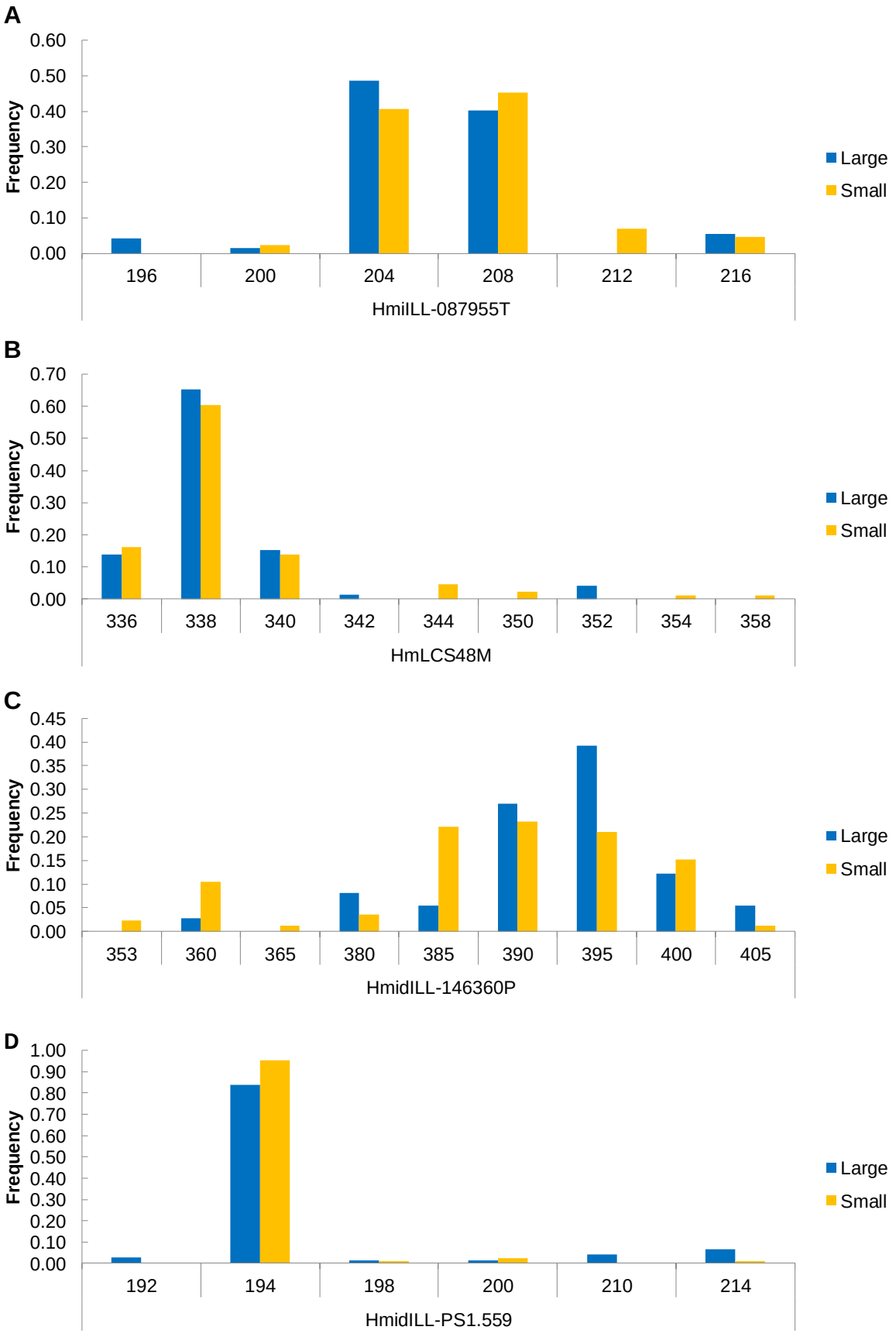
Table S3.8: BLAST results for candidate loci with conserved regions showing significant similarity to multiple known/hypothetical genes.



Figures S3.1A – C: Scatterplots depicting correlation analysis for shell length *versus* live weight (**A**), shell width *versus* –length (**B**), and shell width *versus* live weight (**C**). Trend line equations, correlation coefficients and corresponding significance values, and R^2-values are also indicated.



Figures S3.2A – C: Summary of genetic diversity estimates across the various loci within the large and small groups of the FBC cohort (**A**), and Families A (**B**) and B (**C**), individually. These include the mean number of alleles (A_n), mean number of alleles with a frequency of above 5%, mean effective number alleles (A_e), Shannon's Information Index mean (I), mean number of private alleles and mean unbiased expected heterozygosities (uH_e). Error bars denote standard error.



Figures S3.3A – D: Allele frequencies for the markers with private alleles between the large and small groups of the FBC cohort. They include: HmidILL-085955T (**A**), *HmLCS48* (**B**), *HmidILL.146360* (**C**), and *HmidPS1.559* (**D**).

