THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Improving the utility of genetic markers in fish populations.

Gilbert Michael Macbeth

Associate Diploma in Applied Biology
Bachelor of Applied Science
Graduate Diploma in Computer Science

A thesis submitted for the degree of Doctor of Philosophy at
The University of Queensland in 2015
School of Mathematics and Physics

Abstract

A review of the use of genetic markers applied to aquaculture and wild fish populations was conducted. The review detailed current knowledge and provided a foundation on which to explore new opportunities to improve the utility of genetic markers in fish populations. The current use of markers is diverse and includes but is not limited to: improvements to genetic selection of captive populations (marker assisted selection, genome wide selection, walkback selection), estimation of variance components (heritability, genetic correlations), pedigree identification (relatedness, kinship, inbreeding, genetic tagging), population studies (stock differentiation, migration rates, species identification, invasive species distribution, effective population size, phylogenetics, illegal fishing) and survival estimation.

This research was used to build on the utility of existing applications of genetic markers in fish populations. In the first application genetic markers were used to identify sires of progeny in a novel breeding program using *in vitro* fertilisation of eggs. Computer simulation between family selection at the onset of a breeding program was optimised to yield a 40% increase in growth rate for barramundi (*Lates calcarifer)*. Such a large gain is of significant economic importance to the barramundi aquaculture industry. The breeding program was designed to yield additional genetic gains from long term selection by managing inbreeding.

Following on from this related design a new method using a binary threshold model was developed to rapidly assess genotype by environmental interactions (GxE) by modelling and estimating genetic correlations between environments. The motivation for this study was to improve the ability to estimate how much genetic improvement predicted by a selective breeding program will be realised in the commercial environment. As such, rapid estimates of genetic correlations are important during the very early stages of breeding program investment. The design was suitable for rapid assessment of GxE over one generation with a true 0.70 genetic correlation yielding standard errors as low as 0.07.

While the first two applications assumed sufficient genetic markers to identify sires, an accurate new statistical method was developed to identify individuals using genetic markers when Type I and Type II errors occur. The new theory advances the application of likelihood methods which were implemented in a new software tool called SHAZA. The methods are useful in wildlife forensic studies where missing data can often occur as a result of the breakdown of DNA in field conditions.

The theory developed and implemented in SHAZA was applied to a mark-recapture design in a dataset which had many individual genotypes with missing loci. The new theory was able to recover 80% more pairwise comparisons of individual genotypes compared to discarding missing data. This facilitated the estimation of abundance and proportion of Spanish mackerel (*Scomberomorus commerson*) caught during commercial harvest which was determined with finite estimates. The abundance estimate is potentially useful in fisheries management and ecological monitoring.

The utility of genetic markers to assess an idealised estimate of the abundance of breeding adults was also investigated by estimating effective population size (*Ne*). In this study it was discovered that outlier genotypes on non-conspecific species created a large bias in linkage disequilibrium estimation of *Ne*. Correspondence analysis methods were tested using simulation as a means of identifying non-target species. Simulations showed that the identification and removal of these non-target genotypes was successful in improving the accuracy of *Ne* estimation.

A review of the major findings, their implications, caveats and future research ideas are discussed in the final chapter. One future area of research would be to investigate the utility of estimating the percentage of full sibs in populations that have insufficient genotype data collected for individual assignment tests. A new proposed method is based on a curve of the cumulative number of false positives plotted against the log likelihood ratio of pairwise comparisons. The change in the curve was found to be sensitive to small changes in the percentage of full-sibs in a population.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Publications incorporated as chapters in this thesis

Macbeth G.M., Broderick D, Ovenden J.R., Buckworth R.C. (2011) Likelihood-based genetic mark–recapture estimates when genotype samples are incomplete and contain typing errors. Theoretical Population Biology 80:185-196. http://www.ncbi.nlm.nih.gov/pubmed/21763337

Macbeth G.M., Palmer P.J. (2011) A novel breeding program for improved growth in barramundi *Lates calcarifer* (Bloch) using foundation stock from progeny-tested parents. Aquaculture 318: 325-334. http://www.sciencedirect.com/science/article/pii/S004484861100442%

Macbeth G.M., D Broderick, Buckworth R.C, Ovenden J.R. (2012) Linkage disequilibrium estimation of effective population size in Spanish mackerel (*Scomberomorus commerson*) with immigrants from divergent populations. Genes Genomes and Genetics http://ncbi.nlm.nih.gov/pubmed/23550119

Macbeth G.M., and Wang Y-G. (2014) Rapid assessment of genotype by environmental interactions and heritability for growth rate in aquaculture species using *in vitro* fertilisation and DNA tagging. Aquaculture 434: 397-402. http://www.sciencedirect.com/science/article/pii/S0044848614004189

Macbeth G.M., Broderick D., Buckworth R.C., Ovenden J.R., Wang Y-G. (2015) How many fish under the boat? Estimating abundance of narrow-barred Spanish mackerel (*Scomberomorus commerson*) using a genetic mark-recapture approach.

Publications relevant to the thesis but not forming part of it

Ovenden J.R., Macbeth G.M., Pope L., Thuesen P., Street R., Broderick D. (2014) Translocation between freshwater catchments has facilitated the rapid spread of tilapia in eastern Australia. Biological Invasions http://link.springer.com/article/10.1007/s10530-014-0754-6

Do C., Waples R.S., Peel D., Macbeth G.M., Tillett 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, http://onlinelibrary.wiley.com/doi/10.1111/1755-0998.12157/abstract

Morgan J.A.T., Michael Macbeth, Damien Broderick, Paul Whatmore, Raewyn Street, Dave Welch and Jennifer R. Ovenden (2013) Hybridisation, paternal leakage and mitochondrial DNA linearization in three automalous fish (Scombridae) mitochondrion. Mitochondrion Journal, http://www.ncbi.nlm.nih.gov/pubmed/23774068

Peel D., Waples R.S., Macbeth G.M., Do C., Ovenden J.R. (2012) Accounting for missing data in contemporary genetic effective population size (Ne). Molecular Ecology Resources 13:243-253. http://ncbi.nlm.nih.gov/pubmed/23280157

Publications included in this thesis

Macbeth G.M., Palmer P.J. (2011) A novel breeding program for improved growth in barramundi *Lates calcarifer* (Bloch) using foundation stock from progeny-tested parents. Aquaculture 318: 325-334. Incorporated as Chapter 2.

Contributor	Statement of contribution
Author Macbeth (Candidate)	Experimental design (90%) Wrote paper (95%) Edited paper (80%) Statistical analysis (100%) Biological methodology (40%) Software development (100%)
Author Palmer	Experimental design (10%) Wrote paper (5%) Edited paper (20%) Biological methodology (60%)

Macbeth G.M., and Wang Y. (2014) Rapid assessment of genotype by environmental interactions and heritability for growth rate in aquaculture species using *in vitro* fertilisation and DNA tagging. Aquaculture 434: 397-402. Incorporated as Chapter 3.

Contributor	Statement of contribution
Author Macbeth (Candidate)	Experimental design (100%) Wrote paper (100%) Edited paper (95%) Statistical methodology (80%) Statistical analysis (95%)
Author Wang	Edited paper (5%) Statistical methodology (20%) Statistical analysis (5%)

Macbeth G.M., Broderick D, Ovenden J.R., Buckworth R.C. (2011) Likelihood-based genetic mark–recapture estimates when genotype samples are incomplete and contain typing errors. Theoretical Population Biology 80:185-196. Incorporated as Chapter 4.

Contributor	Statement of contribution
Author Macbeth (Candidate)	Experimental design (100%) Wrote paper (95%) Edited paper (80%) Statistical methodology (90%) Statistical analysis (100%) Software development (100%)
Author Broderick	Genotyping (100%) Wrote paper (5%) Edited paper (5%) Statistical methodology (10%)
Author Ovenden	Edited paper (10%)
Author Buckworth	Sample collection (100%) Edited paper (5%)

Macbeth G.M., Broderick D., Buckworth R.C., Ovenden J.R., Wang Y., (2015) How many fish under the boat? Estimating abundance of narrow-barred Spanish mackerel (*Scomberomorus commerson*) using a genetic mark-recapture approach (paper in preparation). Incorporated as Chapter 5.

Contributor	Statement of contribution
Author Macbeth (Candidate)	Wrote paper (95%) Edited paper (80%) Statistical methodology (90%) Statistical analysis (100%)
Author Broderick	Genotyping (100%) Edited paper (5%)
Author Buckworth	Sample collection (100%) Edited paper (5%)
Author Ovenden	Wrote paper (5%) Edited paper (5%)
Author Wang	Statistical methodology (10%) Edited paper (5%)

Macbeth G.M., D Broderick, Buckworth R.C, Ovenden J.R. (2012) Linkage disequilibrium estimation of effective population size in Spanish mackerel (*Scomberomorus commerson*) with immigrants from divergent populations. Genes Genomes and Genetics http://ncbi.nlm.nih.gov/pubmed/23550119 Incorporated as Chapter 6.

Contributor	Statement of contribution
Author Macbeth (Candidate)	Wrote paper (95%) Edited paper (85%) Statistical methodology (85%) Statistical simulations (100%)
Author Broderick	Genotyping (100%) Edited paper (5%) Statistical methodology (15%)
Author Buckworth	Sample collection (100%) Edited paper (5%)
Author Ovenden	Wrote paper (5%) Edited paper (5%)

Contributions by others to the thesis

The progeny test design in Chapter 2 was inspired from discussions with Dr Roger Lewer with additional comments provided by Professor You-Gan Wang and three anonymous reviewers. Three anonymous reviewers contributed to chapter 3. The likelihood approach in chapter 4 was inspired by discussions with Professor Lorenz Hauser with additional comments provided by Dr Warwick Nash, Dr Hock Seng Lee, Dr George Leigh, Dr Tony Swain, Dr Alison Kelly and Dr Jessica Morgan who commented on a draft prior to additional comments being received by three anonymous reviewers. In chapter 5, sample collection was provided by Charles Bryce, Adrian Donati, laboratory assistance was provided by Raewyn Street and Marcus McHale while valuable input at various stages throughout the study were provided by Dr Simon Hoyle, Everson Paige, Ann Preece, Dr George Leigh and members of the Northern Territory Spanish Mackerel Fishermen's Association with additional comments provided by three anonymous reviewers. In chapter 6 laboratory assistance was provided by Raewyn Street and Marcus McHale with helpful comments provided by Professor Robin Walpes, Dr Phillip England, Dr Frisco Palstra and two anonymous reviewers. Proof reading the thesis was provided by David Macbeth and Dr Michael Tierney.

Statement of parts of the thesis submitted to qualify for the award of another degree

None

Acknowledgements

To all those contemplating a PhD, it is not always the smartest, the brightest and the most intelligent who can aspire to complete a PhD. It can be ordinary people who have a passion for learning. To all my teachers thank you for the opportunity that so many people in this world do not have.

After completing an Associate Diploma in Applied Biology in 1979 I was fortunate to start working as a cadet in a research environment at the Animal Research Institute, Yeerongpilly. I was encouraged by working with enthusiastic scientists and by their dedicated work ethic. I extend my thanks to Dr Cam McPhee who was my supervisor for over 20 years. He introduced me to quantitative genetics with many skills in this specialised area learnt from him over the years. Most of my earlier work with Dr McPhee was involved as a researcher for the pig industry. It was here that I became familiar with the concept of genetic correlations, heritability and the value of genetic selection to improve traits of economic interest. I enjoyed the challenge of establishing the National Pig Improvement Program with the support of the Australian pig industry. From the pig industry I moved to aquaculture as a specialist geneticist. It is here that I met Dr Paul Palmer a fisheries biologist with whom we swapped many ideas on different selective breeding designs leading to one of my PhD papers with Paul as a co-author.

After a few years in the aquaculture industry, which I found a dynamic and exciting area, I ended up working for Dr Jenny Ovenden who was looking for someone with genetics and computer skills to join the Molecular Fisheries group. This was such a good group of people to work with and I enjoyed working with them very much. I was intellectually challenged yet the challenges suited my previous experiences and mindset. I don’t know how Jenny did it but all the interesting projects she managed to bring to the group made this one of the most enjoyable phases of my career. I also thank Jenny for giving me the support and encouragement to start a PhD. Jenny introduced me to Professor You-Gan Wang as a potential PhD supervisor who also encouraged me to apply. Not long after commencing my PhD it was unfortunate that the Queensland Government cutbacks hit the group. Working within this group was Damien Broderick who I delightfully remember all those little intellectual conversations we had all too often in the name of science. It was through work and from these conversations that Damien broadened my knowledge of molecular genetics. I would also like to thank Dr Rik Buckworth and the Fisheries

Research Development Corporation for making the Spanish mackerel data available for this study. There are so many other people that I could thank and I have had many fond memories working with so many wonderful people.

Finally I would like to thank my devoted wife Lisa and our two children Emma and Laura for their support. To my mum and dad who valued education and encouraged me to continue with further studies I offer all my love. My father, who has since passed away, was encouraged by his high school teachers to enrol in University but neither he nor his surviving parent could afford for him to attend. My mother was also not privileged with high school education. I therefore dedicate this PhD to my father Gilbert Alan Macbeth, to my mother Rita Mary Macbeth and to all those people that have encouraged me along the way.

Keywords

DNA tagging, likelihood approach, mark and recapture, microsatellite, aquaculture, selective breeding, linkage disequilibrium, genotype by environmental interactions, progeny test, barramundi *Lates calcarifer*, Spanish mackerel *Scomberomorus commerson*.

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 010202 Biological Mathematics 40%
ANZSRC code: 010402 Biostatistics 40%
ANZSRC code: 010404 Probability theory 20%

Fields of Research (FoR) Classification

FoR code: 0104, Statistics, 50%
FoR code: 0102, Applied Mathematics, 50%

Table of contents

Chapter 5

Chapter 7

List of Figures

List of Tables

Standard terms and expressions

These terms are generally accepted in other scientific literature

AFLPs	Amplified fragment length polymorphisms
c^2	Common environmental effect
CA	Correspondence analysis
CV	Coefficient of variation
DNA	Deoxyribonucleic acid
EBV	Estimated breeding value
FAO	Food and Agriculture Organisation
GWS	Genome wide selection
GxE	Genotype by environmental interaction
H_0	Null hypothesis
H_1	Alternative hypothesis
HWE	Hardy weinberg equilibrium
h^2	Heritability
LD	Linkage disequilibrium
LLR	Log likelihood ratio
MAS	Marker assisted selection
MCMC	Marcov chain monte carlo
mtDNA	Mitochondrial DNA
Ne	Effective population size
PCR	Polymerase chain reaction
PID	Probability of identity
QTL	Qualtitative trait loci
RAPDs	Random amplified polymorphic DNA
REML	Restricted maximum likelihood
r_g	Genetic correlation
s.d.	Standard deviation
s.e.	Stndare error
SNPs	Single nucleotide polymorphisms
SSCP	Single-strand conformation polymorphism

Note: Non-standard terms and expressions are described in each chapter.

Chapter 1

Introduction to genetic markers applied to fish populations

Macbeth GM

The following introduction gives a brief overview on the use of genetic markers applied to aquaculture and wild fisheries populations. This overview, together with an understanding of the broad scope of genetic marker applications, provided the necessary background in which to explore new ideas for further investigation.

1.1 Fish production and exploitation

Management of fish populations started in the form of farming with the introduction of aquaculture in China dating back to 2000-1000 B.C. (Rabanal, 1988). The first domesticated fish being common carp *Cyprinus carpio* (Balon 2004). Today only a few fish species can be considered truly domesticated with many aquaculture populations relying on wild resources (Teletchea et al., 2014). Together fisheries and aquaculture products are now the most traded food commodities worldwide making significant contributions to the world's well-being and prosperity (FAO, 2015) with aquaculture now providing a staggering 41% of total world fisheries output from a base of only 4% in 1970. Today most wild fish populations are fully exploited or have already been overfished (FAO, 2015). This seems to be supported by more recent evidence showing that wild fisheries capture has remained static since about 1985 (FAO, 2015). The importance of sustainable management practices (Fréon et al., 2005), including rebuilding fish populations (Hilborn et al., 2014), will be essential to maximize sustainable yields from wild fish populations. It is here that genetic markers can play an important role in monitoring fish populations with the uptake expected to grow through new partnerships between fisheries managers and geneticists (Ovenden et al., 2015).

1.2 Genetic markers used in fish populations

Molecular markers in aquaculture and fisheries have been used for over 50 years (Ryman and Utter, 1987; Liu and Cordes, 2004) and their use has steadily increased over the last two decades (Park and Moran, 1994; Chauhan and Rajiv, 2010; Dichmont et al., 2012; Abdul-Muneer, 2014). There are many types of genetic markers which can broadly be classed as protein markers such as allozymes and DNA markers such as variable number of tandem repeats loci (VNTRs: microsatellites, minisatellites), random amplified polymorphic DNA (RAPDs), amplified fragment length polymorphisms (AFLPs), single nucleotide polymorphisms (SNPs), single-strand conformation polymorphism (SSCP) (Hauser et al., 2011; Dichmont et al., 2012; Abdul-Muneer, 2014) and mitochondrial DNA markers (mtDNA) (Gold et al., 1993).

Of all the different types of genetic markers, microsatellite markers are perhaps the most widely used in conservation genetics, fisheries management and aquaculture as they are highly polymorphic, Mendelian inherited and have co-dominant transmission (Abdul-Muneer, 2014; Duran et al., 2009). These properties make microsatellites an ideal marker and have become a mainstream tool applied to genetic applications in both wild populations and in the management of captive aquaculture populations. SNPs are also increasingly becoming popular (Fernández et al., 2013) with conclusions based on SNP markers analogous to that of microsatellite markers (Coates et al., 2009).

1.2.1 Genetic markers used in closed aquaculture populations

In managing captive fish populations there is a balance between achieving a response to selection and minimizing rates of inbreeding (Macbeth, 2007). Traditionally, this field has been a branch of quantitative genetics with little or no use of genetic markers. More recently, however, the development in the application of genetic markers in captive fish populations have been steadily increasing. These applications generally focus on four areas:

(i) Selection – to improve the rate of genetic gains,
(ii) Genetic parameter estimation – to estimate past and future genetic gains,
(iii) Inbreeding and relatedness – to monitor loss in genetic variance and
(iv) Family survival – to assist mating strategies.

Examples of some marker applications in aquaculture include:

Selection

(i) Quantitative Trait Loci (QTL). The identification of loci associated with a quantitative trait. Putative QTL genes were identified in rainbow trout, *Oncorhynchus mykiss* (Ozaki et al. (2001) with linkage maps constructed in many other aquaculture species (Yue 2012, Zhanjiang 2007). High density QTL mapping will enble genom-wide-association studies with the ability to detect very small marker-trait associations (Yue 2012) making genomic selection possible (Goddard and Hayes, 2007).

(ii) Marker Assisted Selection (MAS). Markers close to quantitative trait loci can be used in selection programs to give higher rates of genetic response in traits with low heritability (e.g. 0.06) compared to phenotypic selection (Sonesson, 2007). Short-term gains of MAS seem to be at the expense of longer term selection responses (Meuwissen and Sonesson, 2004; Gibson, 1994) with no known breeding programs using this technology (FAO 2007). MAS can however be exploited in a number of ways not offered by traditional selection, i.e. selection of fish at fingerling stage for traits such as sexual maturation or meat quality traits, or for selection of mortality traits such as disease resistance, salinity or temperature tolerances (Yue 2012).

(iii) Genome-Wide Selection (GWS). Dense SNP markers have identified a large proportion of additive genetic variance (Yang et al., 2010). Sonesson and Meuwissen (2009) and Nielsen et al. (2011) had shown that GWS is potentially promising in aquaculture breeding programs with a higher accuracy than best linear unbiased prediction and with lower rates of inbreeding per generation.

(iv) Walk-back selection. The maximum selection response at a given rate of inbreeding can be achieved using Walk-back selection (Doyle and Herbinger, 1994). In this example genotyping of the heaviest fish are continued until sufficient family structure is recovered to manage inbreeding. This application encompasses kinship analysis using genotype samples from the heaviest fish to identify pedigree relationships. In fish species walkback selection has been proposed by Sonesson (2005) and Robinson et al. (2010).

Genetic parameter estimation

(i) Genotype by Environmental interaction (GxE). In an application of marker-based pedigree assignment GxE effects from growth have been estimated in European sea bass, *Dicentrarchus labrax* (Dupont-Nivet et al., 2008; Le Boucher et al., 2013) and in rainbow trout, *Oncorhynchus mykiss* (Pierce et al., 2008; Le Boucher et al., 2011). As outlined in Chapter 3 these GxE effects are implicated in a reduced response to selection.

(ii) Estimation of heterosis. Heterosis is a non-additive component of genetic variation. A positive correlation between heterozygosity at microsatellite loci and salinity tolerance was observed in guppy populations (Shikano and Taniguchi 2003).

(iii) Estimation of heritability. Heritability is expressed as the percentage of phenotypic variation that has an underlying genetic cause and is of primary importance in quantitative genetic studies. Heritability estimates for growth in the tropical abalone *Haliotis asinina* were estimated using parentage assignment from microsatellites (Lucas et al., 2006).