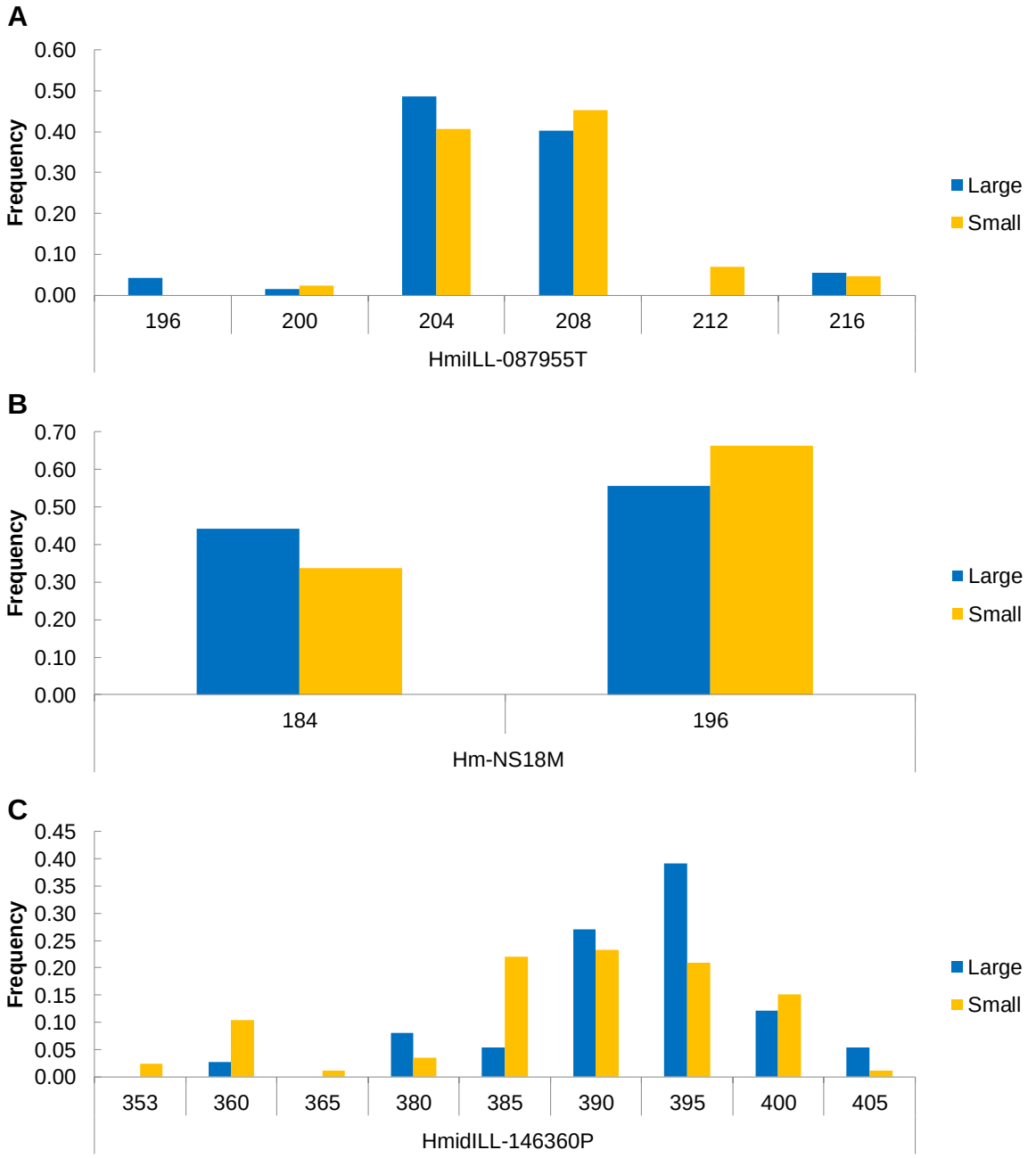


Figure S3.4A – C: Allelic distributions for the most significantly associated loci within the FBC cohort, *HmidILL2.87955* (**A**), *HmNS18* (**B**), and *HmidILL.146360* (**C**).

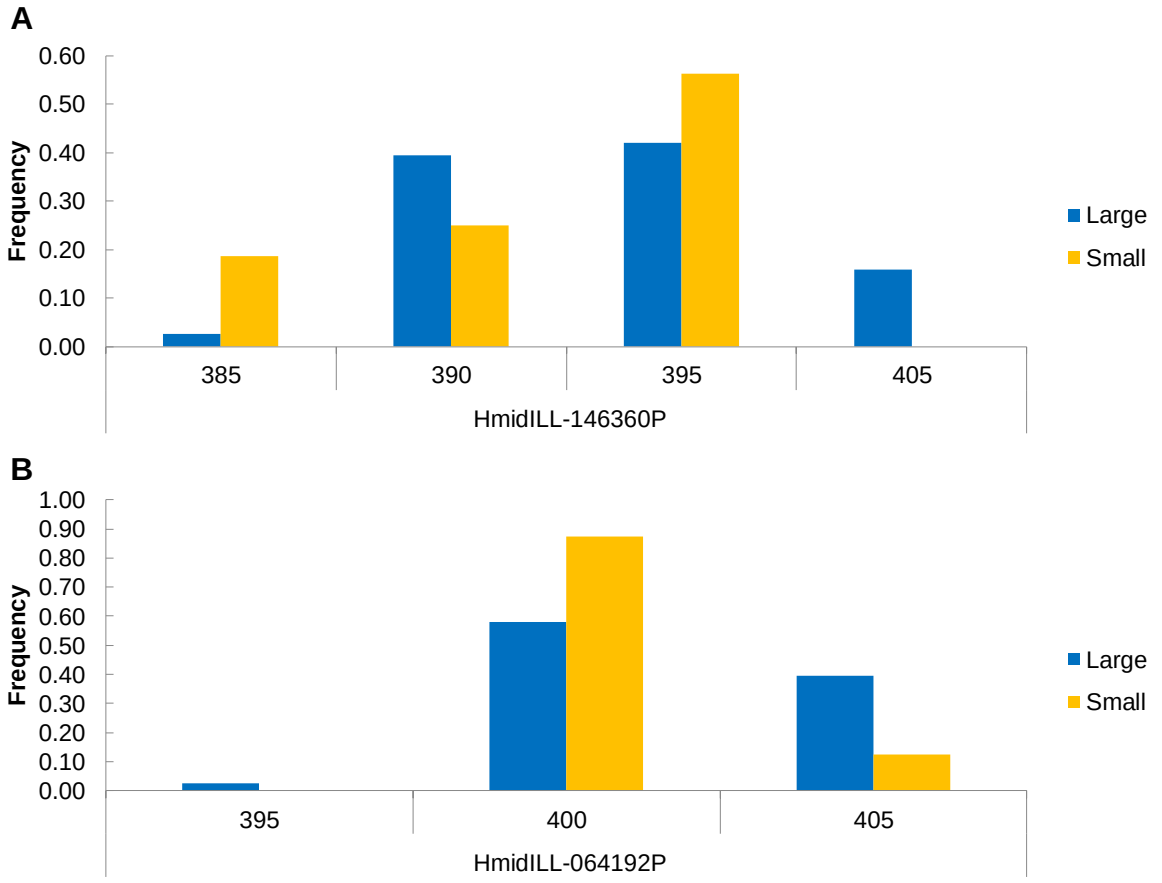


Figure S3.5A – B: Allelic distributions for the most significantly associated loci within Family B, *HmidILL.146360* (**A**) and *HmidILL.64192* (**B**).

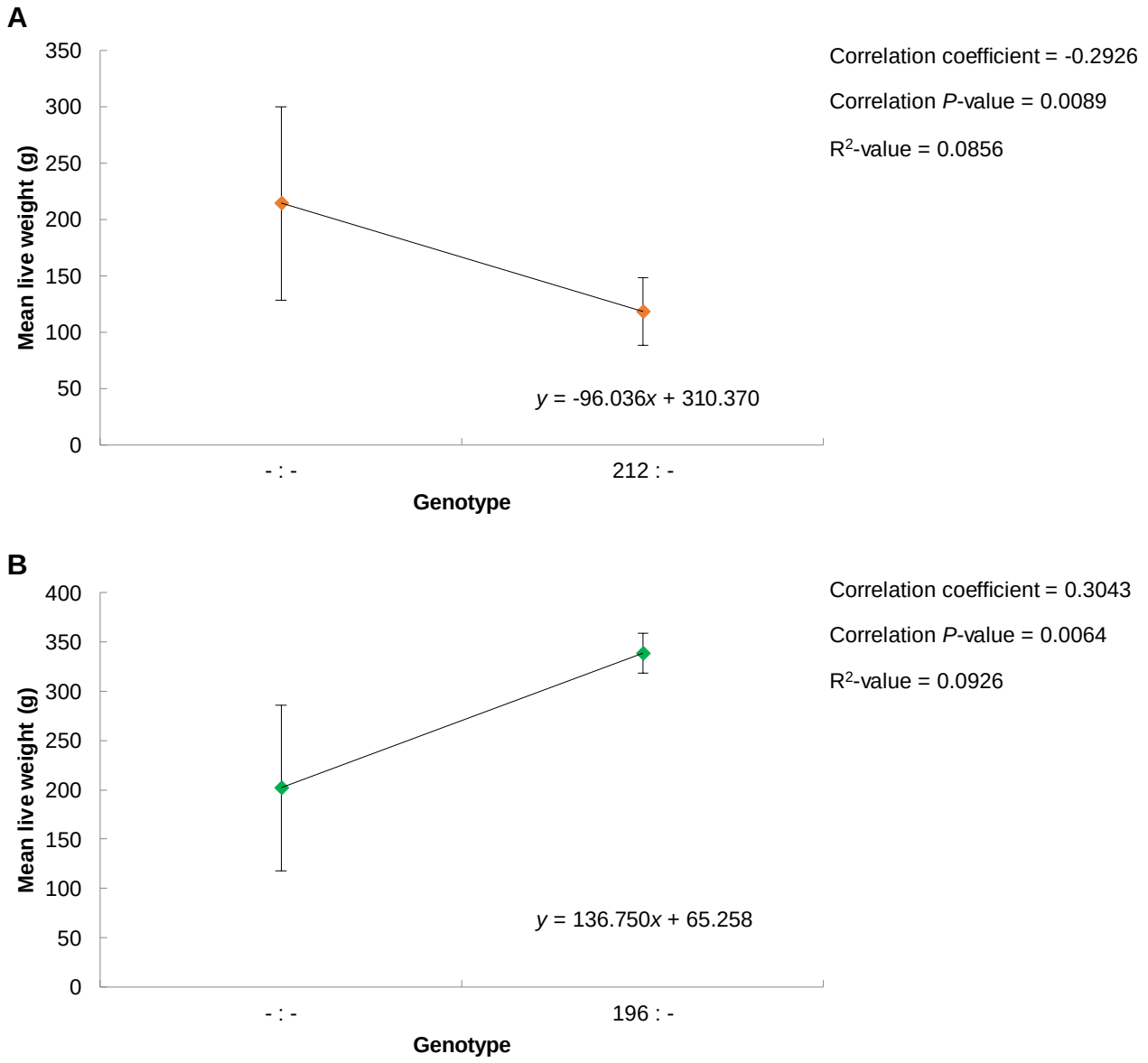


Figure S3.6A – B: Summary results for the linear regressions for the divergent alleles, 212 (**A**) and 196 (**B**), from the *HmidILL2.87955* locus and mean live weight within the FBC cohort. Correlation coefficients, corresponding significance values, and R^2-values are indicated. The graphs, constructed using mean live weight, provide a summary of the regressions. Error bars represent standard deviations. Individuals were defined as either homozygous (A : A) for the allele in question, heterozygous (A : -), or not having the allele at all (- : -).