Inbreeding and relatedness

(i) Kinship analysis. Estimating contribution of parents during mass spawning has been estimated using genetic markers in barramundi (*Lates calcarifer*) (Frost et al., 2006), greater amberjack (*Seriola dumerili*) (Rodriguez-Barreto et al., 2013) and of white sturgeon (*Acipenser transmontanus*) (Rodzen et al., 2003).

(ii) Relatedness. Estimating relatedness of fish (Blonk et al., 2010, Kozfkay et al., 2008) and relatedness of fish trematode parasites (Ndeda et al., 2013) was determined using genetic markers.

(iii) Inbreeding estimation. Microsatellites have been used as a tool to estimate inbreeding in channel catfish (*Ictalurus punctatus*) (Parra-Bracamonte et al., 2011) and in common sole (*Solea solea*) (Blonk et al., 2009). DNA fingerprinting is also a useful tool in estimating rates of inbreeding in aquaculture species during selection (Macbeth 2005).

Family survival

(i) Survival estimation. Survival estimation in rainbow trout using molecular genetic markers (Herbinger et al., 1995) and family survival in oysters (Lind et al., 2009).

1.2.2 Genetic markers used in wild fisheries populations

The use of genetic markers in wild populations is applied to:

(i) Inbreeding and relatedness,
(ii) Spatial analysis and tagging,
(iii) Evolution and species identification and
(iv) Age determination.

Examples of some marker applications in wild fisheries include:

Inbreeding and relatedness

(i) Pedigree analysis. There are a number of methods of pedigree reconstruction including numerous likelihood methods (Marshall et al., 1998; Smith et al., 2001; Wagner et al., 2006; Kalinowski et al., 2007; Wang, 2004, 2011; Reister et al., 2009) and combinatorial reconstruction methods (Sheikh et al., 2010). Pedigree reconstruction has been applied to wild fish populations (Koch et al., 2008; Herbinger et al., 2006; Ford and Williamson, 2010; Aykanat et al., 2014). Close kin analysis used by Bravington and Grewe (2007) is also an application of pedigree analysis.

(ii) Effective population size. Effective population size is an estimate of the capacity of populations to maintain genetic variation and is related to inbreeding. A number of methods utilizing genetic markers have been developed to estimate effective population size (Wang and Whitlock, 2003; Waples and Do, 2008; Zhdanova and Pudovkin, 2008; Wang, 2009; Waples and Waples, 2011). Hatchery supplementation has been found to reduce effective population size in wild salmon (Christie at el., 2012).

Spatial analysis and tagging

(i) Genetic tagging. Genetic mark-recapture methods used to estimate abundance and distribution have been reviewed by Lukacs and Burnham (2005). Using genetic tags, migratory patterns of Northern Atlantic humpback whales have been investigated (Palsboll et al., 1997). *In situ* genetic tagging has also been applied to Spanish mackerel (Buckworth et al., 2012).

(ii) Detection of escaped aquaculture fish populations. Miggiano et al. (2005) had shown in theory that microsatellites could be used to detect gilt seabream escapees. O'Reilly et al. (2006), Glover et al. (2009) and Glover et al. (2013) had found escaped farmed Atlantic salmon, *Salmo salar* using genetic markers. The long term introgression of farmed salmon in wild populations has been identified as a serious conservation issue and has raised global concerns (Glover et al., 2013).

(iii) Migration rates. The theory to estimate migration rates between populations using genetic markers has been developed by Wang and Whitlock (2003) and Paetkau et al. (2004). The conservation of isolated populations may need to be managed differently.

(iv) Spread of invasive fish. Translocation and spread of tilapia has been inferred using microsatellite and mitochondrial markers (Ovenden et al., 2014).

(v) Law enforcement. Illegal fishing (Nielsen et al., 2012). Fish being caught in illegal fishing zones e.g. detection of female crabs being sold illegally (Queensland Government 2010).

Evolution and species identification

(i) Phylogenetics. The study of evolutionary relatedness among groups has been extensively applied to fish populations (Danzmann and Ihssen, 1995; Yue at al., 2009; Fauvelot and Borsa, 2011; Tillett et al., 2011).

(ii) Species identification. 90% of freshwater species from North America can be identified using mitochondrial DNA (April et al., 2011). The range of black tipped sharks *Carcharhinus tilstoni* was extended using DNA species recognition (Ovenden et al., 2010).

(iii) Hybrid identification. Using genetic markers hybrids have been reported in shark populations (Morgan et al., 2011; Morgan et al., 2013) and cichlid populations (Salzburger et al., 2002).

(iv) Stock differentiation. Stock differentiation has implications in stock management and has been detected from the variation in mitochondrial DNA (Park et al., 1993) and SNPs (Candy et al., 2015).

(v) Pathogen identification. DNA identification of virus and bacterial fish pathogens has been demonstrated by Meyers et al. (1992) and Lievens et al. (2011).

Age determination

(i) Age determination. Determination of age is an important parameter in population management in wild fish populations with estimates being difficult to achieve (Campana, 2001). A novel biomarker for age using telomere length has been identified in crustacean species (Godwin et al., 2011) and the Sydney rock oyster (Godwin et al., 2012).

1.3 Proposed research and development of genetic markers

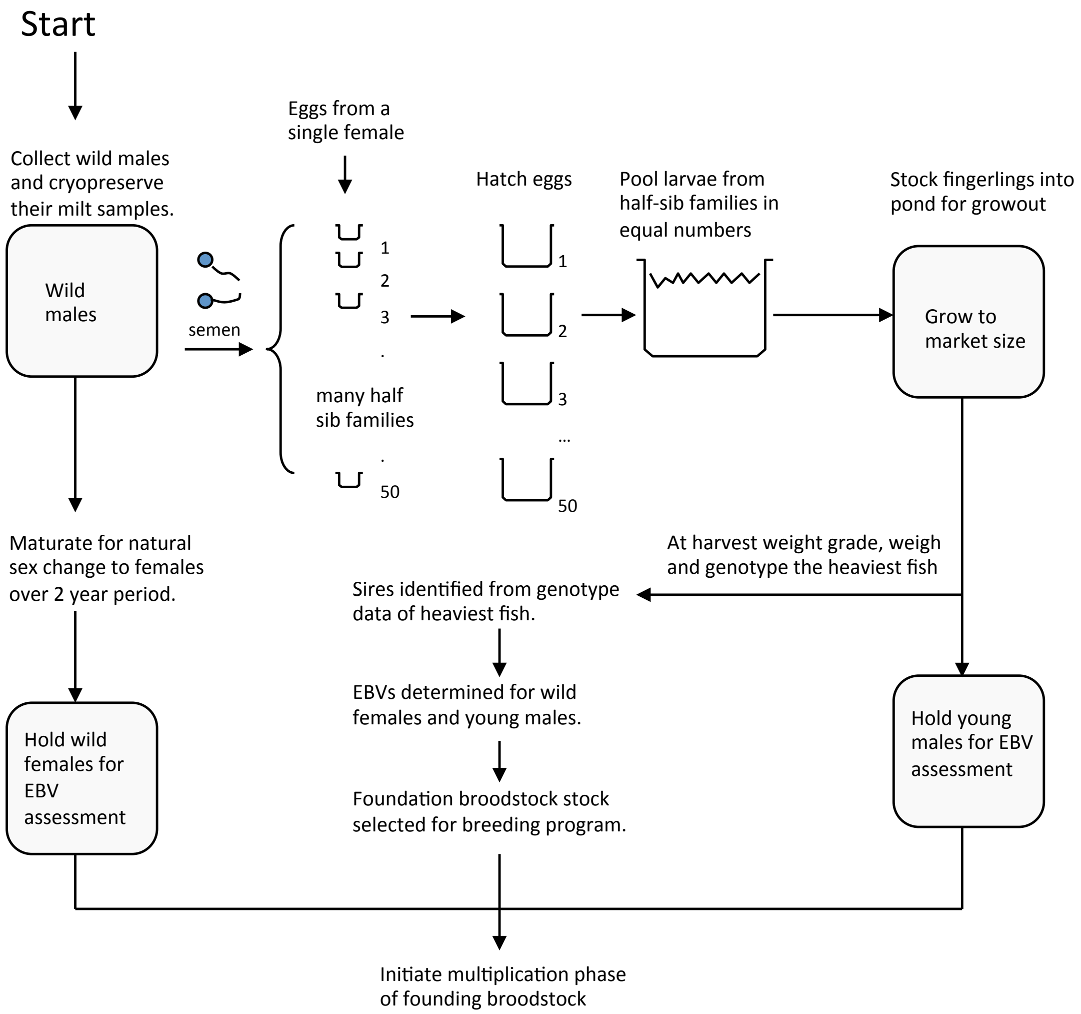
As listed above, the utility of genetic markers applied to aquaculture and fisheries is a diverse area of research and development. Despite the extensive use of molecular genetics in aquaculture and fisheries many techniques are still in the early stages of development. It is this growth and refinement of applications that are a key focus and motivation of this PhD as there are opportunities to expand and improve the existing knowledge base applied genetic markers in fish populations. I have carefully chosen five questions to address based on my review and my expertise in order to demonstrate an improvement in the utility of genetic markers.

1.3.1 How can genetic markers be used to accelerate genetic improvement in fish? (This question is addressed in Chapter 2)

Improvements in growth rate of fish during commercial growout is usually the first trait considered in breeding programs as it is of high economic importance, easily measured and is highly variable and heritable (Macbeth et al., 2002). Selecting the heaviest fish at harvest for use as parents to produce the next generation will yield a cumulative response to selection. A problem with fish is that they may not spawn at the same time with highly variable fertilization rates from parents (Frost et al., 2006) making it difficult to manage inbreeding while applying selection pressure to improve economic traits of interest.

In this study I examine the benefits from using a controlled mating design with artificial mating to improve the rate of genetic gain. One solution to this strategy is to use genetic markers in a novel breeding program that takes advantage of cryopreserved sperm, artificial fertilization and progeny testing to achieve accurate estimated breeding values (EBV) for growth rate. Utilising accurate EBV's, it is thought that large genetic gains can be made by applying a high rate of between family selection at the onset of a breeding program followed by a multiplication phase to increase family numbers for long-term management of inbreeding.

Figure. 1.1. Proposed breeding program designed to achieve large genetic gains in growth rate from accurate estimated breeding values (EBV).

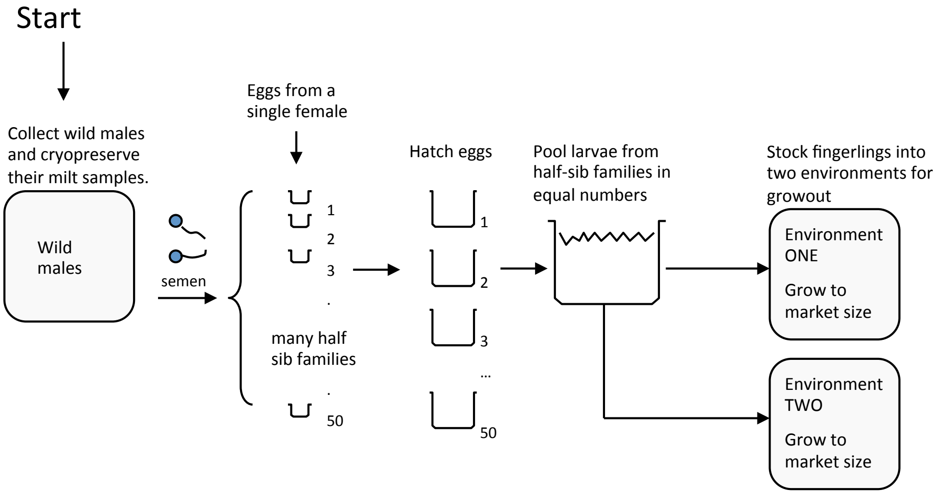

Start
Collect wild males and cryopreserve their milt samples.
Wild males
semen
Eggs from a single female
1
2
3
.
many half sib families
.
50
Hatch eggs
1
2
3
...
50
Pool larvae from half-sib families in equal numbers
Stock fingerlings into pond for growout
Grow to market size
Maturate for natural sex change to females over 2 year period.
At harvest weight grade, weigh and genotype the heaviest fish
Sires identified from genotype data of heaviest fish.
EBVs determined for wild females and young males.
Foundation broodstock stock selected for breeding program.
Hold wild females for EBV assessment
Hold young males for EBV assessment
Initiate multiplication phase of founding broodstock

The initial phase of the proposed breeding program is illustrated in Figure 1.1. The breeding program is novel because it incorporates strain evaluation, progeny testing, and evaluates estimated breeding values prior to forming a foundation population. It can also utilize a higher between family selection intensity than currently used in conventional programs due to the multiplication phase which reduces the rate of inbreeding in subsequent generations.

1.3.2 How can genetic markers be used to improve estimates of genotype by environmental interactions in aquaculture? (This question is addressed in Chapter 3)

In selective breeding programs genotype by environmental interactions (GxE) is a vitally important consideration when assessing the economic benefits. The importance of GxE, when measured by the genetic correlation between the breeding environment and commercial production environment, is that it reflects the proportion of expected genetic gain that is expressed in the production sector. For instance, barramundi (*Lates calcarifer*) may be grown in different environments such as in salt, brackish or sea water, in tropical or subtropical temperatures and in ponds, cages or recirculation tank systems. The genetic correlations of growth between the selected environment and commercial growout may be less than one and therefore not all genetic gains predicted from the selected environment will be expressed in the commercial environment.

Figure 1.2. Proposed mating design to estimate genetic correlations between two environments.


Start
Collect wild males and cryopreserve their milt samples.
Wild males
semen
Eggs from a single female
1
2
3
many half sib families
50
Hatch eggs
1
2
3
...
50
Pool larvae from half-sib families in equal numbers
Stock fingerlings into two environments for growout
Environment ONE
Grow to market size
Environment TWO
Grow to market size

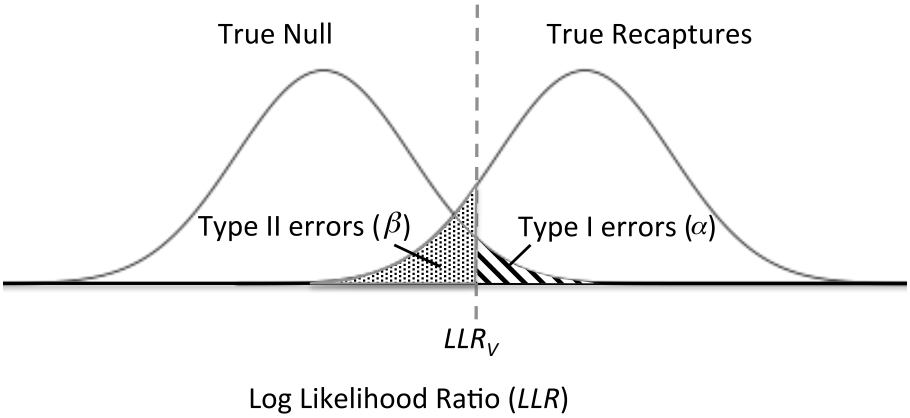
To support a business case in a national fish breeding program the genetic correlation between breeding and commercial environments should be determined as soon as possible. Unfortunately rapid GxE estimates from data collected during single growout period is very difficult to obtain with any precision. I investigated the use of genetic markers to identify sires in a controlled mating design using artificial fertilisation to obtain rapid GxE estimates (Figure 1.2). The rapid estimates of genetic correlation between environments and their standard errors simulated in this study were compared with recent published estimates of genetic correlation as a measure of the usefulness of this methodology.

1.3.3 How can genetic marker identification be improved in genotypes with missing alleles? (This question is addressed in Chapter 4)

In practice the main disadvantages of microsatellites is the existence of null alleles which are alleles that are not observed during standard assays. For example a single null allele at a co-dominant locus would always create an observed homozygote while two null alleles would give missing data at that locus.

This missing genotype data poses a problem when there is insufficient genotype data to distinguish between the log likelihood ratio distributions of true recaptures and true nulls with an overlap occurring between the two distributions (Figure 1.3). In theory there must be a threshold which can minimise the standard error of genotype matches.

Figure 1.3. Distribution of true null and true recaptures based on log likelihood estimates. The threshold value LLR_V is where recapture estimates are determined.

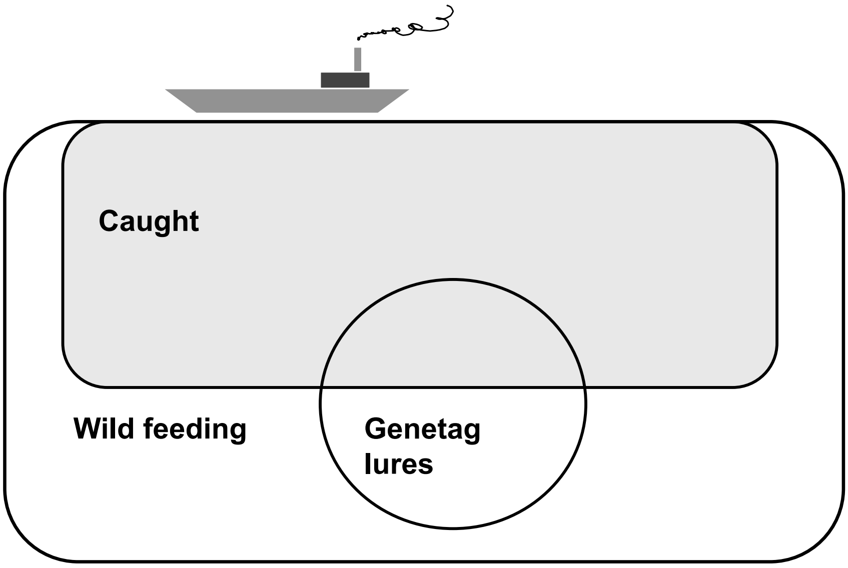


I investigate the use of log likelihood ratios and novel algorithms to estimate the number of Type I and Type II errors. The threshold value (LLR_V) is optimised to obtain the most accurate estimate of recaptures possible by finding a solution that minimises the standard error of the estimates.

1.3.4 How can genetic markers be used to estimate fish abundance? (This question is addressed in Chapter 5)

In a practical application of the theory developed in section 1.3.3 the abundance of wild feeding Spanish mackerel (*Scomberomorus commerson*) below a fishing vessel was determined (Figure 1.4). Special lures called “genetag” lures (Buckworth, 2004) were used to sample DNA tissue of wild feeding fish.

Figure 1.4. Schematic diagram showing the number of fish caught and the number of marked fish using genetag lures that were subsequently caught or remain in the wild.



This study takes the form of a continuous mark-recapture study where DNA tissue from genetag lures are sampled concurrently with the number caught and landed on board. Samples of genotypes from the number caught and the genotypes of genetag lures were determined. The standard error of abundance estimation is determined from the error associated with genetic identification and that of random sampling.

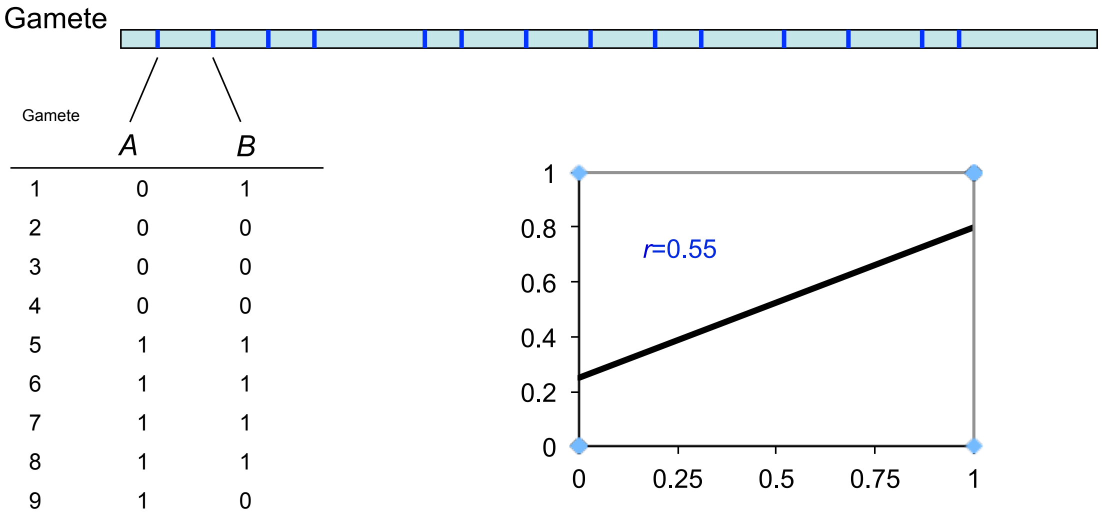
The object of this study was to find finite estimates of abundance when accounting for the total sampling error. Finite estimates of abundance determined by genetic sampling will be a significant achievement and opens up the potential utility of this method for fisheries management.

1.3.5 How can genetic marker estimates of effective population size be improved? (This question is addressed in Chapter 6)

Two new methods that can estimate effective population size (*Ne*) without the need for different samples separated by time over multiple generations include: (i) linkage disequilibrium (LD) estimates (Waples and Do 2008, 2010) and (ii) sibship estimates (Wang 2009; Waples and Waples, 2011).