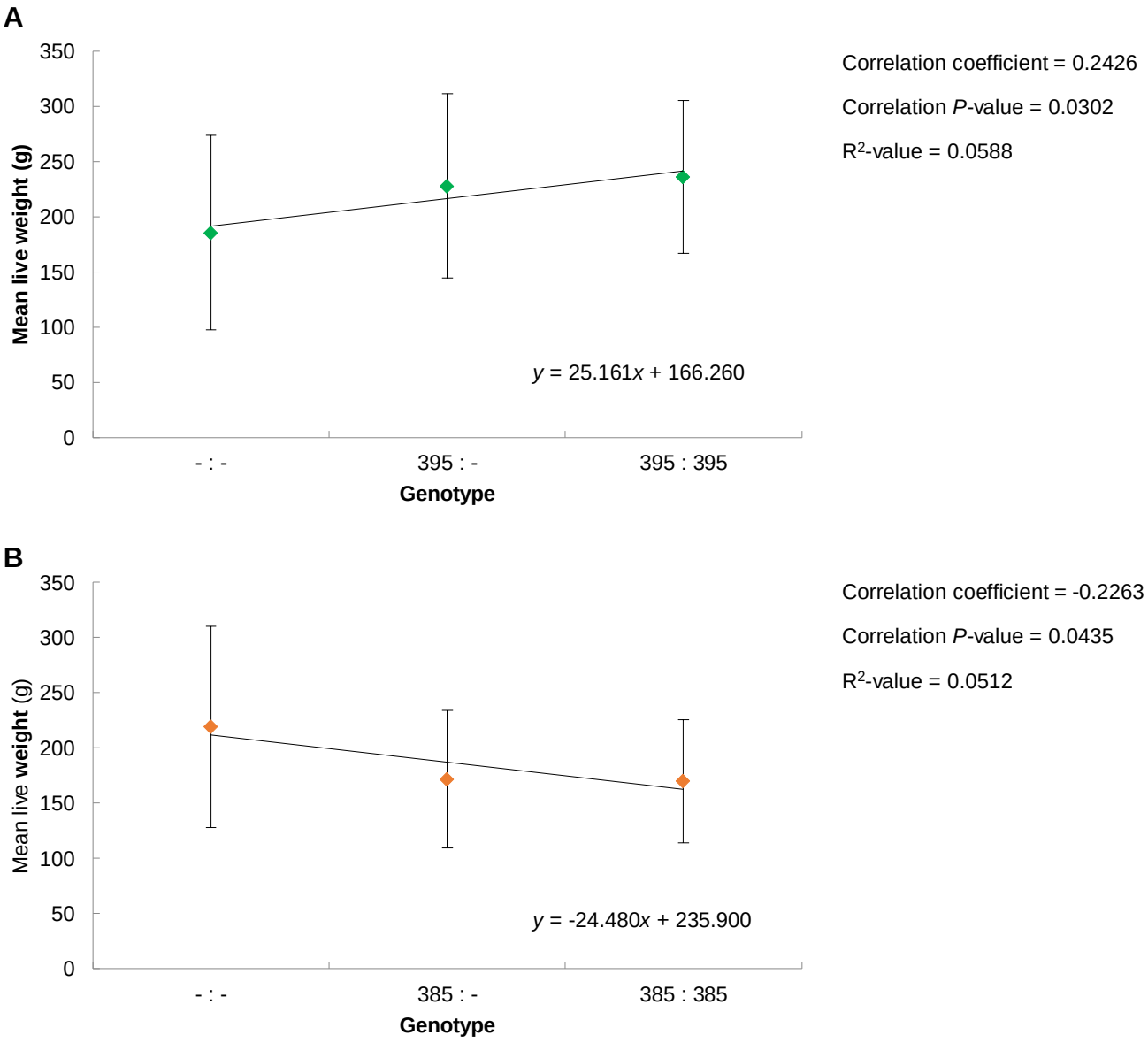


Figure S3.7A – B: Summary results for the linear regressions for the divergent alleles, 395 (**A**) and 385 (**B**), from the *HmidILL.146360* locus and mean live weight within the FBC cohort. Correlation coefficients, corresponding significance values, and R^2-values are indicated. The graphs, constructed using mean live weight, provide a summary of the regressions. Error bars represent standard deviations. Individuals were defined as either homozygous (A : A) for the allele in question, heterozygous (A : -), or not having the allele at all (- : -).

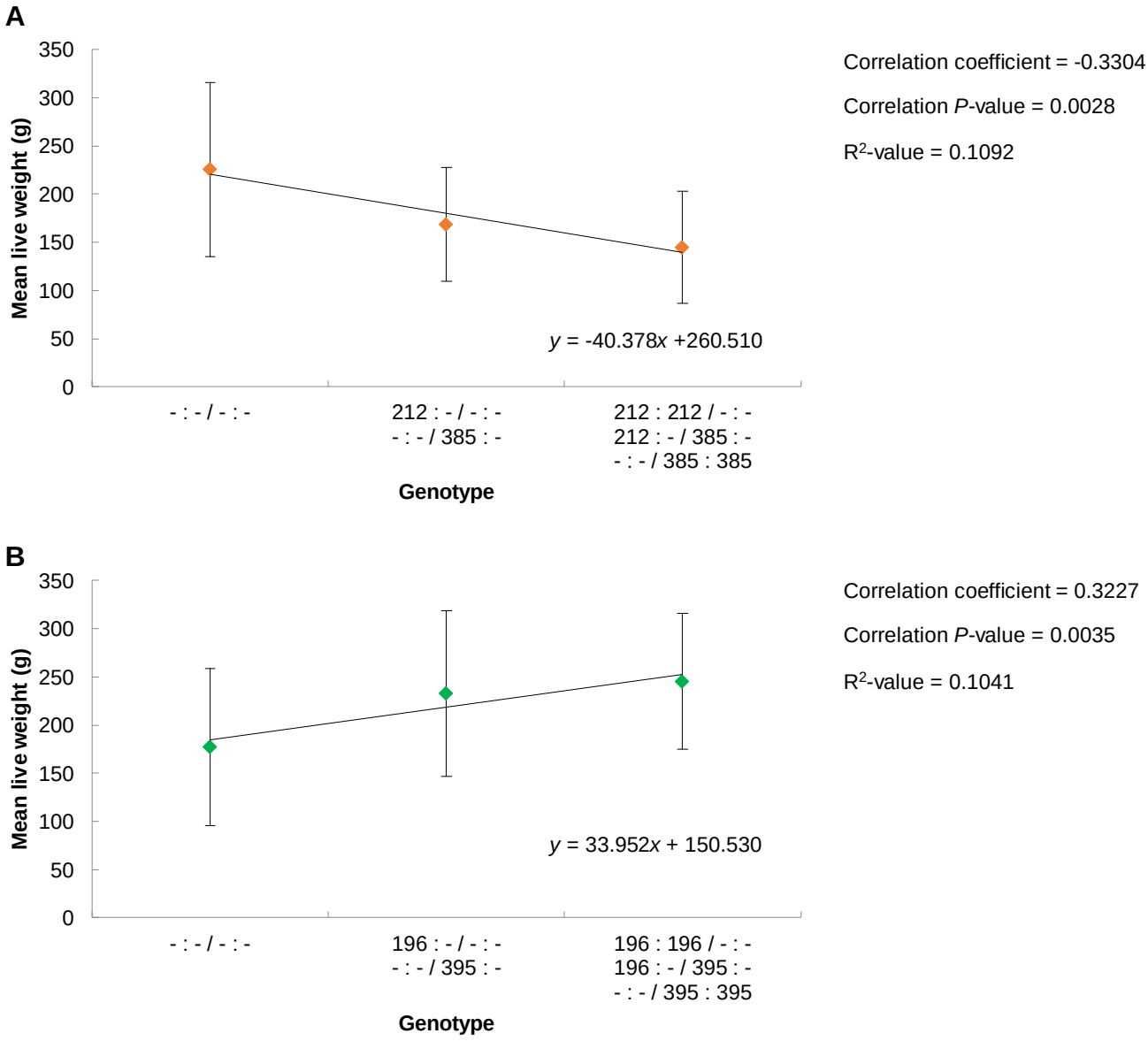


Figure S3.8A – B: Summary results for the linear regressions for pairs of similarly correlating alleles from different loci and live weight within the FBC cohort. Correlation coefficients, corresponding significance values, and R^2-values are indicated. The graphs, constructed using mean live weight, provide a summary of the regressions. Error bars represent standard deviations. Individuals were grouped according to the number of alleles of interest present; i.e. individuals that were homozygous at only one of the loci (A : A / - : -) and those that were heterozygous at both loci (A : - / B : -) were grouped together. Negatively correlating alleles included 212 (*HmidILL2.87955*) and 385 (*HmidILL.146360*) (**A**), while positively correlating alleles included 196 (*HmidILL2.87955*) and 395 (*HmidILL.146360*) (**B**).

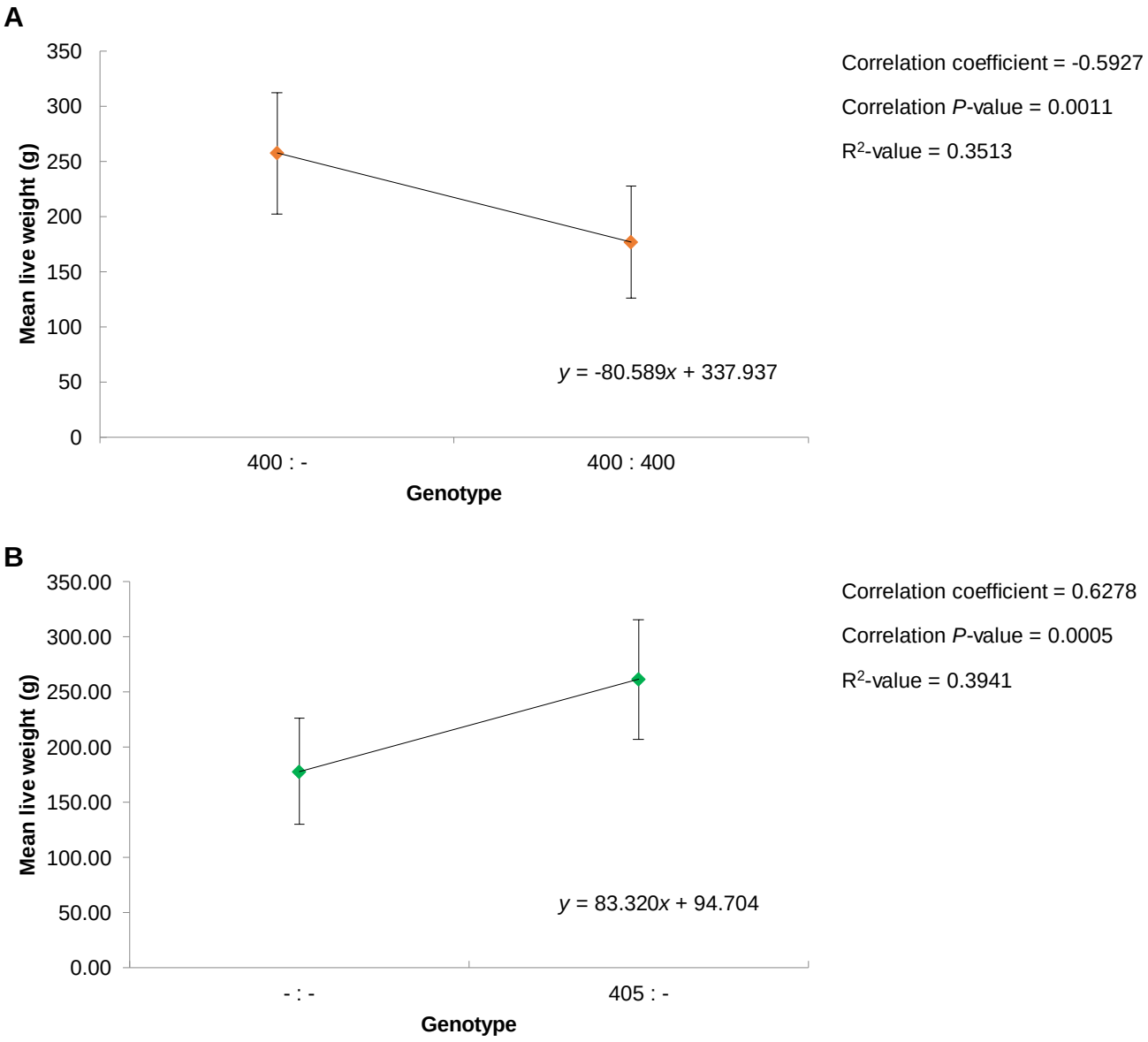


Figure S3.9A – B: Summary results for the linear regressions for the divergent alleles, 400 (**A**) and 405 (**B**), from the *HmidILL.64192* locus and live weight within Family B. Correlation coefficients, corresponding significance values, and R^2-values are indicated. The graphs, constructed using mean live weight, provide a summary of the regressions. Error bars represent standard deviations. Individuals were defined as either homozygous (A : A) for the allele in question, heterozygous (A : -), or not having the allele at all (- : -).

Table S3.1: Multiplex reactions and marker information.