The LD method measures the correlation of allele frequencies between loci (Figure 1.5) as a measure of genetic drift, which is proportional to effective size (Waples and Do, 2008). Recently Waples and England (2011) have shown that LD estimates of *Ne* are robust to equilibrium migration with $\hat{Ne}$ estimating the local effective population size. Published papers utilizing LD to estimate effective population size have generally been applied to populations of less than 5000 (Waples and England, 2011; Waples and Do, 2010). It is not known how robust *Ne* estimates are when migration is not in equilibrium and how accurate the estimates can be when applied to populations of large effective population size.

Figure 1.5. Correlation of two biallelic loci A and B on nine gametes.



Gamete	*A*	*B*
1	0	1
2	0	0
3	0	0
4	0	0
5	1	1
6	1	1
7	1	1
8	1	1
9	1	0

The object of this study was to find finite estimates of *Ne* using microsatellite markers of Spanish mackerel and to investigate the limitations of the LD Ne estimates as the effective population size increases and when non-target genotypes are sampled.

1.4 References

Abdul-Muneer PM (2014) Application of microsatellite markers in conservation genetics and fisheries management: Recent advances in population structure analysis and conservation strategies. Genetics research International http://dx.doi.org/10.1155/2014/691759

April J, Mayden RL, Hanner RH, Bernatchez L (2011) Genetic calibration of species diversity among North America’s freshwater fishes. Procedings of the Natlional Acadamy of Science. USA 108:10602-10607.

Aykanat T, Johnston S, Cotter D, Cross FT, Poole R, Prodohi PA, Reed T, Rogan G, McGinnity P, Primmer CR (2014) Molecular pedigree reconstruction and estimation of evolutionary parameters in a wild Atlantic salmon river system with incomplete sampling: a power analysis. Evolutionary Biology 14:68. http://www.biomedcentral.com/1471-2148/14/68

Balon EK (2004) About the oldest domesticates among fishes. Journal of fish biology 65:1-27.

Blonk RJW, Komen J, Kamstra A, Crooijmans RPMA, van Arendonk JAM (2009) Levels of inbreeding in group mating captive broodstock populations of Common sole, (*Solea solea*), inferred from parental relatedness and contribution. Aquaculture 289:26-31.

Blonk RJW, Komen J, Kamstra A, van Arendonk JAM (2010) Estimating breeding values with molecular relatedness and reconstructed pedigrees in natural mating populations of common sole, Solea solea. Genetics 184:213-219.

Bravington M, Grewe P (2007) A method for estimating the absolute spawning stock size of SBT, using close-kin genetics. In working document CCSBT-SC/0709/18, CCSBT Scientific Committee Meeting, pp 1-7, Hobart Australia.

Buckworth RC (2004) Effects of spatial stock structure and effort dynamics on the performance of alternative assessment procedures for the fisheries of Northern Australia. PhD Thesis, Univ. of British Columbia, 226p.

Buckworth RC, Ovenden JR, Broderick D, Macbeth GM, McPhereson GR, Phelan MJ (2012) GENETAG: Genetic mark-recapture for real-time harvest rate monitoring: Pilot studies in northern Australia Spanish Mackerel fisheries. Northern Territory Government, Australia. Fishery Report No. 107.

Campana SE (2001) Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. Journal of Fish Biology 59:197-242.

Candy JR, Campbell NR, Grinnell MH, Beacham TD, Larson WA, Narum SR (2015) Population differentiation determined from putative neutral and divergent adaptive genetic markers in Eulachon (*Thaleichthys pacificus*, Osmeridae), an anadromous Pacific smelt. Molecular Ecology Resources doi:10.1111/1755-0998.12400

Chauhan T, Rajiv K (2010) Molecular markers and their applications in fisheries and aquaculture. Advances in Bioscience and Biotechnology 1:281-291.

Christie MR, Marine ML, French RA, Waples RS, Blouin MS (2012) Effective size of a wild salmonid population is greatly reduced by hatchery supplementation. Heredity 109:254-260.

Coates BS, Sumerford DV, Miller NJ, Kim KS, Sappington TW, Siegfried BD, Lewis LC (2009) Comparitive performance of single nucleotide polymorphism and microsatellite markers for population genetic analysis. Journal of Heredity 100:556-564.

Danzmann RG, Ihssen PE (1995) A phylogeographic survey of brook charr (*Salvelinus fontinalis*) in Algonquin Park, Ontario based upon mitochondrial DNA variation. Molecular Ecology 4:681-697

Dichmont CM, Ovenden JR, Berry O, Welch DJ, Buckworth RC (2012) Scoping current and future genetic tools, their limitations and their applications for wild fisheries management. CSIRO, Brisbane, pp. 129.

Doyle RW, Herbinger CM (1994) The use of DNA fingerprinting for high-intensity within-family selection in fish breeding, Proceedings of the 5th World Congress on Genetics Applied to Livestock Production, August 7–12 1994, Vol. 19, Guelph, Canada, pp. 364–371.

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

Duran C, Appleby N, Edwards E, Batley J (2009) Molecular genetics markers: Discovery, applications data storage and visualisation. Current Bioinformatics 4:16-27.

FAO (2007) Marker-assisted selection, Current status and future perspectives in crops, livestock, forestry and fish. Eds. Guimaraes EP, Ruane J, Scherf BD, Sonnino A, Dargie JD. ISBN 978-92-5-105717-9

FAO (2012) The state of World fisheries and aquaculture 2014. Fisheries and Aquaculture Department, Food and Agriculture Organization of the United Nations, Rome. http://www.fao.org/3/a-i3720e.pdf

Fauvelot C, Borsa P (2011) Patterns of genetic isolation in a widely distributed pelagic fish, the narrow-barred Spanish mackerel (*Scomberomorus commerson*). Biological Journal of the Linnean Society 104:886-902.

Fernández ME, Goszczynski DE, Lirón JP, Villegas-Castagnasso EE, Carino MH, Ripoli MV, Rogberg-Muñoz A, Posik DM, Peral-Garcia P, Giovambattista G. (2013) Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, tracability and assessment of parentage in an inbred Angus herd. Genetics and Molecular Biology 36:185-191.

Ford MJ, Williamson KS (2010) The aunt and uncle effect revisited—the effect of biased parentage assignment on fitness estimation in a supplemented salmon population. Journal of Heredity 101:33-41.

Fréon P, Cury P, Shannon L, Roy C (2005) Sustainable exploitation of small pelagic fish stocks challenged by environmental and ecosystem changes: a review. Bulletin of Marine Science 76:385-462.

Frost LA, Evans BS, Jerry DR (2006) Loss of genetic diversity due to hatchery culture practices in barramundi (*Lates calcarifer*). Aquaculture 261:1056-1064.

Gibson JP (1994) Short term gain at the expense of long term response with selection of identified loci. Proceedings of the 5th World Congress on Genetics Applied to Livestock Production, August 7–12 1994, Vol. 21, Guelph, Canada, pp. 201–204.

Glover KA, Hansen LP, Skaala Ø. (2009) Identifying the source of farmed escaped Atlantic salmon (Salmo salar): Bayesian clustering analysis increases accuracy of assignment. Aquaculture, 290:37-46.

Glover KA, Pertoldi C, Besnier F, Wennevik V, Kent M, Skaala Ø. Atlantic salmon populations invaded by farmed escapes: quantifying genetic introgression with a Bayesian approach and SNPs. (2013) BMC Genetics 14:74.

Goddard ME, Hayes BJ (2007) Genomic selection. Journal of Animal Breeding and Genetics 124:323-330.

Godwin RM, Frusher S, Montgomery SS, Ovenden JR (2011). Telomere length analysis in crustacean species: *Metapenaeus macleayi*, *Sagmariasus verreauxi*, and *Jasus edwardsii*. ICES J. Marine Science 68:2053-2058.

Godwin R, Brown I, Montgomery S, Frusher S, Green T, Ovenden J (2012) Telomere dynamics in the Sydney rock oyster (Saccostrea glomerata): an investigation into the effects of age, tissue type, location and time of sampling. Marine Biology 159:77-86.

Gold JR, Richardson LR, Furman C, King TL (1993) Mitochondrial DNA differentiation and population structure in red drum (*Sciaenops ocellatus*) from the Gulf of Mexico and Atlantic Ocean. Marine Biology 116:175-185.

Herbinger CM, Doyle RW, Pitman ER, Paquet D, Mesa KA, Morris DB, Wright JM, Cook D (1995) DNA fingerprint based analysis of parental and maternal effects on offspring growth and survival in communally reared rainbow trout. Aquaculture 137:245-256.

Herbinger CM, O'Reilly PT, Verspoor E (2006) Unravelling first-generation pedigrees in wild endangered salmon populations using molecular genetic markers. Molecular Ecology 15:2261-2275.

Hauser L, Baird M, Hilborn R, Seeb LW, Seeb JE (2011) An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. Molecular Ecology Resources 11:150-161.

Hilborn R, Hively DJ, Jensen OP, Branch TA (2014) The dynamics of fish populations at low abundance and prospects for rebuilding and recovery. ICES journal of marine science. http://icesjms.oxfordjournals.org/content/early/2014/03/30/icesjms.fsu035.full.pdf

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Molecular Ecology 16:1099-1106.

Koch M, Hadfield JD, Sefc KM (2008) Pedigree reconstruction in wild cichlid fish populations. Molecular Ecology 17:4500-4511.

Kozfkay CC, Campbell MR, Heindel JA, Baker DJ, Kline P, Powell MS, Flagg T (2008) A genetic evaluation of relatedness for broodstock management of captive, endangered Snake River sockeye salmon, *Oncorhynchus nerka*. Conservation Genetics 9:1421-1430.

Le Boucher R, Quillet E, Vandeputte M, Lecalvez JM, Goardon, L, Chatain B, Medale F, Dupont-Nivet M (2011) Plant-based diet in rainbow trout (*Oncorhynchus mykiss* Walbaum): Are there genotype-diet interactions for main production traits when fish are fed marine *vs*. plant-based diets from the first meal? Aquaculture 321:41-48.

Le Boucher R, Vandeputte M, Dupont-Nivet M, Quillet E, Ruelle F, Vergnet A, Kaushik S, Allamellou JM, Médale F, Chatain B (2013) Genotype by diet interactions in European sea bass (*Dicentrarchus labrax, L.*) in case of a nutritional challenge on totally plant-based diets. Journal of Animal Science 91:44-56.

Lievens B, Frans I, Heusdens C, Justé A, Jonstrup SP, Lieffrig F, Willems KA (2011) Rapid detection and identification of viral and bacterial pathogens using a DNA array-based multiplex assay. Journal of Fish Diseases 34:861-875.

Lind CE, Evans BE, Taylor JJU, Jerry DR (2009) The consequences of differential family survival rates and equalizing maternal contributions on the effective population size (Ne) of cultured silver-lipper pearl oysters, *Pinctada maxima*. Aquaculture Research 41:1229-1242.

Liu ZJ, Cordes JF (2004) DNA marker technologies and their applications in aquaculture genetics. Aquaculture 238:1-37.

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

Lukacs PM, Burnham KP (2005) Review of capture–recapture methods applicable to noninvasive genetic sampling. Molecular Ecology 14:3909-3919.

Macbeth GM, O'Brien L, Palmer P, Lewer R, Garret R, Wingfield M, Knibb W (2002) Selective breeding in barramundi – Technical Report for the Australian Barramundi Farmers Association August 2002 Information Series QI 02067, 37pp

Macbeth GM (2005) Rates of inbreeding using DNA fingerprinting in aquaculture breeding programs at various broodstock fitness levels – a simulation study. Australian Journal of Experimental Agriculture. 45:893-900.

Macbeth M (2007) Optimising between and within family selection for harvest weight in prawns with restricted harvest size. Association for the Advancement of Animal Breeding and Genetics. September 2007.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. Molecular Ecology 7:639-655.

Meuwissen THE, Sonnesson AK (2004) Genotype-assisted optimum contribution selection to maximize selection response over a specified time period. Genetical Research 84:109-116.

Meyers TR, Sullivan J. Emmenegger E, Follett J, Short S, Batts WN, Winton JR (1992) Identification of viral hemorrhagic septicaemia virus isolated from Pacific cod *Gadus macrocephalus* in Prince William Sound, Alaska, USA. Diseases of Aquatic Organisms 12:167-175.

Miggiano E, De Innocentils S, Ungaro A, Sola L, Crosetti D (2005) AFLP and microsatellites as genetic tags to identify cultured gilthead seabream escapees: data from a simulated floating cage breaking event. Aquaculture International 13:137-146.

Morgan JAT, Harry A, Welch D, Street R, White J, Geraghty PT, Macbeth WG, Broderick D, Tobin A, Simpfendorfer CA, Ovenden JR (2011) Detection of interspecies hybridisation in Chondrichthyes: hybrids and hybrid offspring between Australian (*Carcharhinus* *tilstoni*) and common (*C. limbatus*) blacktip shark found in an Australian fishery. Conservation Genetics 1-9

Morgan JAT, Macbeth M, Broderick D, Whatmore P, Street R, Welch D, Ovenden JR. (2013) Hybridisation, paternal leakage and mitochondrial DNA linearization in three automalous fish (Scombridae) mitochondrion. Mitochondrion Journal, http://www.ncbi.nlm.nih.gov/pubmed/23774068

Ndeda VM, Owiti DO, Aketch BO, Onyango DM (2013) Genetic relatedness of Diplostomum species (*Digenea: Diplostomidae*) infesting Nile tilapia (*Oreochromis niloticus L.*) in Western Kenya. Open Journal of Applied Sciences 3:441-448.

Nielsen EE, Cariani A et al. (2012) Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. Nature communications 3:1-6

Nielsen HM, Sonesson AK, Meuwissen THE (2011) Optimum contribution selection using traditional best linear unbiased prediction and genomic breeding values in aquaculture breeding schemes. Journal of Animal Science 89:630-638.

O'Reilly PT, Carr JW, Whoriskey FG (2006) Detection of European ancestry in escaped farmed Atlantic salmon, *Salmo salar* L., in the Magaguadavic River and Chamcook Stream, New Brunswick, Canada. ICES Journal of Marine Science 63:1256-1262.

Ovenden JR, Macbeth GM, Pope L, Thuesen P, Street R, Broderick D (2014) Translocation between freshwater catchments has facilitated the rapid spread of tilapia in eastern Australia. Biological Invasions, http://link.springer.com/article/10.1007/s10530-014-0754-6

Ovenden JR, Morgan JAT, Kashiwagi T, Broderick D, Salini J (2010) Towards a better management of Australia's shark fishery: genetic analysis reveal unexpected ratios of cryptic blacktip species *Carcharhinus tilstoni* and *C. limbatus.* Marine and Freshwater Research 61:253-262.

Ovenden JR, Berry O, Welch DJ, Buckworth RC, Dichmont CM (2015) Ocean's eleven: a critical evaluation of the role of population, evolutionary and molecular genetics in the management of wild fisheries. Fish and Fisheries 16:125-159.

Ozaki A, Sakamoto t, Khoo S, Nakamura K, Coimbra MR, Akutsu T, Okamoto N (2001) Quantitative trait loci (QTLs) associated with resistance/susceptibility to infectious pancreatic necrosis virus (IPNV) in rainbow trout (*Oncorhynchus mykiss*). Molecular Genetics and Genomics 265:23-31.

Paetkau D, Slade R, Burden M, Estoup A (2004) Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. Molecular Ecology 13:55-65.

Palsboll PJ, Allen J, Berube M, Clapham PJ, Feddersen TP, Hammond PS, Hudson RR, Jorgensen H, Katona S, Larsen AH (1997) Genetic tagging of humpback whales. Nature 388:767-769.

Park LK, Brainard MA, Dightman DA (1993) Low levels of intraspecific variation in the mitochondrial DNA of chum salmon (Oncoryhnchus keta). Molecular Marine Biology Biotechnology 2:362-370.

Park LK. and Moran P (1994) Developments in molecular genetic techniques in fisheries. Reviews in Fish Biology and Fisheries 4:272-299.

Parra-Bracamonte GS, Sifuentes-Rincón AM, De La Rosa-Reyna XF, Arellano-Vera W, Sosa-Reyes B (2011) Inbreeding evidence in a traditional channel catfish (*Ictalurus punctatus*) hatchery in Mexico Electronic Journal of Biotechnology doi:10.2225/vol14-issue6-fulltext-7

Pierce LR, Palti Y, Silverstein JT, Barrows FT, Hallermann EM, Parsons JE (2008) Family growth response to fishmeal and plant-based diets shows genotype x diet interaction in rainbow trout (*Oncorhynchus mykiss*). Aquaculture 278:37-42.

Queensland Government (2010) Crab DNA first lands fisherman $45,000 fine. http://www.cabinet.qld.gov.au/MMS/StatementDisplaySingle.aspx?id=69301

Rabanal HR (1988) History of aquaculture, ASEAN/UNDP/FAO Regional Small-Scale Coastal Fisheries Development Project, Manila, Philippines April 1988

Reister M, Stadler PF, Klemm K (2009) FRANz: reconstruction of wild multi-generation pedigrees. Bioinformatics 25:2134-2139.

Robinson NA, Schipp G, Bosmans J, Jerry DR (2010) Modelling selective breeding in protandrous, batch-reared Asian sea bass (*Lates calcarifer*, Bloch) using walkback selection. Aquaculture Research 41:e643-e655.

Rodriguez-Barreto D, Consuegra S, Jerez S, Cejas JR, Martin V, Lorenzo A (2013) Using molecular markers for pedigree reconstruction of the greater amberjack (*Seriola dumerili*) in the absence of parental information. Animal Genetics 44: 596-600.

Rodzen JA, Famula TR, May B (2003) Estimation of parentage and relatedness in the polyploid white sturgeon (*Acipenser transmontanus*) using a dominant marker approach for duplicated microsatellite loci. Aquaculture 232:165-182.

Ryman N, Utter F (1987) Population genetics and fisheries management. Univ. Washington Press, Seattle.

Salzburger W, Baric S, Sturmbauer C (2002) Speciation via introgressive hybridization in East African cichlids? Molecular Ecology 11:619-625.

Sheikh SI, Berger-Wolf TY, Khokhar AA (2010) Combinatorial reconstruction of half-sibling groups from microsatellite data. Journal of Bioinformatics and Computational Biology 8:337-356.

Shikano T, Taniguchi N (2003) DNA markers for estimation of inbreeding depression and heterosis in the guppy *Poecilia reticulata.* Aquaculture Research 34:905-911.

Smith BR, Herbinger CM, Merry HR (2001) Accurate partition of individuals into full-sib families from genetic data without parental information. Genetics 158:1329-1338.

Sonesson A (2005) A combination of walk-back and optimum contribution selection in fish: a simulation study. Genetics Selection Evolution 37:587-599.

Sonesson AK (2007) Within-family marker-assisted selection for aquaculture species. Genetics Selection Evolution 39:310-317.

Sonesson AK, Meuwissen THE (2009) Testing strategies for genomic selection in aquaculture breeding programs. Genetics Selection Evolution 41:1-9.

Teletchea F, Fontaine P (2014) Levels of domestication in fish: Implications for the sustainable future of aquaculture. Fish and Fisheries 15:181-195.

Tillett BJ, Meekan MG, Broderick D, Field IC, Cliff G, Ovenden JR (2011). Pleistocene isolation, secondary introgression and restricted contemporary gene flow in the pig-eye shark, *Carcharhinus amboinensis* across northern Australia. Conservation Genetics. 13:99-115

Wagner AP, Creel S, Kalinowski ST (2006) Estimating relatedness and relationships using microsatellite loci with null alleles. Heredity 97:336-345.

Wang J, Whitlock MC (2003) Estimating effective population size and migration rates from genetic samples over space and time. Genetics 163:429-446.

Wang J (2004) Sibship reconstruction from genetic data with typing errors. Genetics 166:1963-1979.

Wang J (2009) A new method for estimating effective population sizes from a single sample of multilocus genotypes. Molecular Ecology 18:2148-2164.

Wang J (2011) COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. Molecular Ecology Resources 11:141-145.

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

Waples RS, 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 RS, England PR (2011) Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. Genetics 189:633-44.

Waples RS, Waples RK (2011) Inbreeding effective population size and parentage analysis without parents. Molecular Ecology Resources 11:162-171.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery, GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42:565-569.

Yue GH, Zhu ZY, Lo LC, Wang CM, Lin G, Feng F, Pang HY, Li J, Gong P, Liu HM, Tan J, Chou R (2009) Genetic variation and population structure of Asian seabass (*Lates calcarifer*) in the Asia-Pacific region. Aquaculture 293:22-28.

Yue GH (2012) Recent advances of genome mapping and marker-assisted selection in aquaculture. Fish and Fisheries 15:376-396.

Zhdanova O, Pudovkin A (2008) Nb_HetEx: A program to estimate the effective number of breeders. Journal of Heredity 99:694-695.

Zhanjiang JL (2007) Microsatellite markers and assessment of marker utility. Aquaculture Genome Technologies, Blackwell publishing 1:43-57.