MP[a]	Marker	Motif	Marker T_a[b]	Label	MP T_a	Size range	Source
1	*HmidPS1.559*	$(CA)_n(G)(TCAC)_n$	55	VIC	54	105-140	Slabbert *et al.* 2012
	HmidPS1.561	$(TTGC)_n...(TGTT)_n$	55	FAM		115-150	Slabbert *et al.* 2012
	HmLCS5	$(GCTC)_4(ACTC)_3$	55	FAM		472-482	Slabbert *et al.* 2008
	HmidILL2.87955	GTGA	50	FAM		192-222	Rhode *et al.* 2012
	HmLCS48	$(CT)_{14}(CA)_9$	55	VIC		334-352	Slabbert *et al.* 2008
	HmNS18	$(ACCA)(AGG)_n(ACC)_nAG$ $(ACC)_nAG(AAC)$	60	PET		178-500	Slabbert *et al.* 2010
2	*HmNR106*	$(TG)_{15}$	60	FAM	57	329-389	Slabbert *et al.* 2008
	HmidILL.64192	AAATA	61	PET		375-415	Rhode *et al.* 2012
	HmidILL.146360	TTCTT	60	NED		395-445	Rhode *et al.* 2012
3	*HmNSS1*	$(GGGTTA)_n$	56	PET	55	200-260	Slabbert *et al.* 2010
	HmidILL.88398	TTTTG	58	PET		400-435	Rhode *et al.* 2012
	HmidILL.37506	GTGA	56	VIC		370-405	Rhode *et al.* 2012
	HmidILL1.2192	ATAC	58	VIC		415-450	Rhode *et al.* 2012

[a] MP = Multiplex

[b] T_a = Annealing temperature (°C)

Table S3.2: Basic diversity statistics for the large and small groups of the FBC cohort, Family A, and Family B, and for each cohort as a whole. These include: sample size (N), number of alleles per marker (A_n), effective number of alleles per marker (A_e), Shannon's Information Index values (I), observed heterozygosity (H_o), unbiased expected heterozygosity (uH_e), F_{IS} values, and Hardy-Weinberg *P*-values.

Cohort	Size	Marker	N	A_n	A_e	I	H_o	uH_e	F_{is}	HW *P*-value
FBC	**Large**	***HmidlLL2.87955***	36	5	2.4780	1.0693	0.7778	0.6049	-0.3040	0.0046**
		HmLCS5	33	3	2.6115	1.0288	0.4848	0.6266	0.2143	0.0368*
		HmLCS48	36	5	2.1246	1.0315	0.3889	0.5368	0.2653	0.0000**
		HmNS18	35	2	1.9742	0.6866	0.8857	0.5006	-0.7949	0.0000**
		HmidlLL.146360	37	7	3.9283	1.5937	0.7568	0.7556	-0.0152	0.0062**
		HmidlLL.64192	37	3	2.4914	0.9781	0.5405	0.6068	0.0970	0.3238
		HmNR106	34	10	7.2025	2.1082	0.8824	0.8740	-0.0246	0.0683
		HmidPS1.559	37	6	1.4099	0.6742	0.2162	0.2947	0.2563	0.0700
	Small	***HmidlLL2.87955***	43	5	2.6395	1.1404	0.6512	0.6285	-0.0483	0.6455
		HmLCS5	40	4	2.7188	1.0881	0.5250	0.6402	0.1696	0.1538
		HmLCS48	43	7	2.4123	1.2083	0.4884	0.5923	0.1658	0.0432*
		HmNS18	43	2	1.8083	0.6392	0.6744	0.4523	-0.5088	0.0012**
		HmidlLL.146360	43	9	5.4785	1.8301	0.7674	0.8271	0.0612	0.0020**
		HmidlLL.64192	43	3	2.6061	1.0124	0.6512	0.6235	-0.0566	0.9165
		HmNR106	43	10	7.1806	2.0976	0.9535	0.8709	-0.1078	0.0034**
		HmidPS1.559	42	4	1.1015	0.2410	0.0952	0.0932	-0.0338	1.0000
	All	***HmidlLL2.87955***	79	6	2.5891	1.1493	0.7089	0.6177	-0.1736	0.0570*
		HmLCS5	73	4	2.6799	1.0681	0.5068	0.6312	0.1916	0.0060**
		HmLCS48	79	9	2.2823	1.1795	0.443	0.5654	0.2131	0.0000**
		HmNS18	78	2	1.8989	0.6663	0.7692	0.4764	-0.6589	0.0000**
		HmidlLL.146360	80	9	5.0653	1.7988	0.7625	0.8076	0.0248	0.0000**
		HmidlLL.64192	80	3	2.5539	0.9974	0.6000	0.6123	0.0191	0.6797
		HmNR106	77	10	7.3062	2.1244	0.9221	0.8688	-0.0662	0.0004**
		HmidPS1.559	79	6	1.2343	0.4813	0.1519	0.1910	0.1865	0.0364*
Family A	**Large**	***HmidlLL2.87955***	16	2	1.8824	0.6616	0.7500	0.4839	-0.6000	0.0373*
		HmLCS5	16	3	2.7380	1.0458	0.7500	0.6552	-0.1815	0.2298
		HmLCS48	16	3	1.7239	0.6922	0.5000	0.4335	-0.1907	0.1516
		HmNS18	16	2	1.8221	0.6435	0.6875	0.4657	-0.5238	0.0949
		HmidlLL.146360	16	4	2.9767	1.2342	0.6875	0.6855	-0.0353	0.2414
		HmidlLL.64192	16	2	1.7534	0.6211	0.6250	0.4435	-0.4545	0.2339
		HmNR106	16	4	3.8496	1.3666	0.9375	0.7641	-0.2665	0.0256*
		HmidPS1.559	16	1	1.0000	0.0000	0.0000	0.0000	-	-
	Small	***HmidlLL2.87955***	16	2	1.8824	0.6616	0.6250	0.4839	-0.3333	0.3191
		HmLCS5	15	3	2.3810	0.9433	0.6667	0.6000	-0.1494	1.0000
		HmLCS48	16	3	1.3801	0.5386	0.2500	0.2843	0.0922	0.3059
		HmNS18	16	2	1.4382	0.4826	0.3750	0.3145	-0.2308	1.0000
		HmidlLL.146360	16	4	3.0843	1.2420	0.7500	0.6976	-0.1098	0.1916
		HmidlLL.64192	16	3	2.0237	0.7856	0.8125	0.5222	-0.6062	0.0303*
		HmNR106	16	4	3.7101	1.3463	1.0000	0.7540	-0.3690	0.1426
		HmidPS1.559	16	1	1.0000	0.0000	0.0000	0.0000	-	-
	All	***HmidlLL2.87955***	32	2	1.8824	0.6616	0.6875	0.4762	-0.4667	0.0202*
		HmLCS5	31	3	2.5799	1.0054	0.7097	0.6224	-0.1662	0.4267
		HmLCS48	32	3	1.5574	0.6390	0.3750	0.3636	-0.0787	0.1495

		HmNS18	32	2	1.6397	0.5789	0.5313	0.3963	-0.4057	0.0718
		HmidILL.146360	32	4	3.0386	1.2407	0.7188	0.6815	-0.0729	0.4837
		HmidILL.64192	32	3	1.8910	0.7173	0.7188	0.4787	-0.5365	0.0028**
		HmNR106	32	4	3.7996	1.3601	0.9688	0.7485	-0.3174	0.0005**
		HmidPS1.559	32	1	1.0000	0.0000	0.0000	0.0000	-	-
Family B	**Large**	***HmidILL2.87955***	19	2	1.9945	0.6918	0.9474	0.5121	-0.9000	0.0002**
		HmLCS5	19	3	1.8095	0.7142	0.2105	0.4595	0.5294	0.0120*
		HmLCS48	19	3	1.7150	0.7371	0.3684	0.4282	0.1163	0.0094**
		HmNS18	19	2	1.9151	0.6708	0.7895	0.4908	-0.6522	0.0123*
		HmidILL.146360	19	4	2.7876	1.1183	0.6842	0.6586	-0.0670	0.0208*
		HmidILL.64192	19	3	2.0338	0.7791	0.8421	0.5220	-0.6567	0.0035**
		HmNR106	19	4	3.7801	1.3554	0.9474	0.7553	-0.2881	0.0007**
		HmidPS1.559	18	1	1.0000	0.0000	0.0000	0.0000	-	-
	Small	***HmidILL2.87955***	8	2	1.7534	0.6211	0.6250	0.4583	-0.4545	0.4872
		HmLCS5	8	2	1.1327	0.2338	0.1250	0.1250	-0.0667	-
		HmLCS48	8	3	2.1333	0.9003	0.5000	0.5667	0.0588	0.1422
		HmNS18	8	2	1.9692	0.6853	0.8750	0.5250	-0.7778	0.1385
		HmidILL.146360	8	3	2.4151	0.9841	0.8750	0.6250	-0.4933	0.3287
		HmidILL.64192	8	2	1.2800	0.3768	0.2500	0.2333	-0.1429	1.0000
		HmNR106	8	3	2.2456	0.8815	1.0000	0.5917	-0.8028	0.0258*
		HmidPS1.559	8	1	1.0000	0.0000	0.0000	0.0000	-	-
	All	***HmidILL2.87955***	27	2	1.9570	0.6821	0.8519	0.4983	-0.6938	0.0002**
		HmLCS5	27	3	1.5969	0.6172	0.1852	0.3809	0.4057	0.0075**
		HmLCS48	27	3	1.8340	0.7917	0.4074	0.4633	0.0841	0.0005**
		HmNS18	27	2	1.9337	0.6759	0.8148	0.4920	-0.7159	0.0007**
		HmidILL.146360	27	4	2.8092	1.1610	0.7407	0.6562	-0.2705	0.0344*
		HmidILL.64192	27	3	1.8386	0.7080	0.6667	0.4647	-0.5021	0.0348*
		HmNR106	27	4	3.3907	1.2905	0.9630	0.7184	-0.5094	0.0000**

* $P < 0.05$; ** $P < 0.01$

Table S3.3: Estimated null allele frequencies, and 0.025- and 0.975 confidence bounds for the FBC cohort.

Marker	Frequency Estimate	0.025 Bound	0.975 Bound
HmidILL2.87955	0.0345	0.0000	0.1045
HmLCS5*	0.5489	0.4621	0.6354
HmLCS48	0.0894	0.0328	0.1583
HmNS18*	0.4804	0.3827	0.5789
HmidILL.146360	0.0605	0.0176	0.1228
HmidILL.64192*	0.4495	0.3556	0.5453
HmNR106	0.0302	0.0028	0.0801
HmidPS1.559	0.1102	-	-

* Marker with significant null allele frequencies.

Table S3.4: Marker pairs that showed significant linkage disequilibrium ($P < 0.05$) within the FBC cohort.

Marker 1	Marker 2	*P*-value
HmidILL2.87955	*HmLCS5*	0.0092
	HmLCS48	0.0083
	HmidILL.146360	0.0070
	HmNR106	0.0016
	HmidPS1.559	0.0143
HmNR106	*HmLCS5*	0.0018
	HmidILL.146360	0.0009
	HmidILL.64192	0.0236
	HmidPS1.559	0.0382
HmidILL.64192	*HmNS18*	0.0331
	HmidILL.146360	0.0115

Table S3.5: Significance values for the exact G-tests for allelic and genotypic differentiation, and distance values for the allelic and genotypic distance tests, for the large and small groups of the FBC cohort, Family A and Family B, as well as the significance values for the multi-allelic trend test performed for the FBC cohort.

		Exact G-tests		Distance Tests		
Cohort	**Marker**	**Allelic**	**Genotypic**	**Allelic**	**Genotypic**	**Multi-allelic Trend Test**
FBC	***HmidlLL2.87955***	**0.0421***	**0.0278***	0.1298	0.3443	0.0694
	HmLCS5	0.7463	0.7840	0.0701	0.1038	0.6602
	HmLCS48	0.0506	0.1058	0.1169	0.2700	0.2325
	Hm-NS18M	0.1989	**0.0325***	**0.1056***	**0.2113***	**0.0276***
	HmidlLL.146360	**0.0017***	**0.0053***	**0.3089***	**0.6128***	**0.0150***
	HmidlLL.64192	0.8688	0.8613	0.0295	0.1565	0.8564
	HmNR106	0.6920	0.7170	0.1118	0.5212	0.6693
	HmidPS1.559	0.0729	0.1425	**0.1248***	0.1956	0.1621
Family A	***HmidlLL2.87955***	1.0000	1.0000	0.0000	0.1250	
	HmLCS5	0.6156	0.5628	0.0938	0.2458	
	HmLCS48	0.3782	0.3280	0.1563	0.3125	
	Hm-NS18M	0.2587	0.1542	0.1563	0.3125	-
	HmidlLL.146360	1.0000	1.0000	0.0625	0.4375	
	HmidlLL.64192	0.6023	0.4330	0.0938	0.1875	
	HmNR106	0.9437	0.8847	0.0625	0.2500	
	HmidPS1.559	-	-	0.0625	0.0625	
Family B	***HmidlLL2.87955***	0.3701	0.0648	0.1612	0.3224	
	HmLCS5	0.0963	0.1987	0.2533	0.2961	
	HmLCS48	0.7062	0.6510	0.1118	0.1316	
	Hm-NS18M	1.0000	1.0000	0.0428	0.0855	-
	HmidlLL.146360	**0.0417***	**0.0392***	0.3026	0.5592	
	HmidlLL.64192	0.0723	**0.0129***	**0.2961***	**0.5921***	
	HmNR106	0.0688	0.0549	0.3322	0.4013	
	HmidPS1.559	-	-	0.0556	0.0556	

* Significant *P*-value ($P < 0.05$)

Table S3.6: F-statistics for the single-locus F-tests performed for shell width and –length and live weight in the FBC cohort, Family A and Family B.