Chapter 2

A novel breeding program for improved growth in barramundi *Lates calcarifer* (Bloch) using foundation stock from progeny-tested parents.

Macbeth G.M., Palmer P.J. (2011) A novel breeding program for improved growth in barramundi *Lates calcarifer* (Bloch) using foundation stock from progeny-tested parents. Aquaculture 318: 325-334.

2.1 ABSTRACT

Rapid genetic gains for growth in barramundi (*Lates calcarifer*) appear achievable through a breeding program using foundation stock from progeny tested broodstock. The potential gains of this novel breeding design were investigated using biologically feasible scenarios tested with computer simulation models. This breeding design involves the production of a large number of full-sib families using artificial mating which are then compared in common growout conditions. The estimated breeding values of their paternal parents are calculated using a binomial probit analysis to assess their suitability as foundation broodstock. The program can theoretically yield faster rates of genetic gain compared to other breeding programs for aquaculture species. Assuming a heritability of 0.25 for growth, foundation broodstock evaluated in two years had breeding values for faster growth ranging from 21% to 51% depending on the genetic diversity of stock under evaluation. As a comparison it would take between nine and twenty-two years to identify broodstock with similar breeding values in an alternative barramundi breeding program.

2.2 Introduction

Barramundi (Lates calcarifer) also known as Asian sea-bass is an increasingly important tropical aquaculture species of the Asia-Pacific region and it is inevitable that breeding programs for this species will soon commence (Macbeth et al., 2002; Wang et al., 2008). Previously published papers showing genetic gains for barramundi could not be identified and only one simulated breeding program has recently been reported (Robinson et al., 2010). At the onset of any new breeding program in aquaculture there is much to be gained by assessing wild genetic diversity as different strains may be more suitable for commercial production. The walk-back selection program for growth rate proposed by Robinson et al. (2010) did not attempt to evaluate the potentially diverse strains from different geographic locations prior to breeding. In species other than barramundi regional sampling of strains has revealed a 52% difference between low and high growth in six strains of *Labeo rohita* (Reddy et al., 2002), a 73% difference in weight in five strains of *Onorhynchus mykiss* (Overturf et al., 2003) and a 104% difference in weight at 105 days between Abbassa and Maryout tilapia strains (Elghobashy, 2001). Differences within lines can also be large with

Brody et al. (1976) reporting differences between the means of half-sibs as large as 30% of the overall mean in *Cyprinus carpio*.

If the breeding values of wild fish from different regions are evaluated prior to establishing a breeding program then there is the potential to make significant genetic gains in the short term. Common practice in barramundi hatcheries is to source replacement broodstock from the wild when required, but some hatcheries are starting to use selected commercially grown fish. As with other aquaculture species a breeding program is usually initiated with one or perhaps combined strains randomly sampled as foundation parents. To address the uncertainty in strain selection a two stage selection approach has been applied in the past where strains are first evaluated (Elghobashy, 2001) prior to selection with the best strains then selected for a foundation population. However, this strategy can be costly and can take considerably more time than simply forming a synthetic line of mixed strains. More recently in barramundi there have been attempts to find genetic markers linked to quantitative trait loci (QTL) of economic importance as a potential means of screening foundation broodstock (Wang et al., 2007). However, again this method is costly and is restricted to a small number of QTL with large effects and so ignores the potentially largest component of genetic variance from cumulative effects of many genes with smaller effects.

In an alternative strategy the high accuracy of progeny testing (Robertson, 1957) could be used to evaluate wild fish. This strategy has been under consideration for many years since Wohlfarth et al. (1961) used it to assess growth in carp. Later Brody et al. (1976) advocated large scale progeny tests but Gjedrem (1983) suggested it would “increase generation interval markedly”. Five years later Gall (1988) mentioned that there was no evidence that progeny testing had been successfully implemented in fish breeding and since then it has received little attention in aquaculture for testing of quantitative traits such as growth rate.

Barramundi is ideally suited to progeny testing because their high fecundity in both females (up to 46 x 10^6 eggs per female; Davis, 1984) and males (up to 10-15 ml of semen; Maneewong, 1986; Palmer, 2000) which allows many progeny to be tested for each parent. Artificial fertilisation would be essential, because large numbers of synchronous natural spawns are difficult to achieve in practice for this species. Artificial fertilisation can also eliminate maternal effects and eliminate age differences which could potentially give fish a size advantage they never relinquish (Tave, 1995). It is proposed to screen potential foundation broodstock for growth using genotype identification and phenotypic observations in a progeny test framework where families are produced by artificial fertilisation.

While copious quantities of semen can be collected from wild males captured on spawning grounds, this is generally only possible a short time before spawning in captive males (Hogan et al., 1987). The potential to strip-spawn eggs and artificially inseminate them with cryopreserved semen from multiple sires has been successfully demonstrated in *L. calcarifer* (Palmer et al*.,* 1993) and enables the progeny of many half-sib families to be grown for accurate breeding value determination of sires. The protandrous sex reversal of *L. calcarifer,* male to female at 3-8 years of age (Moore, 1979; Davis, 1982), offers a novel approach in which wild broodstock females can be accurately evaluated prior to selecting them as foundation parents from progeny testing of their paternal full-sib families. The breeding values of young males can also be determined with relatively high precision by combining information from their own phenotype with the relatively accurate breeding values of their progeny tested sires. Thus, young males can also be evaluated as possible foundation broodstock providing inbreeding is managed.

In general, to manage inbreeding to less than 1% per generation (Goddard, 1992; Meuwissen and Woolliams, 1994) many more broodstock are needed for a selective breeding program compared to the relatively low numbers of broodstock that are needed solely to produce fingerlings for industry. This has been the most important factor that has, up to now, inhibited the establishment of a barramundi breeding program in Australia. In designing a suitable program for selective breeding in barramundi, as with other large aquaculture species, it is important to consider minimising broodstock numbers to manage costs while having sufficient numbers to manage inbreeding.

Minimising broodstock numbers is one method of reducing costs but what is perhaps more important is to maximise early genetic gains (Smith, 1978). An option to improve the rate of early genetic gains is explored using a mating plan with intense between-family selection of potential foundation stock accurately identified from progeny tested wild barramundi.

Stochastic computer methods were used to evaluate the progeny test scheme proposed here under a range of simulated parameter values. It was examined how a progeny test scheme could be implemented for barramundi to estimate heritability, to assess geographic strains, and to achieve rapid genetic gains while managing inbreeding for long term selection. To assist the successful implementation of the scheme a description of husbandry methods is also presented in detail.

2.3 Methods

The computer simulation study of the general breeding design had five basic stages:

(i) collection of wild males and their milt for use in the progeny test,

(ii) evaluation of wild broodstock through a progeny test,

(iii) selection of foundation stock from the very best progeny test sires (which change to females) and the very best young males from the progeny test,

(iv) multiplication of the best foundation stock to create sufficient families to manage long term inbreeding and

(v) performing ongoing selection in subsequent generations.

2.3.1 Progeny test design

The breeding design involved the stripping of eggs from two hatchery females and artificial insemination of multiple sires to initiate a progeny test. The number of progeny tested sires simulated (*NPT*) was either 50, 100 or 200 per dam with no fish between the two progeny test groups being related. In an example with *NPT*=50 and two dams, sires 1 … 50 were crossed with dam one and sires 51 … 100 were crossed with dam two. The number of fingerlings reared to 100 mm from each dam was kept constant at 60,000 to emulate a small hatchery run with the size of each full-sib family equal to 60,000/*NPT*. In practice more than 60,000 should be reared to account for mortality from fertilisation to 100mm and good husbandry should be used to minimise mortality (see section *2.8.*).

The 60,000 fingerlings from each dam were not mixed at any time. For each dam the 100 mm fingerlings were then randomly sub-sampled into two replicates each with a stocking group size per dam (*SGS*) of 5,000, 15,000 or 30,000 resulting in 50,000, 30,000 and none being discarded respectively. The sub-sampling creates some variability in the number in each full-sib family between replicates and emulates a realistic on-farm sampling event. The two replicates were considered a minimum to reduce the risk of experimental failure with only one replicate required to achieve genetic gains.

While the number of barramundi females used in strip spawning (*NFS*) can be varied to suit a number of experimental designs it was demonstrated how a minimum of two dams can be used successfully to achieve large genetic gains. The reasons why only two dams were chosen was that manual stripping of eggs from a female is a demanding task with precision timing of egg collection required (Palmer, 2000) and to demonstrate that two dams is sufficient to manage long term inbreeding.

2.3.2 Simulation of data

Assuming the foundation stock were unrelated, the true breeding value (*A*) was determined using a simulated heritability (h_s^2) which was assigned in different simulations as either 0.2, 0.25, 0.3 or 0.4. The equations were simplified by expressing phenotypic variance $\sigma_P^2 = \sigma_a^2 + \sigma_e^2 = 1$ giving the additive genetic variance $\sigma_a^2 = h_s^2$ and the error variance $\sigma_e^2 = 1 - \sigma_a^2$. True breeding values for *i*=1 . . . (*NPT* x *NFS*) progeny test sires were determined by $A_i = N(0, \sigma_a^2)$, and the true breeding values for *j*=1 . . . *NFS* dams as $A_j = N(0, \sigma_a^2)$. True breeding values for the *k*th offspring (*k*=1 . . . *K*) from the *i* th sire and *j* th dam were determined by $A_{ijk} = (A_i + A_j)/2 + M_{ijk}$ with the Mendelian sampling variation estimated as $M_{ijk} = N(0, \sigma_a^2 / 2)$. The phenotype of the *ijk*th progeny was determined as $P_{ijk} = A_{ijk} + N(0, \sigma_e^2)$.

The sensitivity of the progeny test was examined with a different number of full-sibs (*K*) within each *i* th sire and *j* th dam combination. Using *NPT*=50, simulations undertaken were (a) even full-sib family size *K*=1200 yielding 60,000 fingerlings which were randomly sorted into two replicates of 30,000, (b) variable full-sib family size using five groups of 10 progeny tested sires each with *K* equal to 1920, 1560, 1200, 840 and 480 yielding 60,000 fingerlings which were randomly sorted into two replicates of 30,000 and (c) variable full-sib family survival from 100% to 60% using five groups of 10 progeny tested males each with *K* equal to 1200, 1080, 960, 840 and 720 yielding 48,000 fingerlings prior to random sorting into two replicates of 24,000. In option (c) total survival was assumed to be known with progeny breeding values determined using *SGS*=24,000 samples. Due to random sampling into two replicate groups the number of full-sibs per family were approximately *K*/2 in each replicate. For each combination of parameters simulated the progeny test was repeated in 250 computer trials each with two replicates.

2.3.3 Geographic sampling

An examination was made of effectiveness of the progeny test to identify superior strains collected from genetically isolated populations in the wild. Five strains were simulated with populations having mean genetic differences μ equal to -2, -1, 0, 1 or 2 standard deviations for growth in a commercial environment. When modelling regional sampling *NPT*=50 sires were used comprising 10 sires per strain sampled within each of the two spawning groups. The genetic value of the *i*th progeny tested sire was expressed as $A_i = N(0,\sigma_a^2) + \mu$.

2.3.4 Statistical analysis

Estimation of heritability and estimated breeding values were calculated using a probit sire model which is essentially the “threshold” model in animal breeding (Gianola and Foulley, 1983). The threshold point was determined by the heaviest number genotyped (*NG*) using either 200, 400 or 800 fish selected on phenotype at final harvest for each replicate within each dam. The *NG* fish were genotyped with sire identified (and thus also assigned to its full-sib family) and the record assigned a threshold score of one. Records for all remaining fish, the stocking group size (*SGS*) less the heaviest genotyped (*NG*), were created and assigned a threshold score of zero by assuming each sire contributed to *SGS*/*NPT* full-sib samples in total. In the case where variable family sizes were modelled equal full-sib contributions per sire were assumed when setting up the analysis as the variation in the experimental contributions were not known. Variance parameters were estimated by residual maximum likelihood (REML) by defining the binary score as the random effect in package ASREML (Gilmour et al*.,* 2001) with heritability from the probit analysis calculated as: $h^2 = 4\sigma_s^2 / (\sigma_s^2 + 1)$ where σ_s^2 is the estimated sire variance. In matrix notation the model can be written as $y = Za + e$ where *y* is a vector containing threshold scores of zero or one, *a* is a vector of additive genetic effects of sires, *Z* is the incidence matrix relating random sire effects to observations and *e* is a vector of random errors.

The estimated mean and standard deviation (σ_{h^2}) of h^2 were determined from 500 simulation trials. Assuming the average h^2 estimate was determined from the mean of four estimates obtained from each of two dams by two replicate groups, the standard error of the mean h^2 was determined as: $\sigma_{h^2} / \sqrt{4} = \sigma_{h^2} / 2.$

2.3.5 Evaluation of foundation broodstock

The estimated breeding values from the *i*th progeny tested sire ($\hat{A}_i$) was obtained from the *sln* output file of ASREML (Gilmour et al., 2001). Pre-stocking tank effects (effects prior to 100 mm) were not simulated and assumed to be non-significant as fingerlings are in practice graded and mixed between tanks up to ten times. All variation in fingerling weight prior to stocking was assumed to be non-genetic with procedures put in place to minimise phenotypic variation (section 2.3.8). As barramundi are protandrous hermaphrodites the best wild-captured progeny tested sires, identified as those with the highest $\hat{A}_i$ values for harvest weight, change sex to functional females. To speed up selection response young males reared from the progeny test as foundation males (first generation sires) were mated to the best progeny tested sires which are now females. Within each dam the estimated breeding value of the *ik*th progeny (*k*th full-sib from the *i*th progeny tested sire) was estimated as: $\hat{A}_{ik} = \hat{A}_i / 2 + i_{ik} h^2 (1 - r) / \sqrt{(1 - r.h^2)}$ where h^2 is the heritability, *r*=0.5 is the genetic co-ancestry for full-sibs and i_{ik} is the within-family selection differential in phenotypic standard deviation units. In practice the weight from all offspring from a dam are not individually recorded with i_{ik} estimated using $i_{ik} = (P_{ik} - \overline{P}) / \sigma_P - \hat{A}_{i.} / 2$ where $\hat{A}_i$ is the estimated breeding value of the *i*th sire determined from the probit sire model, P_{ik} is the harvest weight of the *ik*th male and the phenotypic mean ($\overline{P}$) and variance (σ_P^2) of offspring weights determined from sampling. For all simulations $\overline{P} = 0$ and $\sigma_P = 1.0$ with $i_{ik} = P_{ik} - \hat{A}_{i.} / 2$.

2.3.6 Ongoing selection response

After establishment of the progeny test, which is only implemented in the initial generation to assess breeding values of foundation broodstock, ongoing selection was deployed in all following generations using within-family selection. This design assumed a selection intensity of 1:1000 (*i*=3.37 standard deviations) with 24 families and a cumulative inbreeding rate restricted to 1/(2*Ne*)=0.52% per generation where effective population size *Ne*=2*N* (Falconer 1972) given variance in family size is zero and *N*=2x24 parents. For illustrative purposes long-term genetic improvement was expressed as the improvement in two-year harvest weight of 2.5kg, heritability h^2=0.25, a coefficient of variation of 25% (consistent with 19.7% and 27.6% in barramundi; Wang et al., 2008) and a generation length of three years.

Using these parameters a deterministic rate of within-family selection response was estimated as $i\sigma_P h^2(1-r)\Big/\sqrt{(1-r.h^2)}$ =0.55 genetic standard deviations (Falconer, 1972).

2.3.7 Inbreeding

Coefficients of inbreeding of the different designs were determined using methods described in Meuwissen and Luo (1992) implemented in the Animal Breeder's Tool Kit (Golden et al*.,* 1992).

2.3.8 Implementation of progeny test design and husbandry

It is believed a practical design would consist of the collection of at least 100 wild males (e.g. *NPT*=50 for each of two dams) from an assortment of geographic regions and possibly including some broodstock males from industry. This collection would be undertaken during their summer spawning season, so that semen can be simultaneously harvested and cryopreserved in liquid nitrogen. It is suggested that semen be stored in several (at least two) separate cryovials per male (0.2 mL per vial: Palmer et al., 1993). In practice males should be pit tagged and held in captivity pending estimated breeding value (*EBV*) assessment for each sire using the progeny test scheme. At this stage, males in captivity should be screened for noda virus (Parameswaran et al., 2008) by testing semen and blood taken at time of capture or other strategic times during transfer and handling.

Matings for the progeny test are created using the cryopreserved semen of 2*NPT* wild males and strip-spawned eggs from two induced females. These are arranged using a controlled insemination process which creates two unrelated groups of *NPT* full-sib families. Two induced females are seen as a minimum for management of inbreeding. The strip spawning process is made easier through the use of hatchery females with a track record of consistent spawning under repeatable environmental conditions. Females with oocyte diameters of about 0.4 mm can be induced to spawn with single aqueous injections of luteinising hormone-releasing hormone analogues (Garcia, 1989; Garrett and Connell, 1991). Under optimal conditions ovulation generally occurs 36-38 hours after injection, which allows stripping times to be predicted. Using this approach Palmer (2000) achieved multiple successful artificial inseminations using a mechanically-assisted approach to the mixing of stripped eggs and thawed cryopreserved semen. According to this design fertilisation is performed simultaneously in separate chambers each containing one cryovial of thawed semen (0.2 ml) and 20 ml of eggs for each full-sib family. The "dry" method of fertilisation is used where semen and eggs are mixed before an equal volume of seawater is added to activate the sperm and inseminate the eggs. At 3,000 eggs ml^{-1} and 50% mortality from unfertilised and unhatched eggs, this approach would yield approximately 30,000 larvae per full-sib family. Using this procedure, two unrelated groups are created with each group having a different mother and *NPT* different fathers. These two groups are not mixed during the progeny test and are reared separately.

About 10 minutes after mixing with seawater the inseminated eggs are incubated in aerated 3 litre hemispherical bowls until the embryos hatch. This approach allows fertilisation rates to be assessed during the pre-hatch incubation period. Subsequent viability estimates for each bowl can assist in determining the volumetric stocking that can provide approximately equal numbers of hatched larvae from each full-sib family into a communal larval rearing facility. Appropriate biosecurity measures could be applied, such as the disinfection of fertilised eggs with ozonated water to reduce the incidence of infections including noda-viruses and infectious pancreatic necrosis virus (Grotmol et al., 2003).

Barramundi fingerlings typically require grading when they reach a length of 20 mm to avoid cannibalism which occurs when size differences are greater than 67% (Parazo et al., 1991). Gradings may then be required as often as every three to seven days with all fish pooled and sorted on size (girth) into about five tanks. As fish grow and during each grading process the same five tanks are used to reallocate the separate grades. The variance of fingerling size could be minimised by suppressing growth in larger grades using tank temperatures lower than their optimum 28^{o}C to 32^{o}C range (Glencross and Felsing, 2006). For example Bermudes et al. (2010) reported differential growth rates in fingerlings with temperatures below 29^{o}C. The lower phenotypic variance will reduce the need to cull outlier fingerlings to retain approximately the same number in each full-sib family group. At 30 mm size the fingerlings are transferred to larger grow-out facilities with grading continued where necessary until they reach 100 mm. At this length communal stocking for growth assessment occurs.

A final mechanical grading at around 250g to manage competition and cannibalism is recommended. The heaviest 20% of individuals are restocked for final growout, preferably until a commercial harvest weight of up to 2.5kg (or 2 years), with the remaining fish discarded. The heaviest number genotyped (*NG*), within each of the two dams and two replicates, are held in captivity for *EBV* assessment after being pit-tagged, weighed, and tissue sampled for sire identification using genotyping. Caudal fin clips provide non-destructive tissue samples for this identification procedure (Frost et al., 2006).

2.4 Results

2.4.1 Heritability

Heritability (h^2) estimates using a binomial probit analysis from genotyping the heaviest (largest) *NG* progeny were consistent with the simulated heritability (Table 2.1). This confirms that the probit analysis is a suitable way of determining heritability for continuous traits such as harvest weight and that the simulation was implemented correctly.

Table 2.1. Average heritability (h^2) and (standard deviation) from 500 h^2 estimates determined by simulating *NG*=400 genotyped and *NPT*=50 progeny tested sires. The standard error of the mean heritability (±) was determined from four estimates derived from two strip spawns each with two replicates.

	Simulated heritability (h_s^2)					
	0.2		0.3		0.4	
Stocked (*SGS*)	Estimated heritability (h^2)					
15,000	0.200	(0.058)	0.302	(0.072)	0.406	(0.094)
		±0.029		±0.036		±0.048
30,000	0.199	(0.059)	0.300	(0.076)	0.413	(0.089)
		±0.030		±0.038		±0.043

2.4.2 Genetic gains

The progeny test was used to evaluate potential foundation stock prior to the commencement of a breeding program. It was first determined if a variation in the size of each full-sib family (*K*) has an impact on breeding value estimation. In practice the true breeding values are not known but as this is a simulation the true breeding values can be determined from the best animals selected on estimated breeding values from the binomial probit analysis. The average true breeding values of the best progeny-tested fish ranked on ($\hat{A}_i$) and the best young males within each full-sib family ranked on ($\hat{A}_{ijk}$) from one replicate are shown in Table 2.2. The results indicate that the progeny test is a robust evaluation method. The large variation in full-sib family size (Table 2.2b and 2.2c) had little effect on the true breeding values of both young males and progeny tested sires compared to no variation in family size (Table 2.2a). In these runs the best progeny-tested (PT) dams (originally wild males) had breeding values that were less variable than the best young males. The accuracy of breeding value determination for the first generation young males were higher than what could be achieved without pedigree information of $h = \sqrt{0.30} = 0.55$.