Cohort	Marker	Shell Width (mm)	Shell Length (mm)	Live weight (g)
FBC	***HmidlLL2.87955***	**2.2324***	**2.0691***	**2.3847***
	HmLCS5	0.1685	0.1785	0.1311
	HmLCS48	1.4385	1.4578	1.761
	Hm-NS18M	2.8438	2.7304	2.5545
	HmidlLL.146360	1.6724	1.6012	1.717
	HmidlLL.64192	0.7992	0.6961	0.5119
	HmNR106	1.4454	1.3191	1.4833
	HmidPS1.559	1.3105	1.1325	1.5403
Family A	***HmidlLL2.87955***	0.3047	0.2203	0.5254
	HmLCS5	**3.7064***	2.2896	**2.9472***
	HmLCS48	0.1698	0.0336	0.1246
	Hm-NS18M	2.6925	**5.5495***	**6.7178***
	HmidlLL.146360	0.6109	0.6032	0.8576
	HmidlLL.64192	0.2634	0.6309	0.6502
	HmNR106	1.0001	1.0646	0.8043
	HmidPS1.559	0.2791	0.0379	0.6072
Family B	***HmidlLL2.87955***	3.3789	3.0749	4.0885
	HmLCS5	1.1467	0.8232	1.3458
	HmLCS48	0.5029	0.3240	0.8222
	Hm-NS18M	0.0015	0.0767	0.2604
	HmidlLL.146360	**2.7968***	2.2063	2.0316
	HmidlLL.64192	**7.8068***	**6.6209***	**7.8472***
	HmNR106	1.5125	1.2347	0.9949
	HmidPS1.559	0.2151	0.5184	0.4987

* Significant *P*-value ($P < 0.05$)

Table S3.7: BLAST results for candidate loci under selection with significant similarity to known genes.

Marker	BLAST hit	Organism	Accession No.	*E*-value	% Coverage	% Identity	Gene function
HmidlLL2.87955	ABC transporter 11, family A	*Dictyostelium discoideum*	AF465313.1	7.0e–22	35	73	Transmembrane lipid transport
HmidlLL.146360	Activin A receptor, type I	*Homo sapiens*	NG_008004.1	1.0e–06	11	83	Signal transduction for growth and differentiation factors
HmidlLL.64192	D-lactate dehydrogenase	*Lottia gigantea*	XM_009062238.1	5.0e–47	18	68	Pyruvate metabolism
HmidlLL.37506	Ras-related protein Rab-1A	*Haliotis discus*	EF103367.1	0.00	83	95	Regulation of vesicular transport
HmidlLL1.2192	14-3-3 protein zeta	*Haliotis diversicolor*	KF881014.1	0.00	84	95	Signal transduction

Table S3.8: BLAST results for candidate loci with conserved regions showing significant similarity to multiple known/hypothetical genes.

Marker	BLAST hits	Organism	Accession No.	*E*-value	% Coverage	% Identity	Gene function/identity
HmNS18	Hypothetical protein	*Branchiostoma floridae*	XM_002593551.1	3.0e–04	16	89	Endomembrane protein
	Dwil\GK19912 mRNA	*Drosophila willstoni*	XM_002064022.1	1.0e–03	26	79	Leucine zipper DNA-binding protein
	Dvir\GJ23227 mRNA	*Drosophila virilis*	XM_002053629.1	1.0e–03	28	79	Signal transduction
	Hypothetical protein	*Dictyostelium purpureum*	XM_003287599.1	4.0e–03	28	79	Intracellular signalling
	Hypothetical protein	*Bipolaris zeicola*	XM_007712198.1	4.0e–03	27	76	DNA-binding protein
HmNR106	Hypothetical protein	*Salpingoeca rosetta*	XM_004995275.1	8.0e–08	14	86	Cytoskeletal organisation
	Dipeptidyl-peptidase 6	*Homo sapiens*	NG_033878.1	3.0e–07	13	87	Serine-type peptidase
	Hypothetical protein	*Helobdella robusta*	XM_009028564.1	3.0e–07	13	85	Cytoskeletal organisation
	Enabled homolog, transcript variant 4	*Mus musculus*	NM_001083121.2	3.0e–07	13	85	Cytoskeletal organisation
	TK/SYK protein kinase	*Salpingoeca rosetta*	XM_004988091.1	3.0e–07	13	85	Intracellular signalling
HmLCS5	Mucin 5AC	*Homo sapiens*	KC800812.1	2.0e–18	67	66	Extracellular matrix structural constituent
	ATPase α-subunit	*Haliotis rubra*	AY043205.1	2.0e–18	12	89	Energy metabolism
	Goose-type lysozyme	*Haliotis discus*	JX912535.1	7.0e–18	12	88	Innate immunity
	Hemocyanin	*Haliotis tuberculata*	AJ252741.1	8.0e–17	13	86	Oxygen transport
	Cellulase	*Haliotis discus*	AB125892.1	3.0e–16	12	86	Carbohydrate metabolism
HmidILL.88398	Solute carrier family 17 (sodium phosphate), member 2	*Bos taurus*	NM_001038118.1	1.0E-31	2	94	Transmembrane sodium ion transport
	Interferon-inducible double stranded RNA dependent activator	*Homo sapiens*	BC009470.1	1.0E-17	2	84	Immune response to viral infection
	Carnitine O-palmitoyltransferase 1	*Homo sapiens*	AJ420748.2	5.0E-16	2	84	Fatty acid beta-oxidation