The phenotype P_{ijk} of the top eight young males averaged 3.15, 3.10, 3.05, 3.08, 3.02, 3.02, 3.03 and 2.98 phenotypic standard deviations in the control (Table 2.2a) with 50 full-sib families of approximately even size of $\approx$600 at harvest. If the top four of these young males were selected the selection intensity would be 3.10 phenotypic standard deviations. The phenotypes of young males in models of Table 2.2b and Table 2.2c were similar to those of Table 2.2a reported above. If the top four PT dams were selected from the 50 tested, the selection intensity would be 8%.

The results from Table 2.2a have common simulated variables that can be compared with different stocking group size (*SGS*), simulated heritability (h_s^2), number genotyped (*NG*) and number of progeny tested sires per dam (*NPT*) presented in Table 2.3. If foundation broodstock were selected from the top four ranked sires (young males grown during the progeny test) and top four ranked PT dams within each of the two 30,000 stocked spawning groups (Table 2.2a) the 16 breeding values would be superior to the sampled wild population by 0.97$\underline{+}$0.09 phenotypic standard deviations for weight at harvest (or approximately 24% faster growth assuming a coefficient of variation of 25%). In the case where each stocking group size (*SGS*) was reduced to 15,000 the improved broodstock from the top four males and females was reduced to 0.89$\underline{+}$0.10 phenotypic standard deviations (Table 2.3a).

The breeding values of young males were more sensitive to a reduction in heritability than the progeny tested dams. When 30,000 fingerlings (100 mm) were stocked into the growout system and the simulated heritability was reduced from h^2=0.30 (Table 2.2a) to h^2=0.20 (Table 2.3b) the average true breeding values of the top four ranked young males were reduced by 27% while the true breeding values of the top four ranked progeny-tested dams were reduced by 21%.

Table 2.2. Sensitivity of progeny test from three different trials, (a) control – each full-sib family of approximate even size of $\approx$600 at harvest, (b) variable full-sib family size formed from five groups of 10 progeny tested sires each with $\approx$960, $\approx$780, $\approx$600, $\approx$420 and $\approx$240 and (c) simulated variable full-sib family survival of 100% to 60% using five groups of 10 progeny tested males each with $\approx$600, $\approx$540, $\approx$480, $\approx$420 and $\approx$360. Ranks indicate true breeding values (*A*) of potential first-generation (G1) broodstock from the best 'young males' within each full-sib family ranked on $\hat{A}_{ijk}$ and the best 'PT dams' (previously progeny tested sires prior to sex reversal) ranked on $\hat{A}_i$. True breeding values are shown with average and standard deviation (in brackets) determined from 500 REML analysis from data collected in one replicate of a single strip spawn. All trials assume a heritability h^2=0.30, heaviest number of fish genotyped *NG*=400, number of progeny tested sires *NPT*=50 and a stocking group size *SGS*=30,000. The correlation between true breeding values and estimated breeding values ($r_{A,\hat{A}}$) are also listed.

(a)

rank	1	2	3	4	5	6	7	8	$r_{A,\hat{A}}$
young	1.11	1.00	0.97	0.90	0.82	0.85	0.82	0.78	
males	(0.46)	(0.46)	(0.45)	(0.45)	(0.44)	(0.44)	(0.45)	(0.46)	0.62
PT	1.18	0.97	0.84	0.76	0.70	0.63	0.60	0.52	
dams	(0.29)	(0.25)	(0.23)	(0.24)	(0.23)	(0.23)	(0.22)	(0.24)	0.87

(b)

rank	1	2	3	4	5	6	7	8	$r_{A,\hat{A}}$
young	1.12	0.99	0.95	0.92	0.86	0.85	0.81	0.78	
males	(0.54)	(0.56)	(0.56)	(0.54)	(0.55)	(0.53)	(0.53)	(0.53)	0.62
PT	1.16	0.97	0.84	0.77	0.68	0.64	0.57	0.54	
dams	(0.30)	(0.25)	(0.24)	(0.25)	(0.23)	(0.23)	(0.24)	(0.24)	0.87

(c)

rank	1	2	3	4	5	6	7	8	$r_{A,\hat{A}}$
young	1.11	0.97	0.89	0.86	0.83	0.81	0.80	0.79	
males	(0.47)	(0.45)	(0.45)	(0.46)	(0.44)	(0.45)	(0.45)	(0.45)	0.61
PT	1.17	0.93	0.82	0.75	0.66	0.60	0.56	0.52	
dams	(0.29)	(0.26)	(0.24)	(0.22)	(0.23)	(0.24)	(0.23)	(0.25)	0.86

Table 2.3. True breeding values (*A*) of potential first-generation (G1) broodstock from a single dam with the best 'young males' within each full-sib family ranked on $\hat{A}_{ijk}$ and the best 'PT dams' (previously progeny tested sires prior to sex reversal) ranked on $\hat{A}_i$. Average and standard deviation (in brackets) determined from 500 REML analysis at a given stocking group size (*SGS*), simulated heritability (h_s^2), heaviest number of fish genotyped (*NG*) and number progeny tested (*NPT*). The correlations between true breeding values and estimated breeding values ($r_{A,\hat{A}}$) are also listed.

(a) Changing *SGS* with h_s^2=0.30, *NG*=400 and *NPT*=50.

SGS	rank	1	2	3	4	5	6	7	8	$r_{A,\hat{A}}$
5,000	young	1.03	0.90	0.83	0.80	0.79	0.72	0.73	0.71	
	males	(0.46)	(0.47)	(0.46)	(0.47)	(0.47)	(0.47)	(0.46)	(0.48)	0.60
	PT	1.12	0.90	0.79	0.73	0.64	0.61	0.57	0.52	
	dams	(0.34)	(0.29)	(0.28)	(0.28)	(0.29)	(0.27)	(0.28)	(0.28)	0.83
15,000	young	1.10	0.95	0.94	0.87	0.83	0.80	0.78	0.74	
	males	(0.46)	(0.46)	(0.45)	(0.43)	(0.43)	(0.43)	(0.43)	(0.44)	0.62
	PT	1.16	0.94	0.83	0.74	0.69	0.61	0.56	0.51	
	dams	(0.34)	(0.27)	(0.25)	(0.25)	(0.24)	(0.25)	(0.24)	(0.24)	0.86

(b) Changing h_s^2 with *NG*=400, *NPT*=50 and *SGS* =30,000.

h_s^2	rank	1	2	3	4	5	6	7	8	$r_{A,\hat{A}}$
0.20	young	0.84	0.73	0.68	0.64	0.58	0.59	0.56	0.53	
	males	(0.39)	(0.37)	(0.38)	(0.37)	(0.39)	(0.37)	(0.42)	(0.39)	0.56
	PT	0.94	0.77	0.66	0.60	0.54	0.50	0.46	0.41	
	dams	(0.25)	(0.23)	(0.21)	(0.21)	(0.23)	(0.22)	(0.22)	(0.22)	0.83
0.40	young	1.42	1.27	1.21	1.17	1.08	1.10	1.08	1.01	
	males	(0.52)	(0.52)	(0.51)	(0.52)	(0.52)	(0.52)	(0.51)	(0.52)	0.67
	PT	1.38	1.15	1.00	0.89	0.81	0.74	0.68	0.63	
	dams	(0.31)	(0.26)	(0.25)	(0.24)	(0.23)	(0.23)	(0.23)	(0.24)	0.89

(c) Changing *NG* with h_s^2=0.30, *NPT*=50 and *SGS* =30,000.

NG	rank	1	2	3	4	5	6	7	8	$r_{A,\hat{A}}$
200	young	1.11	0.98	0.93	0.89	0.85	0.83	0.82	0.79	
	males	(0.45)	(0.46)	(0.47)	(0.47)	(0.44)	(0.46)	(0.46)	(0.45)	0.60
	PT	1.16	0.94	0.81	0.71	0.66	0.61	0.54	0.49	
	dams	(0.31)	(0.27)	(0.26)	(0.27)	(0.26)	(0.25)	(0.27)	(0.27)	0.81
800	young	1.12	1.04	0.94	0.91	0.87	0.84	0.85	0.79	
	males	(0.47)	(0.48)	(0.47)	(0.46)	(0.46)	(0.46)	(0.45)	(0.46)	0.63
	PT	1.21	1.00	0.88	0.78	0.70	0.64	0.59	0.54	
	dams	(0.28)	(0.25)	(0.22)	(0.22)	(0.22)	(0.21)	(0.20)	(0.20)	0.92

(d) Changing *NPT* with h_s^2=0.30, *NG*=400 and *SGS* =30,000.

NPT	rank	1	2	3	4	5	6	7	8	$r_{A,\hat{A}}$
100	young	1.15	1.06	1.00	0.98	0.90	0.90	0.88	0.82	
	males	(0.44)	(0.47)	(0.45)	(0.44)	(0.46)	(0.44)	(0.47)	(0.44)	0.60
	PT	1.27	1.10	0.97	0.88	0.83	0.79	0.76	0.70	
	dams	(0.31)	(0.26)	(0.26)	(0.25)	(0.26)	(0.26)	(0.25)	(0.26)	0.80
200	young	1.15	1.04	0.99	0.96	0.90	0.89	0.91	0.88	
	males	(0.48)	(0.47)	(0.46)	(0.46)	(0.47)	(0.46)	(0.46)	(0.46)	0.58
	PT	1.30	1.14	1.04	0.97	0.90	0.87	0.84	0.82	
	dams	(0.32)	(0.30)	(0.31)	(0.32)	(0.30)	(0.31)	(0.31)	(0.31)	0.69

Increasing the number genotyped per dam (*NG*) from 200 to 800 increased the breeding values of the top four ranked young males and top four PT dams by 3% and 7% respectively (Table 2.3c).

Theoretically more than 50 sires per dam can yield higher genetic gains (Table 2.2a and Table 2.3d). Compared to using 200 sires per dam (Table 2.3d) instead of 50 sires (Table 2.2a) the average of the top four young males and top four PT dams improved by 10%.

2.4.3 Genetic gains from geographic sampling

The true breeding values were significantly higher in Table 2.4 than those in Tables 2.2 and 2.3 because the genetic merit of superior strains was detected through the progeny test. The breeding values of all sires in Table 2.4 were lower than the best geographic region (μ= 2) as their mother used in the strip spawn had an average breeding value of zero and contributed half her genes to the sires. Additional genetic response could be achieved by sampling (and possibly re-evaluating) foundation sires from the best performing geographic region. Even though more progeny test sires would theoretically yield a higher response (Table 2.3d) the simulations presented in Table 2.4 were derived using a more manageable 50 sires per dam.

If foundation broodstock were selected from the top four ranked sires (young males grown during the progeny test) and the top four ranked PT dams within each of the two spawning groups the 16 breeding values would average 2.12±0.09 phenotypic standard deviations better than the average strain (Table 2.4) and be approximately 53% faster growing assuming a coefficient of variation of 25% and h^2=0.30.

When simulating a heritability of 0.25 the 16 breeding values averaged 2.05±0.09 phenotypic standard deviations or 51% faster growth (*NPT*=50, *SGS*=30,000, *NG*=400) with $r_{A\hat{A}}$ equal to 0.86 and 0.92 for young males and PT dams respectively.

Table 2.4. The effect of geographic sampling with five simulated strains. True breeding values (*A*) of potential first-generation (G1) broodstock from a single strip spawn with the best 'young males' within each full-sib family ranked on $\hat{A}_{ijk}$ and the best 'PT dams' (previously progeny tested sires prior to sex reversal) ranked on $\hat{A}_i$. Average and standard deviation (in brackets) determined from 500 REML analysis at a given stocking group size (*SGS=30,000*), simulated heritability (h_s^2=0.30), heaviest number of fish genotyped (*NG=400*) and number progeny tested (*NPT=50*). The correlation between true breeding values and estimated breeding values ($r_{A,\hat{A}}$) are also listed.

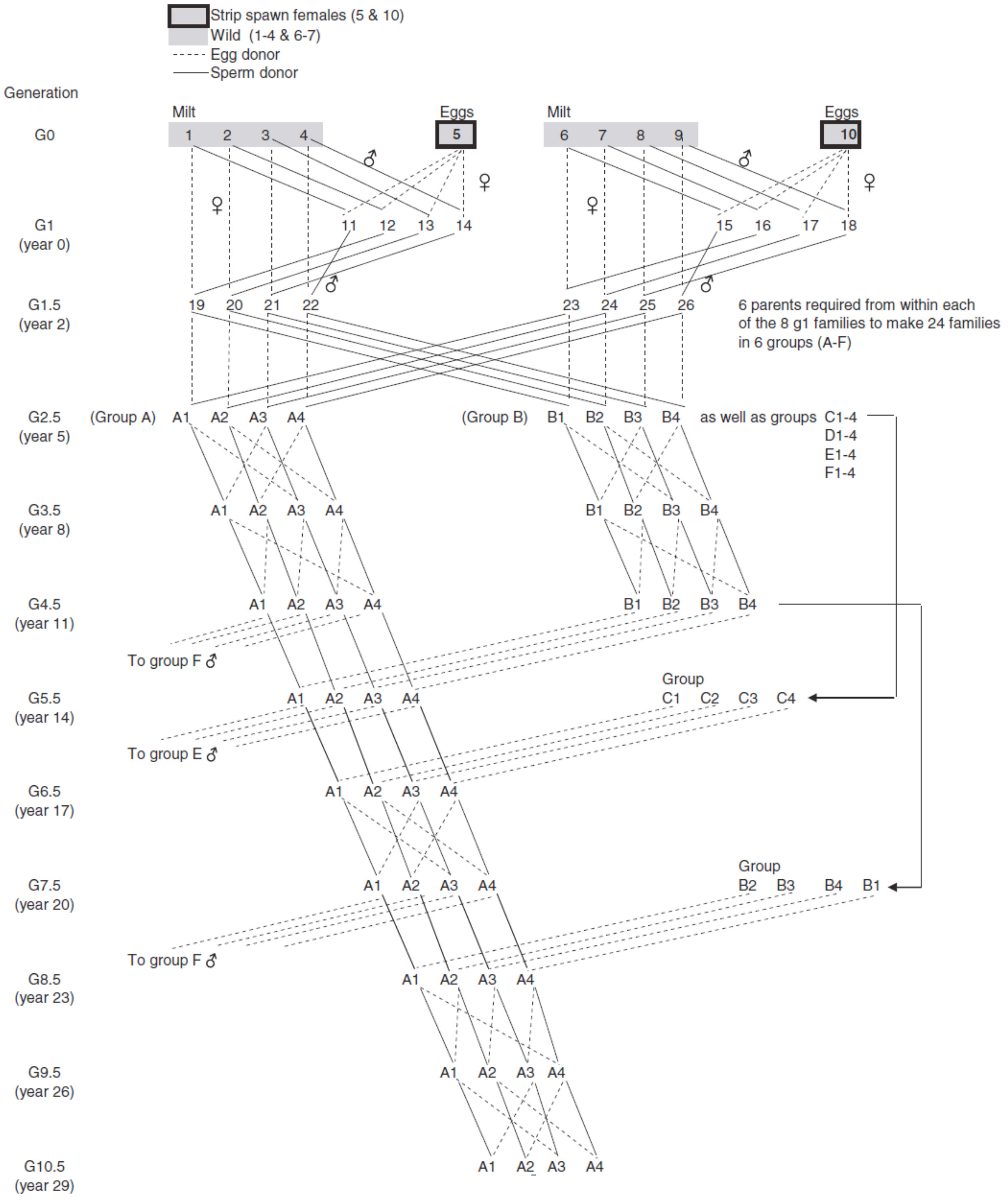
rank	1	2	3	4	5	6	7	8	$r_{A,\hat{A}}$
young	1.90	1.78	1.69	1.62	1.56	1.51	1.47	1.39	
males	(0.46)	(0.45)	(0.44)	(0.44)	(0.46)	(0.45)	(0.46)	(0.47)	0.86
PT	2.82	2.55	2.35	2.21	2.08	1.98	1.88	1.74	
dams	(0.32)	(0.29)	(0.26)	(0.25)	(0.24)	(0.26)	(0.25)	(0.27)	0.92

2.4.4 Managing inbreeding for long-term selection response

In order to maximise genetic gains it is necessary to balance the need to select the very best evaluated fish with the need to minimise inbreeding at below 1% per generation (Goddard, 1992; Meuwissen and Woolliams, 1994). A mating plan that satisfies this inbreeding constraint was designed (Figure 2.1). The plan is broken down into five phases with a brief description of each phase given below.

1) *Generation 0 (collection phase).* Collection of fish for the progeny test which is made up of two hatchery females (for use in strip spawning) and 100 males chosen from the wild (industry males could also be evaluated from semen collected).
2) *Generation 1 (progeny test phase).* Progeny are created from the eggs of two dams artificially mated with semen from 100 wild unrelated males using 50 of these per spawn. Only the best four wild males from each dam (males 1, 2, 3, 4 and 6, 7, 8, 9) are selected for future breeding at harvest weight.
3) *Generation 1.5 (foundation phase – 8 families).* At two years of age four young males with the highest *EBV*s are selected within each of the two 50 full-sib family groups from first generation (G1 males 11, 12, 13, 14 and 15, 16, 17, 18). The former wild males with the highest *EBV*s, now progeny-tested wild females, (G0) are backcrossed to the young males to create a total of 8 foundation families numbered 19 to 26. The response from the backcross reflects 1.5 generations of selection which was called G1.5 with the next generation with both parents from G1.5 called G2.5.
4) *Generation 2.5 (multiplication phase – 24 families).* Six animals are selected as broodstock replacements from within each of the 8 foundation families in G1.5. This yields 48 broodstock fish (6 x 8=48), or 24 broodstock pairs to produce 24 families for ongoing selection. The families are divided into six groups (A, B, C, D, E, F) each containing four families.
5) *Generation 3.5 onwards (ongoing selection phase – 24 families).* There are many mating designs possible. An example of how matings can be made between groups and also within groups to manage inbreeding and minimise gene flow between groups to improve biosecurity risks is illustrated (Figure 2.1). As indicated one male and one female are used from each family (within-family mating design). The mating design is flexible; for example, two male parents from one family could be mated to two females from other families. This system accrues inbreeding at a rate of 0.52% per generation.

Figure 2.1. A possible mating design illustrating how inbreeding can be managed by selecting the best progeny test (PT) dams and best young males from the progeny test in the first generation (G1) with multiplication of families occurring in the backcross generation in year two (G1.5).

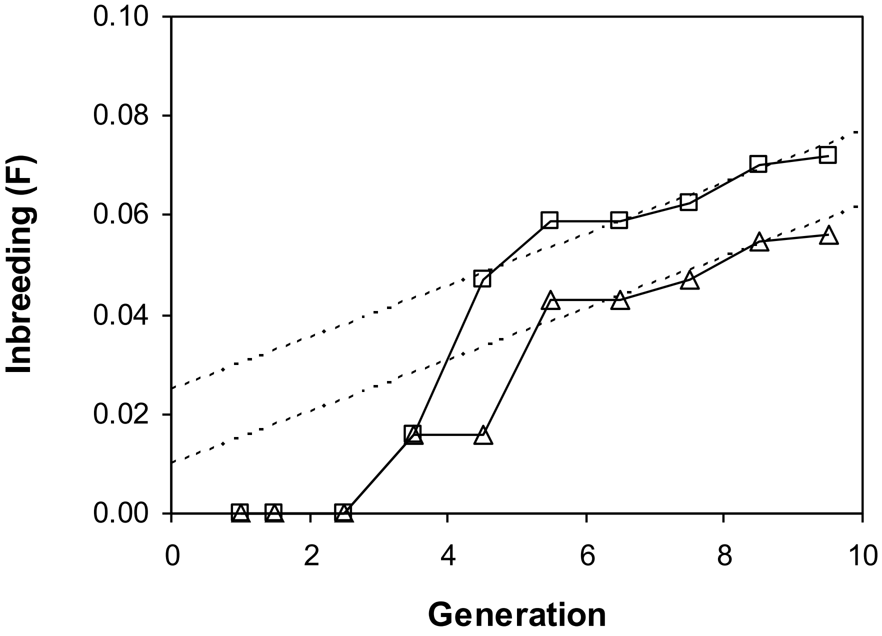


The mating design uses 16 founding fish comprising four of the best sires and four of the best dams from each of the two spawning groups. This breeding plan is one of many alternatives and requires a wild backcross labelled generation G1.5. The progeny from the backcross are multiplied into 24 families which are maintained in subsequent generations with two offspring per family contributing to each generation. The 24 families are formed during the multiplication phase (G2.5) by selecting six parents from each of eight full-sib families produced from the 16 founding parents.

The mating plan illustrated in Figure 2.1 assumes all eight first-generation sires being related to the first generation dams. The simulated results showed that these eight sires will only be related to all first generation dams 7% of the time (h_s^2=0.30, *NG*=400, *NPT*=50, *SGS*=30,000) with this increasing to 21% when genetic strains were modelled. Constraining the best first generation sires (young males from the progeny test) so that none are related to first generation dams reduced inbreeding. This also reduced the selection response in the first generation by 16% and 7% of gains made with 15,000 and 30,000 fish stocked respectively with one strain sampled, h_s^2=0.30, *NG*=400 and *NPT*=50.

Cumulative inbreeding is expected to lie within the two extremes shown in Figure 2.2 with long-term inbreeding accumulating at the rate of 0.52% per generation. With 24 families inbreeding can be completely avoided up until 3.5 generations where different levels of relatedness from the sampling of sires and dams create two divergent patterns (Figure 2.2). Figure 2.2 also indicates the extrapolation of the long term inbreeding rate back to generation zero which is an estimate of the cost of the progeny test in terms of inbreeding and is estimated at 1.0% to 2.5%. This means that the theoretical asymptotic selection response would be 97.5% to 99.0% compared to the response from the within-family selection program had the progeny test not been used to evaluate and multiply superior foundation broodstock.

Figure 2.2. Average inbreeding level over the first 10 generations of selective breeding assuming none (Δ) or all (□) of the eight sires (young males from the progeny test) in the first generation were related to the eight progeny test dams. The dashed line illustrates the long term inbreeding rate at 0.52% per generation showing the cost of the progeny test at between 1.0% and 2.5% inbreeding at generation zero.

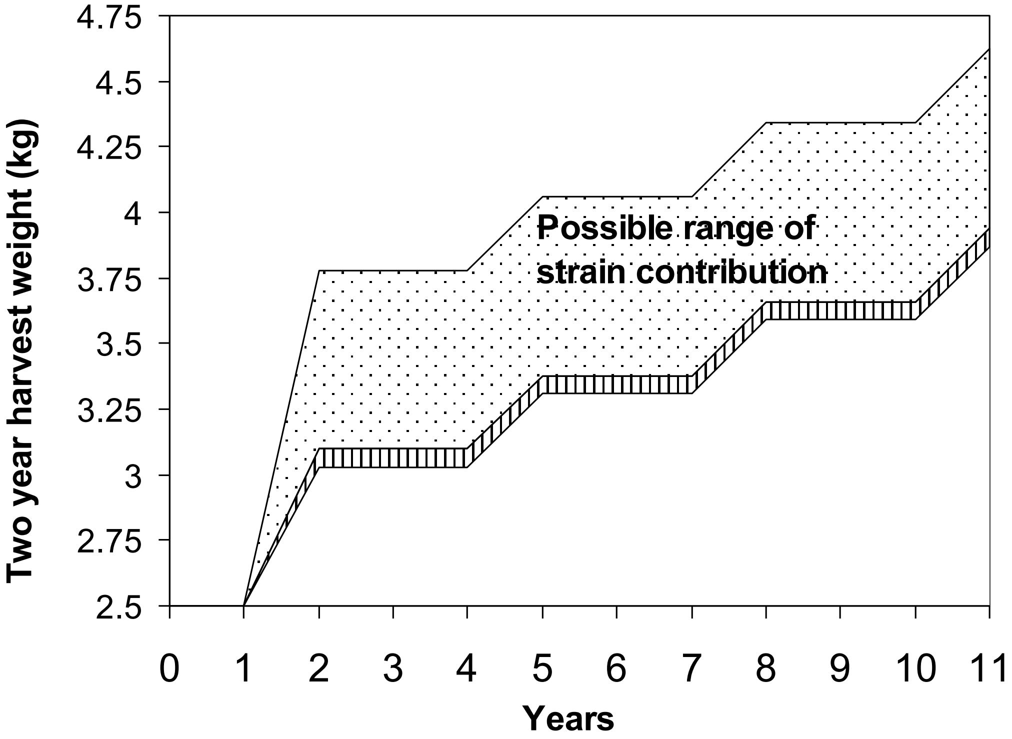


2.4.5 Long-term selection response

The additive components of selective improvement from progeny testing followed by within-family selection are illustrated in Figure 2.3. The improvement shown in this graph assumes a similar mating design to that shown in Figure 2.1 with the top four males and top four females from each of the two dams used as foundation stock.

The genetic contribution of each dam donor is one-eighth of the founding population genotypes and one round of selection of these females contributes a small component of total genetic gains (Figure 2.3). A large component of the progeny test gains came from the ability to identify superior strains from diverse geographic locations. The foundation population had true breeding values for faster growth, shown in year two of Figure 2.3 of between 21% (no strain differences) and 51% (with strain differences) of the base population.

Figure 2.3. Broodstock breeding values for two-year harvest weight from the progeny test showing the range of expected improvement from geographical sampling (shaded area with dots) and the cumulative improvement from a within-family selection program using a high selection intensity of 1:1000 and a generation length of three years. The optional additive gains from sourcing strip-spawned donor females from one round of selection (area with vertical lines) is included. Simulated parameters were: h_s^2=0.25, *NG*=400, *NPT*=50, *SGS*=30,000 and a coefficient of variation of 25%.



2.5 Discussion

There is nothing new about progeny testing in aquaculture (Wohlfarth et al., 1961) or in using artificial mating to estimate genetic parameters (Dupont-Nivet et al., 2008). What is novel in this study is the combination of strain evaluation, genetic parameter estimation and progeny testing to evaluate potential foundation broodstock using a carefully controlled mating design.

The necessity to regularly grade barramundi until they reach a size of about 250 g makes traditional genetic parameter estimation challenging. This problem was overcome by sampling only the heaviest animals for genetic parameter estimation and using a binomial probit analysis which also enabled an accurate progeny test evaluation of the best sires.

Also novel is the way accurately evaluated animals are multiplied to manage inbreeding with the object of obtaining rapid genetic gains. Not seen in any other fish breeding design is the high between-family selection intensity of 8% achieved in the first generation which is much higher than what is conventionally achieved.

2.5.1 Heritability

Using natural matings, where eggs and sperm are released by the fish directly into the water column of the spawning tank, Wang et al. (2008) estimated the heritability of harvest weight for barramundi as 0.22±0.16 and 0.25±0.18 from two factorial crosses. These heritability estimates could have been underestimated due to selective culling of graded fingerlings (Blonk et al., 2010). Despite the short duration of the progeny test design it gave a more precise heritability estimate of 0.30±0.04 from simulated data as many more sires could be evaluated in an experimental design that also minimised the variance in sib numbers.

2.5.2 Genetic gains

The upper range of genetic gains predicted by the models was due to geographic sampling of wild strains. Large differences between strains are possible (Elghobashy, 2001; Overturf et al., 2003). Using the difference between two tilapia strains of 87 g and 178 g (Elghobashy, 2001) and assuming a 25% heritability, a 25% coefficient of variation and estimating the phenotypic standard deviation from the average weight of these strains, the difference between the two strains in genetic standard deviations is equal to 5.5. Similar calculations from five strains of *Oncorhynchus mykiss* (Overturf et al., 2003) revealed 4.1 genetic standard deviations between the fastest and slowest growing strains fed ad libitum. The simulations assumed a maximum strain difference of 4 genetic standard deviations. Considering the estuarine spawning of semi-isolated populations of barramundi that occurs over 8,000 km along the northern Australian coastline (Shaklee and Salini, 1985; Keenan, 1994; Chenoweth, 1998), with the species distribution extending across the Indo-Pacific region (Norfatimah et al., 2009; Yue et al., 2009), it seems reasonable that the weight range in the simulated strains is a realistic upper limit of what could be expected. The magnitude of strain differences in barramundi is of course impossible to predict without experimental trials.

What is known with greater precision is the genetic gains expected from the progeny test assuming no strain differences and known heritability. In an alternative barramundi breeding program designed with a different set of constraints Robinson et al. (2010) reported a 7.1% gain per generation. The progeny test design (Figure 2.2) is capable of identifying in just two years superior broodstock with breeding values that could take between, 3.0 and 7.2 generations to achieve with this alternative design, or 9 to 22 years assuming a three year generation interval.

The reason that this program works so well in the first generation is that

(i) it can manage a much higher between-family selection intensity compared to what can be managed in subsequent generations in any other contemporary design with little effect on inbreeding when utilised with the mating design,

(ii) the progeny test is more accurate than first generation phenotypic selection used in all other contemporary designs, particularly when the heritability is low, and

(iii) the upper range of genetic response from the progeny test was due to the ability to select the most favourable alleles from the best performing strains, rather than crossing strains prior to evaluation. In this way an elite foundation population from the best wild strains has been formed.

2.5.3 Managing inbreeding for long-term selection response

There are challenges in rearing barramundi in captivity and strip spawning and artificial fertilisation is yet another challenge. However, previous work has proven its feasibility (Palmer et al., 1993). It is likely that after the progeny test artificial fertilisation will only be used sparingly and only when necessary to manage desired family matings that could not otherwise be achieved naturally.

Perhaps the largest challenge in achieving a desirable long-term response to selection is to manage inbreeding through natural mating. A possible mating plan given in the Appendix illustrates how inbreeding can be minimised through chosen family matings. In practice more flexibility in mating design may be required. After the multiplication phase (see Appendix), the long-term inbreeding can be managed by following two simple rules: manage individual matings so that inbreeding is less than 25% and use two different parents in each of the 24 families to contribute to the next generation.

Managing inbreeding using walk-back selection (Sonesson, 2005) may not be practical in barramundi due to the difficulty in synchronising spawns and the highly variable sire contribution rate (Frost et al., 2006). The challenge is exacerbated as barramundi are cannibalistic (Parazo et al., 1991) and it may not be practical to mix families of ages more than one day old as they may have a size advantage they never relinquish (Tave, 1995). This is one reason why the simpler within-family selection design seems more practical for barramundi which when applied with a very high selection intensity (1:1000) can theoretically yield genetic gains of 11.4% per generation (*CV*=25%, h^2=25%).

2.5.4 Managing matings for long term selection with protandry (sex change)

A further improvement to the current design could be achieved if the sex of the protandrous broodstock were closely monitored. Assuming 24 families, and perhaps three to four fish per family maintained as potential broodstock, matings should be possible as soon as sufficient numbers of females are available. It would not matter if two males (or two females) were selected from one family so long as the long term inbreeding rate is minimised by using two parents per family to contribute to each generation. Short term inbreeding can be managed by avoiding the mating of close relatives. The advantage of this approach is that it can be applied to minimise generation length from 3 years (males 2 years and females 4 years) to perhaps 2.5 years, potentially yielding a further 20% improvement in the rate of genetic gain. To improve the chances of detecting functional young females more (three or four) broodstock per family would be useful in this protandry mating design.

2.5.5 Cost considerations in design

The cost of running aquaculture breeding programs is high, particularly for species with large broodstock size such as barramundi, with costs increasing in proportion to the size of broodstock facilities required to maintain the program. The progeny test scheme to select foundation stock, using high-intensity between-family selection, was designed as a one-off procedure. The scheme continues with a long-term within-family selection program using only 24 families. As such the scheme achieves significant selective gains over few generations while minimising the number of broodstock that have to be maintained, hence capping the ongoing costs of the breeding program which leads to larger discounted net returns.

As an alternative to long-term within-family selection, it is possible to implement between-family selection at a later stage to allow selection for traits that can only be recorded on sacrificed fish. However, this design would increase cost several fold because more broodstock are required. As such, it would require an economic assessment of additional costs and projected returns prior to a decision being made on its implementation. In comparison with the model proposed by Robinson et al. (2010), which requires 100-200 broodstock (50-100 full-sib families), the new design described in this chapter would require 72-96 broodstock assuming three to four broodstock per family were maintained to guarantee selection of two offspring from each of the 24 families. Therefore the costs of maintaining broodstock in the new program could be as little as half of that proposed by Robinson et al. (2010). In addition, after the progeny test phase, this program does not require the on-going costs and logistical problems associated with genotyping using walk-back selection.

2.6 Conclusion

The *Lates calcarifer* progeny test design described in this chapter theoretically identifies superior foundation broodstock in two years. The same level of genetic improvement could take nine to 22 years of selection in other proposed barramundi breeding programs. Despite the extra effort involved in sourcing potential foundation stock for the progeny test, and the challenging nature of the husbandry activities that are required, this scheme compares favourably with the risks that accrue over a much longer time frame that would otherwise be needed to provide similar benefits in contemporary designs.

2.7 Acknowledgements

The progeny test approach was inspired from discussions with Dr Roger Lewer while he was an employee of the Queensland Department of Primary Industries and Fisheries. Thanks to the many useful comments from referees and reviewers. Funding support was provided by the Department of Employment, Economic Development and Innovation.

2.8 References

Bermudes M, Glencross B, Austen K, Hawkins W (2010) The effects of temperature and size on the growth, energy budget and waste outputs of barramundi (*Lates calcarifer*). Aquaculture 306:160-166.

Blonk RJW, Komen H, Kamstra A, van Arendonk JAM (2010) Effects of grading on heritability estimates under commercial conditions: A case study with common sole, *Solea solea*. Aquaculture 300:43-49.

Brody T, Moav R, Abramson ZV, Hulata G, Wohlfarth G (1976) Applications of electrophoretic genetic markers to fish breeding. II. Genetic variation within maternal half-sibs in carp. Aquaculture 9:351-365.

Chenoweth SF, Hughes JM, Keenan CP, Lavery S (1998) Concordance between dispersal and mitochondrial gene flow: isolation by distance in a tropical teleost, *Lates calcarifer* (Australian barramundi). Heredity 80:187-197.

Davis TLO (1982) Maturity and sexuality in barramundi, *Lates calcarifer* (Bloch), in the Northern Territory and South-eastern Gulf of Carpentaria. Australian Journal of Marine Freshwater Research 33:529-545.

Davis TLO (1984) Estimation of fecundity in barramundi, *Lates calcarifer* (Bloch), using an automatic particle counter. Australian Journal of Marine Freshwater Research 35:111-118.

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

Elghobashy H (2001) Aquaculture genetics research in Egypt. p. 29-34, In: Gupta, MV, Acosta, BO (Eds.) Fish genetics research in member countries and institutions of the International Network on Genetics in Aquaculture. ICLARM Conf. Proc. 64, 179 p.

Falconer DS (1972) Introduction to quantitative genetics. Oliver and Boyd, Edinburgh.

Frost LA, Evans BS, Jerry DR (2006) Loss of genetic diversity due to hatchery culture practices in barramundi (*Lates calcarifer*). Aquaculture 261:1056-1064.

Gall GAE (1988) Heritability and selection schemes for rainbow trout: female reproductive performance. Aquaculture 73:57-66.

Garcia LMB (1989) Spawning response of mature female sea bass, *Lates calcarifer* (Bloch), to a single injection of luteinizing hormone-releasing hormone analogue: effect of dose and initial oocyte size. Journal of Applied Ichthyology 5:177-184.

Garrett RN, Connell MRJ (1991) Induced breeding of barramundi. Austasia Aquaculture 5:10-12.

Gianola D, Foulley JL (1983) Sire evaluation for ordered categorical data with a threshold model. Genetic Selection Evolution 15:201-224.

Gilmour AR, Cullis BR, Welham SJ, Thompson R (2001) ASREML User's Manual. New South Wales Agriculture, Orange Agricultural Institute, Orange, NSW, Australia.

Gjedrem T (1983) Genetic variation in quantitative traits and selective breeding in fish and shellfish. Aquaculture 33:51-72.

Glencross BD, Felsing M (2006) Influence of fish size and water temperature on the metabolic demand for oxygen by barramundi, *Lates calcarifer* (Bloch), in freshwater. Aquaculture Research 37:1055-1062.

Goddard ME (1992) Optimal effective population-size for the global population of black-and-white dairy-cattle. Journal of Dairy Science 75:2902-2911.

Golden BL, Snelling WM, Mallinckrodt CH (1992) Animal breeder's tool kit user's guide and reference manual. Colorado State Univ. Colorado State University Agriultural Experimental Station Technical Bulletin. LTB92-2

Grotmol S, Dahl-Paulsen E, Totland GK (2003) Hatchability of eggs from Atlantic cod, turbot and Atlantic halibut after disinfection with ozonated seawater. Aquaculture 221:245-254.

Hogan AE, Barlow CG, Palmer PJ (1987) Short-term storage of barramundi sperm. Australian Fisheries 46:18-19.

Keenan CP (1994) Recent evolution of population structure in Australian Barramundi, *Lates calcarifer* (Bloch): An example of isolation by distance in one dimension. Australian Journal of Marine Freshwater Research 45:1123-1148.

Macbeth GM, O'Brien L, Palmer P, Lewer R, Garret R, Wingfield M, Knibb W (2002) Selective breeding in barramundi – Technical Report for the Australian Barramundi Farmers Association August 2002 Information Series QI 02067, 37pp.

Maneewong S (1986) Induction of spawning in sea bass (*Lates calcarifer*) in Thailand. In: Management of wild and cultured sea bass/barramundi (*Lates calcarifer*): Proceedings of an international workshop held in Darwin, N.T. Australia, 24-30 September 1986. (Eds. J. W. Copland and D. L. Grey) pp. 116-119. (ACIAR Proceedings No. 20).

Meuwissen THE, Luo Z (1992) Computing inbreeding coefficients in large populations. Genetics Selection Evolution 24:305-313.

Meuwissen THE, Woolliams JA (1994) Effective sizes of livestock populations to prevent a decline in fitness. Theoretical and Applied Genetics 89:1019-1026.

Moore R (1979) Natural sex inversion in giant perch (*Lates calcarifer*). Australian Journal of Marine Freshwater Research 30:803-813.

Norfatimah MY, Siti Azizah MN, Othman AS, Patimah I, Jamsari AFJ (2009) Genetic variation of *Lates calcarifer* in Peninsular Malaysia based on the cytochrome b gene. Aquaculture Research 40:1742-1749.

Overturf K, Casten MT, LaPatra SL, Rexroad C III, Hardy RW (2003) Comparison of growth performance, immunological response and genetic diversity of five strains of rainbow trout (*Onorhynchus mykiss*). Aquaculture 217:93-106.

Palmer PJ (2000) Gamete storage and culture techniques for the barramundi, *Lates calcarifer* (Bloch). PhD Thesis, The University of Queensland.

Palmer PJ, Blackshaw AW, Garrett RN (1993) Successful fertility experiments with cryopreserved spermatozoa of Barramundi, *Lates calcarifer* (Bloch) using dimethylsulfoxide and glycerol as cryopreservants. Reproduction Fertility and Development 5:285-293.

Parameswaran, V, Rajesh Kumar S, Ishaq Ahmed VP, Sahul Hameed AS (2008) A fish nodavirus associated with mass mortality in hatchery-reared Asian Sea bass, *Lates calcarifer*. Aquaculture 275:366-369.

Parazo MM, Aliva EM, Reyes Jr DM (1991) Size-dependent and weight-dependent cannibalism in hatchery-bred sea bass (*Lates calcarifer* Bloch). Journal of Applied Ichthyology 7:1–7.

Reddy PVGK, Gjerde B, Tripathi SD, Jana RK, Mahapatra KD, Gupta SD, Saha JN, Sahoo M, Lenka S, Govindassamy P, Rye M, Gjedrem T (2002) Growth and survival of six stocks of rohu (*Labeo rohita*, Hamilton) in mono and polyculture production systems. Aquaculture 203:239-250.

Robertson A (1957) Optimal group size in progeny testing and family selection. Biometrics 13:442-450.

Robinson NA, Schipp G, Bosmans J, Jerry DR (2010) Modelling selective breeding in protandrous, batch-reared Asian sea bass (*Lates calcarifer*, Bloch) using walkback selection. Aquaculture Research 41:e643-e655.

Shaklee JB, Salini JP (1985) Genetic variation and population subdivision in Australian barramundi, *Lates calcarifer* (Bloch). Australian Journal of Marine and Freshwater Research 36:203-218.

Sonesson A (2005) A combination of walk-back and optimum contribution selection in fish: a simulation study. Genetics Selection Evolution 37:587-599.

Smith C (1978) The effect of inflation and form of investment on the estimated value of genetic improvement in farm livestock. Animal Production 26:101-110.

Tave D (1995) Selective breeding programmes for medium-sized fish farms. FAO Fisheries Technical Paper 352.

Wang CM, Zhu ZY, Lo LC, Feng F, Lin G, Yang WT, Li J,Yue GH (2007) A microsatellite linkage map of barramundi, *Lates calcarifer*. Genetics 175:907-915.

Wang CM, Lo LC, Zhu ZY, Lin G, Feng F, Li J, Yang WT, Tan J, Chou R, Lim HS, Orban L, Yue GH (2008) Estimating reproductive success of brooders and heritability of growth traits in Asian sea bass (*Lates calcarifer*) using microsatellites. Aquaculture Research 39:1612-1619.

Wohlfarth G, Moav R, Lahman M (1961) Genetic improvement of carp. III. Progeny tests for differences in growth rate, 1959-60. Bamidgeh 13:40-54.

Yue GH, Zhu ZY, Lo LC, Wang CM, Lin G, Feng F, Pang HY, Li J, Gong P, Liu HM, Tan J, Chou R (2009) Genetic variation and population structure of Asian seabass (*Lates calcarifer*) in the Asia-Pacific region. Aquaculture 293:22-28.

Chapter 3

Rapid assessment of genotype by environmental interactions and heritability for growth rate in aquaculture species using *in vitro* fertilisation and DNA tagging.

Macbeth G.M., and Wang Y-G. (2014) Rapid assessment of genotype by environmental interactions and heritability for growth rate in aquaculture species using *in virtro* fertilisation and DNA tagging. Aquaculture 434: 397-402.

3.1 ABSTRACT

Commercial environments may receive only a fraction of the expected genetic gains for growth rate as predicted from the selection environment. This fraction is the result of undesirable genotype-by-environment interactions (GxE) which can be measured by the genetic correlation (r_g) of growth between the different environments. Rapid estimates of genetic correlation achieved in one generation are notoriously difficult to estimate with precision. A new design is proposed where genetic correlations can be estimated by utilising artificial mating from cryopreserved semen and unfertilised eggs stripped from a single female. Traditional phenotype analysis of growth was compared to a threshold model where only the largest fish are genotyped for sire identification. The threshold model was robust to differences in family mortality of up to 30%. The design is unique as it negates potential re-ranking of families caused by an interaction between common maternal environmental effects and growing environment. The design is suitable for rapid assessment of GxE over one generation with a true 0.70 genetic correlation yielding standard errors as low as 0.07. Different design scenarios were tested for bias and accuracy with a range of heritability values, genetic correlation levels, family survival rates, number of half-sib families created, number of progeny within each full-sib family, number of fish genotyped and number of fish stocked.

3.2 Introduction

Genotype by environmental interactions (GxE) are important in many fish species such as barramundi or Asian seabass (*Lates calcarifer*). For example, the commercial environments in which barramundi are grown in after stocking as fingerlings include a combination of:

(i) ponds, cage and recirculation tanks,
(ii) fresh water, brackish and sea water growout and
(iii) tropical and sub-tropical temperatures.

These GxE interactions, measured by genetic correlations, are important to selective breeding as they cause a re-ranking of breeding values expressed in the different environments.

Significant GxE interactions for growth in aquaculture have occurred as a result of differences in temperature, salinity, stocking density and between recirculation, pond and cage facilities (Sylven et al., 1991; Myers et al., 2001; Ponzoni et al., 2005; Saillant et al., 2006; Eknath, 2007; Khaw et al., 2009; Mas-Muñoz et al., 2013) with estimates for body weight at harvest in different environments as low as 0.19±0.13 (Sae-Lim et al., 2013). There is also an emerging need for GxE assessment of growth rate in newly formulated diets with significant re-ranking of family performances between diets observed by Pierce et al. (2008), Dupont-Nivet et al. (2009) and Boucher et al. (2011).

There is a critical point at which the genetic correlation for growth is so low that more than one breeding program must be considered. This will principally be an economic decision based on the value of production in the different environments. Robertson (1959) suggested genetic correlations below 0.80 were significant whereas Ponzoni (2005) considered a genetic correlation of 0.58±0.14 for pond and cage growout in tilapia was not sufficient evidence to justify separate breeding programs. From this study the 95% confidence interval of the genetic correlation ranged from 0.31 to 0.85. In another study Domingos et al. (2013) suggested that GxE interactions were insignificant when the estimate was 0.98 even though, due to the large standard error, there was a 5% chance that the genetic correlation between fresh and saltwater growth could be as low as 0.55. If more accurate estimates of genetic correlation were available then perhaps a second breeding program could become justified. It is clear that minimising the standard error of genetic correlations is an important objective that should be achieved at the earliest stages of a breeding program and that estimates with large standard errors are clearly not informative for decision making.

Optimum designs for estimating GxE have been reported recently (Sae-Lim et al., 2010) with a model that assumed equal numbers within each family stocked prior to the evaluation. This will require separate spawning and hatching of fingerlings prior to families being pooled in different growout environments which will contribute to common family environmental effects. In some fish species the maternal common environmental effect (c^2) for growth can be as high as 0.21 Khaw et al. (2009) but are generally below 0.10 (Dupont-Nivet et al., 2010; Doupe, 2004; Gall and Huang, 1988). Even with estimates below 0.10, omitting maternal or family components can substantially inflate heritability estimates (Tosh et al., 2010; Winkelman and Peterson, 1994). The confounding of common environmental effects with additive genetic effects is a statistical problem particularly when only one generation of performance measurements are available.

A new design is proposed for rapid assessment of heritability and GxE using artificial fertilisation. Cryopreserved semen from many males is used to fertilise eggs of a single female to create many half-sib families which are identified using DNA tagging. This design:

(i) is not influenced by maternal variance as there is only one dam,

(ii) is suitable for large fish species as it negates the difficulty and cost of achieving multiple synchronised spawns,

(iii) can be suitable in species where grading of size is required to reduce cannibalism and

(iv) can be utilised when no prior pedigree and production records are available.

As reported by Macbeth and Palmer (2011) barramundi, *L. calcarifer,* is ideally suited to this design as their high fecundity in both females with up to 46 x 10^6 eggs per spawn (Davis 1984) and males with up to 10-15ml of milt (Palmer, 2000) allows many offspring to be tested for each sire. Collection of sperm has now been demonstrated in up to 30 marine fish species (Suquet et al., 2000) with egg collection (Sahoo et al., 2008) and artificial fertilisation successful in producing multiple sires (Palmer et al*.,* 1993; Quinton et al., 2007; Dupont-Nivet, 2008).

This paper examines the artificial mating design under a range of options to assess the impact on the precision and bias of heritability and genetic correlation estimates.

3.3 Methods

3.3.1 Experimental design

Due to the difficulty of stripping eggs from multiple females simultaneously and to eliminate common maternal variance an initial breeding design involving the stripping of eggs from one hatchery female (dam) was considered and using artificial mating from cryopreserved semen to establish 25 to 400 half-sib families (i.e. 25 to 400 sires). It was assumed each family was incubated separately until hatched with equal number of larvae per family pooled in a single nursery tank and grown to fingerling stage in a common environment. The simulations modelled a total of 60,000 fingerlings with subsamples ranging from 2,000 to 15,000 randomly allocated fingerlings selected to grow into each of two environments (E_1 and E_2). As a result of this subsampling procedure random variation in the number within each half-sib family is created to emulate sampling in a field trial.

Six different design configurations were simulated with parameters of each listed in Table 3.1. In these configurations the average number per family in each environment is the number stocked divided by the number of sires. For configurations 1 to 5, survival in each environment was assumed to be 100%. In configuration 6 the sensitivity of the experimental design to differing family survival rates ranging from 60% to 100% was investigated. No correlation of survival with growth was assumed. In the first scenario the survival rate of each half-sib family in both environments was the same with families 1..50 having a 100%

survival rate and families 51..100 having a survival rate chosen between 60% to 90%. The second scenario is similar to the first but with families in environment 2 swapped such that families 1..50 will have the lower survival rate and families 51..100 with the 100% survival rate. This scenario has half-sib families with low survival rates in E_1 equal to high survival rates in E_2 and vice versa emulating a survival by environment interaction.

Table 3.1. List of six configurations of simulation parameters used to estimate heritability ($\hat{h}^2$) and genetic correlation ($\hat{r}_g$) from 100 simulations. Simulation parameters include: heritability (h^2), genetic correlation between environments (r_g), number of unique sires (S), the threshold sample size taken from the heaviest fish sampled for DNA parentage analysis in each of two environments (*nDNA*) and the number of fish stocked per environment (*Stk*) taken from a pooled sample of 60,000 fingerlings. Figures enclosed by brackets indicate the range of simulated values.

Configuration		
	Simulated parameters	Presented in:
1	h^2=0.30, *Stk*=15000, *nDNA* =800, *S*=50, r_g ={0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.90}.	(Figure 3.1)
2	h^2=0.30, *Stk*=15000, *nDNA* =800, *S*={25, 50, 100, 150, 200, 400}, r_g ={0.70, 0.80}.	(Table 3.2)
3	h^2=0.30, *Stk* =15000, *S*=50, r_g =0.70, *nDNA* ={3200, 1600, 800, 400, 200}.	(Table 3.3)
4	h^2=0.30, *S*=50, r_g =0.70, *Stk*={15000, 10000, 5000, 2000}, *nDNA* ={1600, 800}.	(Table 3.4)
5	*Stk*=15,000, r_g=0.80, *S*=50, *nDNA* ={400, 800}, h^2={0.1, 0.2, 0.3, 0.4,0.5}.	(Table 3.5)
6	h^2=0.30, *Stk*=15,000, r_g =0.70, *S*=50, *nDNA* =800. Variable family survival rates.	(Table 3.6)

3.3.2 Simulation of genetic parameters

Phenotypes of progeny were generated using artificial mating formed by mixing eggs from a single female (dam *j*=1) with semen from unrelated males (*i*th sire 1..*S*) to create *S* half-sibs. The phenotypic variance was modelled as $\sigma_P^2 = \sigma_{a1}^2 + \sigma_{e1}^2 = 1$ giving the additive genetic variance in environment one $\sigma_{a1}^2 = h^2$ and the error variance in environment one $\sigma_{e1}^2 = 1 - \sigma_{a1}^2$. True breeding values $A_i = N(0, \sigma_{a1}^2)$ for the *i*th sire of each half-sib family were determined by sampling a normal distribution with mean 0 and variance σ_{a1}^2. The true breeding value of the only dam *j* was determined as $A_j = N(0, \sigma_{a1}^2)$. True breeding values for the *k*th full-sib offspring (*k*=1…*K*) within the *i*th sire and dam *j* were determined by $A_{ijk} = (A_i + A_j)/2 + M_{ijk}$ which accounts for half of the additive genetic variation each generation being derived from Mendelian sampling with $M_{ijk} = N(0, \sigma_{a1}^2 / 2)$. The phenotype of the *ijk*th progeny was determined as $P_{ijk} = A_{ijk} + N(0, \sigma_{e1}^2)$.

Sire, dam and offspring genotypes in environment two (A') were simulated using a true genetic correlation (r_g) as $A_i' = r_g A_i + N(0, \sigma_{a1}^2(1 - r_g^2))$, $A_j' = r_g A_j + N(0, \sigma_{a1}^2(1 - r_g^2))$ and $A_{ijk}' = (A_i' + A_j')/2 + M_{ijk}$ respectively giving the correlation of offspring breeding values A_{ijk}' and A_{ijk} equal to r_g. This assumed the additive genetic variance in environment two (σ_{a2}^2) was the same as that in environment one with $\sigma_{a2}^2 = \sigma_{a1}^2 = h^2$ and error variance $\sigma_{e2}^2 = \sigma_{e1}^2$ with the phenotype of fish *ijk* in environment two defined as $P_{ijk}' = A_{ijk}' + N(0, \sigma_{e2}^2)$. For modelling purposes the same *M_{ijk}* for both environments was used as each of the *ijk*th fish could only be grown in one environment.

3.3.3 Statistical analysis

The analysis was achieved using a probit model which is equivalent to the "threshold" model in animal breeding (Gianola and Foulley, 1983). The binary threshold point was determined from the largest *nDNA* fish from *H* fish harvested in each environment. The *nDNA* fish were genotyped for sire identification and assigned a threshold score of one. A threshold score of zero was assigned to all the remaining (*H- nDNA*) fish that were not genotyped. Due to random sampling the exact number of fish per sire represented in each environment was unknown and therefore an equal number of offspring per sire (*H*/*S*) was assumed in the analysis so that each sire, 1..*S*, had the same sum of 'zero' and 'one' threshold scores.

Variance-covariance parameters were estimated using the generalised linear model procedure glmmPQL in R (R Development Core Team, 2011). As expected from the simulation model preliminary results revealed no significant differences in the heritability between the two environments (*P*>0.05). It was therefore assumed the variances $\sigma_{a1}^2 = \sigma_{a2}^2$ and $\sigma_{e1}^2 = \sigma_{e2}^2$ leading to a computationally efficient univariate model: E(*B*)= Φ (*sire*/*E*+*r*) where *B* is the binary threshold score indicated by 1 for above the threshold or 0 for below the threshold, the function Φ is the cumulative standard normal distribution, *sire* is the random sire effect nested within environment (*E*) and *r* is the residual error. Note that E(*B*) is also the probability of *B*=1. The generalised linear model implemented in R had the form: glmmPQL(*B*~ 1, random= ~1|*sire*/*E*, data=yy, family=binomial(link="probit")). For comparison it was assumed all animals were weighed and DNA fingerprinted for sire identification using the statistical model *W*=*sire*/*E* + *r* where *W* is the phenotypic weight, *sire* is a random sire effect nested within environment (*E*) and *r* is the residual error. The R implementation of the model was:
glmmPQL(*W* ~ *E*, random= ~1|*sire*/*E*, data=yy, family=gaussian) where *W* is the phenotypic weight. The fixed effect of environment (*E*) in this model accounts for the phenotypic scaling caused by the interaction between the maternal genetic effect of the dam and the two environments.

The glmmPQL model partitioned three variance components:

(i) the residual variance ($\hat{\sigma}_r^2 = 1 - \hat{\sigma}_a^2 / 2$) which in the sire model includes the Mendelian variance component and is therefore different from the simulated error variance,

(ii) the additive genetic sire variance shared in common between the two environments ($\hat{\sigma}_{s\_between}^2 = \hat{r}_g \hat{\sigma}_a^2 / 4$) and

(iii) the additive genetic sire variance not shared between the two environments ($\hat{\sigma}_{s\_within}^2 = (1 - \hat{r}_g)\hat{\sigma}_a^2 / 4$)

giving the total sire variance $\hat{\sigma}_s^2 = \hat{\sigma}_{s\_within}^2 + \hat{\sigma}_{s\_between}^2 = \hat{\sigma}_a^2 / 4$. This gave heritability equal to $\hat{h}^2 = \hat{\sigma}_a^2 \big/ (\hat{\sigma}_s^2 + \hat{\sigma}_d^2 + \hat{\sigma}_r^2) = 4\hat{\sigma}_s^2 \big/ (2\hat{\sigma}_s^2 + \hat{\sigma}_r^2)$ assuming the dam variance ($\hat{\sigma}_d^2$) equals the sire variance, and the genetic correlation equal to $\hat{r}_g = \hat{\sigma}_{s\_between}^2 \big/ \hat{\sigma}_s^2$. The procedure used to estimate $\hat{r}_g$ above was computationally efficient as it negates the need to determine the numerator relationship matrix required by standard bivariate analysis.

The standard error (s.e.) of $\hat{r}_g$ and $\hat{h}^2$was determined from the standard deviation (s.d.) their estimates from 100 simulation runs. The only exception was in Figure 1 where $s.e. = s.d./\sqrt{2}$ which assumed each environment was replicated twice using a different set of sires to illustrate what could realistically be achieved in one growth period. Bias was determined from the difference between true r_g and h^2 values and the means of their estimates $\hat{r}_g$ and $\hat{h}^2$ from 100 simulation runs using a *t*-test.

3.4 Results

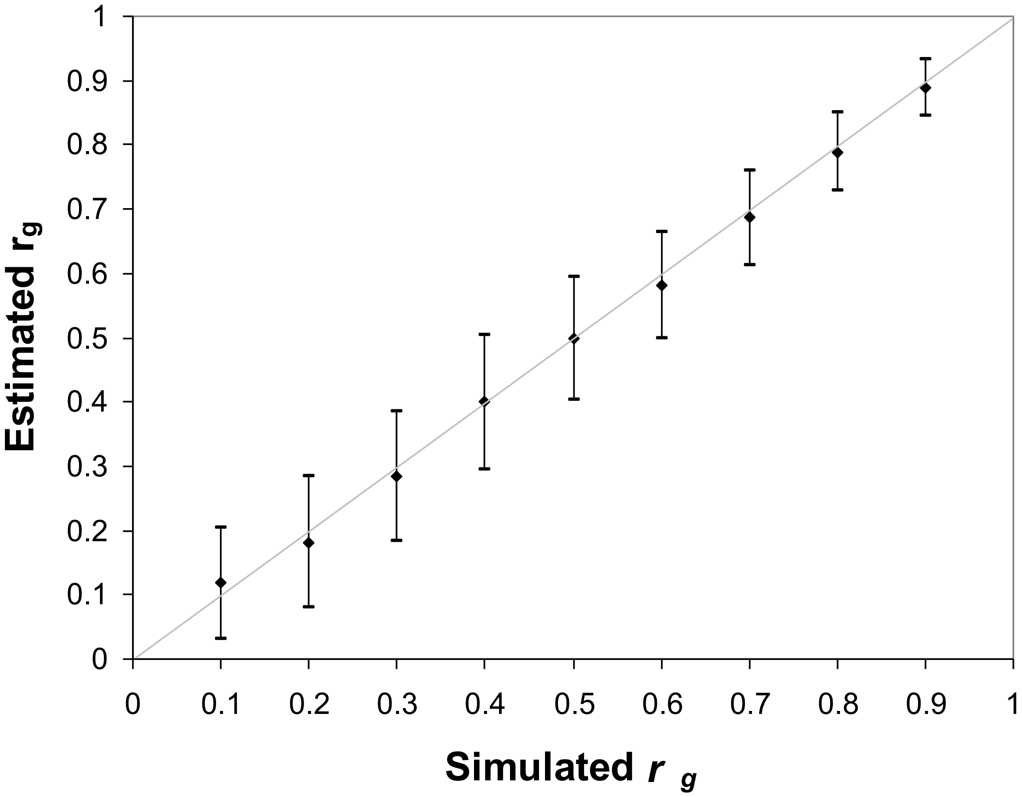
3.4.1 Variable genetic correlations with 50 sires (Configuration 1).

A list of simulation parameters for configuration 1 is provided in Table 3.1. This configuration consists of 50 different sires with the heaviest 800 genotyped for sire identification from 15000 fish stocked in each of two environments. Average estimates of genetic correlations from this configuration are shown in Figure 3.1 with standard errors assuming configuration 1 was replicated twice. The lowest standard error occurred when the true genetic correlation was high (e.g. 0.90±0.04) with standard errors generally increasing as the genetic correlation decreased. Over the nine points the genetic correlation deviated from the expected simulated value by -0.008 (standard error *s.e.*=0.004). This bias was small relative to the standard errors depicted graphically by the deviation from the grey regression line (Figure 3.1).

If the predicted genetic correlation was 0.80 then it could be determined using a 95% confidence limit that the true genetic correlation should be above 0.68 which was calculated as 0.80 less twice the standard error (Figure 3.1).

The true heritability in this design was 0.30 with mean estimates ($\bar{\hat{h}}^2$) and standard errors of 0.30±0.04, 0.30±0.05, 0.30±0.05, 0.30±0.05, 0.31±0.05, 0.30±0.05, 0.31±0.06, 0.30±0.05 and 0.31±0.05 for simulated genetic correlations of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9 respectively. In this configuration the average heritability estimates were biased higher at 0.004 (*s.e.*=0.001). In an analysis similar to configuration 1, by doubling the number of sires to 100 and doubling *nDNA* to 1600, no significant increase in the precision of genetic correlation estimates were detected (*P*>0.05).

Figure 3.1. Simulated versus estimated genetic correlation (r_g). Standard errors assume configuration 1 was replicated twice. See configuration 1 simulation parameters Table 3.1: (h^2=0.30, *Stk*=15000, *sires S*=50, *nDNA*=800).



3.4.2 Variable number of sires (Configuration 2).

See Table 3.1 for list of simulation parameters for configuration 2. The main differences from configuration 1 were that two values of r_g (0.70 and 0.80) were used and the number of sires per dam was increased at five levels from 25 to 400 (Table 3.2). The general trend was for the standard error of both $\bar{\hat{r}}_g$ and $\bar{\hat{h}}^2$ to decrease with increased number of sires. Interestingly an increase in the number of sires of 200 and above did not improve precision with $\bar{\hat{h}}^2$ being significantly underestimated. There was also a trend for $\bar{\hat{r}}_g$ to be underestimated as the number of sires increased.. This result suggests that when there are too few full-sibs per family, the family means are determined with less precision by the binomial threshold method, which in turn caused a bias in the genetic parameter estimates.

Table 3.2. Variable number of sires giving average genetic correlation ($\bar{\hat{r}}_g$) and average heritability ($\bar{\hat{h}}^2$) with standard errors (s.e.) at two genetic correlation levels (r_g). See configuration 2 simulation parameters Table 3.1: (h^2=0.30, *Stk*=15000, *nDNA*=800).

	r_g=0.70				r_g=0.80			
Sires (S)	$\bar{\hat{r}}_g$	(s.e.)	$\bar{\hat{h}}^2$	(s.e.)	$\bar{\hat{r}}_g$	(s.e.)	$\bar{\hat{h}}^2$	(s.e.)
25	0.70	(0.13)	0.29	(0.07)	0.77	(0.10)	0.30	(0.07)
50	0.68	(0.10)	0.31	(0.05)	0.78	(0.09)	0.30	(0.05)
100	0.67	(0.10)	0.31	(0.04)	0.77	(0.08)	0.31	(0.05)
150	0.66	(0.09)	0.33	(0.04)	0.74	(0.09)	0.33	(0.04)
200	0.63	(0.09)	0.33	(0.04)	0.73	(0.08)	0.33	(0.04)
400	0.56	(0.09)	0.39**	(0.04)	0.66	(0.08)	0.38**	(0.04)

** (P<0.01) different from simulated h^2.

3.4.3 Variable number DNA tagged per dam (Configuration 3).

As the number DNA tagged from the heaviest fish harvested was increased there was a general trend for a reduction in the standard error of $\bar{\hat{r}}_g$ values (Table 3.3). The large bias in $\bar{\hat{r}}_g$ when 200 were DNA tagged of 0.70-0.63=0.07 was not significantly different from the true simulated value of 0.70 due to the large standard error of these estimates from the simulation runs. It appears that at least 400 DNA samples should be taken to reduce the risk of estimation bias with little gain in precision and accuracy above 800 DNA samples.

Table 3.3. Sensitivity of threshold sample size (*nDNA*) on average genetic correlation ($\bar{\hat{r}}_g$) and average heritability ($\bar{\hat{h}}^2$) with standard errors (s.e.). See configuration 3 simulation parameters Table 3.1: (h^2=0.30, r_g=0.70, *Stk*=15000, sires *S*=50).

nDNA	$\bar{\hat{r}}_g$	(s.e.)	$\bar{\hat{h}}^2$	(s.e.)
200	0.63	(0.16)	0.32	(0.07)
400	0.69	(0.12)	0.31	(0.05)
800	0.69	(0.10)	0.31	(0.05)
1600	0.69	(0.09)	0.30	(0.05)
3200	0.69	(0.09)	0.29	(0.05)

3.4.4 Variable number stocked (Configuration 4).

In the binomial analysis a general increase in the precision and accuracy of $\bar{\hat{r}}_g$ and $\bar{\hat{h}}^2$ occurred as the number stocked in each environment increased (Table 3.4). As expected the estimated genetic parameters from the phenotypic analysis was consistently more accurate than the binomial analysis. Individual weighing and genotyping all 2000 fish gave a similar standard error of r_g=0.09 compared to grading and genotyping the heaviest 400 with a larger number of fish stocked. When comparing both methods to 2000 genotyped the binomial design could be replicated five times to produce the same number genotyped. If the binomial design was replicated five times with different sires the standard error of r_g would reduce to $0.09/\sqrt{5}=0.04$. There was no apparent advantage in stocking over 10000 fish when the binary threshold *nDNA* was 400.

Table 3.4. Sensitivity of the number of fingerlings stocked per environment (*Stk*) on estimates of average genetic correlation ($\bar{\hat{r}}_g$) and average heritability ($\bar{\hat{h}}^2$) with standard errors (s.e.). Results from a binomial analysis with threshold size *nDNA*=400 or a phenotypic analysis requiring all individual fish to be weighed and genotyped for sire identification. See configuration 4 simulation parameters Table 3.1: (h^2=0.30, r_g=0.70, *sires S* =50).

	Binomial analysis				Phenotypic analysis			
Stk	$\bar{\hat{r}}_g$	(s.e.)	$\bar{\hat{h}}^2$	(s.e.)	$\bar{\hat{r}}_g$	(s.e.)	$\bar{\hat{h}}^2$	(s.e.)
2000	0.58	(0.17)	0.34	(0.07)	0.71	(0.09)	0.30	(0.04)
5000	0.69	(0.09)	0.32	(0.05)	0.70	(0.07)	0.30	(0.03)
10000	0.68	(0.09)	0.31	(0.06)	0.70	(0.06)	0.30	(0.03)
15000	0.68	(0.09)	0.31	(0.06)	0.68	(0.06)	0.29	(0.04)

3.4.5. Variable heritability (Configuration 5).

Variable heritability was examined with 50 sires and sampling either the largest 400 or 800 for DNA genotyping of sires (Table 3.5). A general increase in accuracy was observed when estimating genetic correlations as the true heritability increased in magnitude. The converse was true for heritability with accuracy increasing as the true heritability decreased from 0.5 to 0.1. In all cases the estimated genetic correlations were equal to or slightly lower than the simulated r_g value of 0.80 but not significantly different from that value (*P*>0.05). These results confirm that the binomial threshold model is a robust method of estimating genetic correlations within the realistic range of heritability from 0.1 to 0.5 for growth rate.

Table 3.5. Sensitivity of heritability on estimates of average genetic correlation ($\bar{\hat{r}}_g$) and average heritability estimates ($\bar{\hat{h}}^2$) with standard errors (s.e.) at two threshold sizes (*nDNA*). See configuration 5 simulation parameters Table 3.1: (r_g=0.80, *Stk*=15000, *sires S*=50).

	nDNA=400				*nDNA*=800			
Heritability	$\bar{\hat{r}}_g$	(s.e.)	$\bar{\hat{h}}^2$	(s.e.)	$\bar{\hat{r}}_g$	(s.e.)	$\bar{\hat{h}}^2$	(s.e.)
0.1	0.77	(0.18)	0.10	(0.02)	0.77	(0.13)	0.10	(0.02)
0.2	0.78	(0.13)	0.20	(0.05)	0.78	(0.10)	0.20	(0.04)
0.3	0.79	(0.11)	0.30	(0.06)	0.79	(0.07)	0.31	(0.05)
0.4	0.80	(0.08)	0.41	(0.07)	0.80	(0.07)	0.40	(0.07)
0.5	0.77	(0.10)	0.51	(0.09)	0.78	(0.07)	0.50	(0.08)

3.4.6 Variable family survival (Configuration 6).

In this configuration the sensitivity of the binomial threshold model to differences in family survival was investigated. The experimental design was sufficiently robust to estimate genetic correlations with variable survival rates of 60% and 100% provided the survival rates of each half-sib family were the same across both environments (Table 3.6). The design was sensitive to interaction survival rates where families having good survival rates in one environment had poor survival rates in the other environment and vice versa. In this worst case interaction survival rates of 60% and 100% had significantly lower genetic correlation estimates than the simulated value of 0.70 (Table 3.6).

Table 3.6. Sensitivity of survival on average genetic correlation ($\bar{\hat{r}}_g$) and average heritability ($\bar{\hat{h}}^2$) with standard errors (s.e.). The percentage family survival across 100 full-sib families indicated by sires 1..50 and sires 51..100 (see Table 3.1 configuration 6: h^2=0.30, r_g=0.70, *Stk*=15000, *sires S*=100, *nDNA*=800).

Environment 1		Environment 2					
Sires		Sires		$\bar{\hat{r}}_g$ (s.e.)		$\bar{\hat{h}}^2$ (s.e.)	
1..50	51..100	1..50	51..100				
Control							
100	100	100	100	0.70	0.09	0.32	0.05
Same family survival in each environment (no interaction)							
100	90	100	90	0.69	0.10	0.32	0.04
100	80	100	80	0.68	0.10	0.32	0.04
100	70	100	70	0.70	0.09	0.34	0.04
100	60	100	60	0.72	0.09	0.36	0.04
Different family survival in each environment (with interaction)							
100	90	90	100	0.67	0.09	0.32	0.05
100	80	80	100	0.61	0.10	0.33	0.04
100	70	70	100	0.52	0.11	0.33	0.04
100	60	60	100	0.42**	0.10	0.37	0.05

** (P<0.01) different from simulated r_g.

3.5. Discussion

The results in this study demonstrate that it is feasible to estimate genetic correlations using a binomial threshold model. In this design there may be families not represented in the top threshold of fish sampled and therefore it is important to understand the limitations and also advantages of this design as reflected by the precision and bias of estimating both genetic correlations and heritability.

One advantage of this design is the absence of maternal common variance which is expressed as a proportion of total phenotypic variance and denoted as c^2 (Montaldo et al., 2012). The existence of c^2 and potential re-ranking of families due to an interaction between c^2 and environment is ignored in studies estimating genetic correlations (Fishback et al., 2002; Ponzoni, 2005; Sae-Lim et al., 2010). This non-genetic re-ranking of families in different environments will ultimately cause an underestimate of GxE. As there is no c^2 effect in the design presented in this study the genetic correlations are potentially less biased than other proposed designs.

The drawback with no c^2 effect is that the heritability estimates are potentially inflated by ($1+c^2$) if applied to selection environments where c^2 is present. However in other experimental designs where c^2 is present genetic variances tend to be overestimated when c^2 has been unaccounted in statistical models (Tosh et al., 2010). Dupont-Nivet et al. (2009) went to the extent of creating maternal clones for genetic variance component assessment to minimise maternal effects. Other strategies include reducing the age span of experimental fish (Pierce et al., 2008) or by minimising the size range of eggs between spawns (Dupont-Nivet et al., 2009). The potential source of bias caused by differences in maternal effects is simply eliminated in this design because c^2 is equal to zero as only one dam is used. One potential concern using a design with a single dam is that genotype by genotype interactions could inflate the sire variance. Generally genotype by genotype interactions are of little concern but could be alleviated in species where eggs from multiple females can be collected simultaneously. In this case equal weights of eggs from each female may be mixed prior to fertilization in order to average the effect of c^2 from multiple dams for each sire.

Another source of experimental variation is the common tank effect, which is equivalent to a paternal effect in this design. This tank effect was minimised using *in vitro* fertilisation allowing fertilization of all half-sib families to occur within minutes of each other and by mixing families soon after hatching (Macbeth and Palmer, 2011). This protocol minimises the potential for any paternal tank effects. If in practice, fingerlings cannot be reared in the same tank then significant tank effects could be included in the statistical model. Small tank effects are unlikely to affect h^2 and r_g estimates as families can be mixed immediately after hatching prior to rearing in different tanks. The estimates are also reasonably robust to environmental changes affecting family representation in the heaviest group of fish genotyped as inferred by the survival sensitivity analysis.

There are many similarities between the results of Sae-Lim et al. (2010) and the present study. For example,

(i) a reduction in the standard error of r_g occurs as h^2 increases,

(ii) a interaction of family survival and environment can cause a bias and a reduction in the accuracy of r_g and

(iii) there is an optimised family size that minimises the standard error of r_g.

The design can give similar standard errors of r_g to that reported by Sae-Lim et al. (2010). When only 2000 are stocked it is recommend all animals are genotyped and weighed with an analysis performed on phenotypic data. The binomial analysis, with larger number of fish stocked, is expected to perform well in aquaculture species where commercial grading on size is required during the experimental trial period, as practiced in species such as *L. calcarifer* (Macbeth and Palmer, 2011).

A uniform family survival is desirable in the probit analysis as all families were assumed to be represented in equal stocking proportions. As there is no dam variance expressed in this design and with all eggs exactly the same age, and sampled from the same size distribution, it is unlikely that there will be a large variation in family survival. Additionally due to one dam used in the design only three quarters of the total additive genetic variation is available to contribute to survival variation between the half-sib families produced. It is also less likely that significant family survival by environment interaction will occur due to both the reduction in total genetic variance and elimination of c^2 variance components. Overall the analysis appeared robust to differing family mortality rates up to 30% although an interaction between family survival and environment seemed to cause more bias in r_g and h^2 estimates.

Designs using synchronous spawning are difficult to achieve in practice (Dupont-Nivet et al., 2008; Quinton et al., 2007; Boucher et al., 2011) making it necessary to pool spawns over a period of more than one day (Pierce et al., 2008). Mixing fish born more than one day apart is undesirable as older fish may maintain a size advantage they never relinquish (Tave, 1995) causing biased estimates of genetic parameters. Also if batch spawning is used the highly variable fertilisation rates among males and females in mating tanks (Frost et al., 2006; Nissling et al., 2002) may cause many half-sib families to be over-represented with the downstream risk of increasing the standard errors of genetic parameter estimates. With controlled artificial mating the design is less variable and is therefore achieved with less risk of experimental failure.

In many of the simulated designs with r_g=0.70 a standard error of 0.09 could be achieved which would reduce to $0.09/\sqrt{2}=0.06$ if two replicates of each environment were evaluated with a different set of sires. This standard error is better than published estimates of r_g reported within the range of 0.5 to 0.8 including: 0.51+0.19 for *Dicentrarchus labrax* (Salliant et al., 2006), 0.56±0.34 for *Solea solea* (Mas-Muñoz et al., 2013), 0.58±0.14 for *Oreochromis niloticus* (Ponzoni et al., 2005), 0.70±0.10 for *D. labrax* (Dupont-Nivet et al., 2008), *0.73±0.13 for Oncorhynchus mykiss* (Pierce et al., 2008), 0.75±0.09 for *Oncorhynchus mykiss* (Dupont-Nivet et al., 2010), 0.82±0.21 for *Coregonus lavaretus* L.

(Quinton et al., 2007), 0.67±0.12 *Oncorhynchus mykiss* Walbaum (Boucher et al., 2011) and 0.74±0.21 to 0.84±0.15 for *Oreochromis niloticus* (Khaw et al., 2009).

The experimental design used here to estimate genetic correlations is also suitable for a novel design which can achieve rapid genetic gains during the establishment phase of a breeding program (Macbeth and Palmer, 2011). In existing breeding programs the design could be used to compare the ranking of nucleus breeding sires to other sires outside the breeding program as part of a screening program to increase genetic gains and or reduce inbreeding.

3.6. Conclusion

Both threshold and phenotypic weight analysis are a viable option to estimate the heritability and genetic correlations for growth rate using multiple half-sib families created from artificial fertilisation of eggs from a single female. The threshold model was robust to differences in family mortality of up to 30%. The advantages include:

(i) the elimination of bias from family re-ranking caused by c^2 by environmental interaction,

(ii) the standard errors from this design are on average better than published estimates using natural mating designs,

(iii) rapid assessments of genetic parameters over one growing period can be obtained,

(iv) tighter mating control reducing the risk of unequal representation of families,

(v) families can be mixed and reared immediately after hatching to reduce fixed tank effects and

(vi) threshold analysis suitable for species that are graded during commercial operations.

3.7 Acknowledgements

Funding was supported from the Department of Agriculture Fisheries and Forestry, Queensland Government, Australia.

3.8 References

Boucher RL, Quillet E, Vandeputte M, Lecalvez JM, Goardon L, Chatain B, Medale F, Dupont-Nivet M (2011) Plant-based diet in rainbow trout (*Oncorhynchus mykiss* Walbaum): Are there genotype-diet interactions for main production traits when fish are fed marine *vs*. plant-based diets from the first meal? Aquaculture 321:41-48.

Davis TLO (1984) Estimation of fecundity in barramundi, *Lates calcarifer* (Bloch), using an automatic particle counter. Australian Journal of Marine Freshwater Research 35:111-118.

Domingos JA, Smith-Keune C, Robinson N, Loughnan S, Harrison P, Jerry DR (2013) Heritability of harvest growth traits and genotype-environment interactions in barramundi, *Lates calcarifer* (Bloch). Aquaculture 402-403:66-75.

Doupe RG (2004) Selection for faster growing back bream *Acanthopagrus butcheri*. PhD thesis, Murdoch University.

Dupont-Nivet M, Karahan-Nomm B, Vergnet A, Merdy O, Haffray P, Chavanne H, Chatain B, Vandeputte M (2010) Genotype by environment for growth in European sea-bass (*Dicentrarchus labrax* L.) are large when growth rate rather than weight is considered. Aquaculture 306:365-368.

Dupont-Nivet M, Medale F, Leonard J, Le Guillou S, Tiquet F, Quillet, E, Geurden I (2009) Evidence of genotype-diet interactions in the response of rainbow trout (*Oncorhynchus mykiss*) clones to a diet with or without fishmeal at early growth. Aquaculture 295:15-21.

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

Eknath AE, Bentsen HB, Ponzoni RW, Rye M, Nguyen NH, Thodesen J, Gjerde B (2007) Genetic improvement of farmed tilapias: composition and genetic parameters of a synthetic base population of *Oreochromis niloticus* for selective breeding. Aquaculture 273:1–14.

Fishback AG, Danzmann RG, Ferguson MM, Gibson JP (2002) Estimates of genetic parameters and genotype by environment interactions for growth traits of rainbow trout (*Oncorhynchus mykiss*) as inferred using molecular pedigrees. Aquaculture 206:137-150.

Frost LA, Evans BS, Jerry DR (2006) Loss of genetic diversity due to hatchery culture practices in barramundi (*Lates calcarifer*). Aquaculture 261:1056-1064.

Gall GAE, Huang N (1988) Heritability and selection schemes for rainbow trout: female reproductive performance. Aquaculture 73:57-66.

Gianola D, Foulley JL (1983) Sire evaluation for ordered categorical data with a threshold model. Genetic Selection Evolution 15:201-224.

Khaw HL, Bovenhuis H, Ponzoni RW, Rezk MA, Charo-Karisa H, Komen H (2009) Genetic analysis of Nile tilapia (*Oreochromis niloticus*) selection line reared in two input environments. Aquaculture 294:37-42.

Macbeth GM, Palmer PJ (2011) A novel breeding program for improved growth in barramundi *Lates calcarifer* (Bloch) using foundation stock from progeny-tested parents. Aquaculture 318:325-334.

Mas-Muñoz J, Blonk R, Schrama JW, van Arendonk J, Komen H (2013) Genotype by environment interaction for growth of sole (*Solea solea*) reared in an intensive aquaculture system and in a semi-natural environment. Aquaculture 411:230-235.

Montaldo H, Castillo-Juárez H, Campos-Montes G, Pérez-Enciso M (2012) Effect of the data family structure, tank replication and the statistical model, on the estimation of genetic parameters for body weight at 28 days of age in the Pacific white shrimp (*Penaeus (Litopenaeus) vannamei* Boone, 1931). Aquaculture Research 44:1715-1723.

Myers, JM, Heggelund PO, Hudson G, Iwamoto RN, (2001) Genetics and broodstock management of coho salmon. Aquaculture 197:43-62.

Nissling A, Westin L, Hjerne O (2002) Reproduction success in relation to salinity for three flatfish species, dab (*Limanda limanda*), plaice (*Pleuronectes platessa*), and flounder (*Pleuronectes flesus*), in the brackish water Baltic Sea. Journal of Marine Science 59:93-108.

Palmer PJ (2000) Gamete storage and culture techniques for the barramundi, *Lates calcarifer* (Bloch). PhD Thesis, The University of Queensland.

Palmer PJ, Blackshaw AW, Garrett RN (1993) Successful fertility experiments with cryopreserved spermatozoa of Barramundi, *Lates calcarifer* (Bloch) using dimethylsulfoxide and glycerol as cryopreservants. Reproduction Fertility and Development 5:285-293.

Pierce L, Palti Y, Silverstein J, Barrows F, Hallerman E, Parsons J (2008) Family growth response to fishmeal and plant-based diets shows genotype × diet interaction in rainbow trout (*Oncorhynchus mykiss*). Aquaculture 278:37-42.

Ponzoni RW, Hamzah A, Kamaruzzaman N, Hooi LK (2005) Live weight genetic parameters in two production environments in the GIFT strain of Nile tilapia (*Oreochromis* *niloticus*). Association for the Advancement of Animal Breeding and Genetics 16:202-205.

Quinton CD, Kause K, Ruohonen K, Koskela J (2007) Genetic relationships of body composition and feed utilisation traits in European whitefish (*Coregonus lavaretus* L.) and implications for selective breeding in fishmeal- and soybean meal-based diet environments. Journal of Animal Science 85:3198-3208.

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Robertson A (1959) The sampling variance of the genetic correlation coefficient. Biometrics 15:469-485.

Sae-Lim P, Kause A, Mulder HA, Martin KE, Barfoot AJ, Parsons JE, Davidson J, Rexroad III CE, van Arendonk JAM, Komen H (2013) Genotype-by-environment interaction of growth traits in rainbow trout (*Oncorhynchus mykiss*): A continental scale study. Journal of Animal Science 91:5572-5581.

Sae-Lim P, Komen H, Kause A (2010) Bias and precision of estimates of genotype-by-environment interaction: A simulation study. Aquaculture 310:66-73.

Sahoo SK, Giri SS, Chandra S, Mohapatra BC (2008) Evaluation of breeding performance of Asian catfish *Clarias batrachus* at different dose of HCG and latency period combinations. Turkish Journal of Fisheries and Aquatic Sciences 8:249-251.

Saillant E, Dupont-Nivet M, Haffray P, Chatain B, (2006) Estimates of heritability and genotype-environment interactions for body weight in sea bass (*Dicentrarchus labrax* L.) raised under communal rearing conditions. Aquaculture 254:139-147.

Suquet M, Dreanno C, Fauvel C, Cosson J, Billard R (2000) Cryopreservation in marine fish. Aquaculture Research 31:231-243.

Sylven S, Rye M, Simianer H (1991) Interaction of genotype with production system for slaughter weight in rainbow trout (*Oncorhynchus mykiss*), Livestock Production Science 28:253-263.

Tave D (1995) Selective breeding programmes for medium-sized fish farms, FAO Fisheries Technical Paper 352.

Tosh JJ, Garber AF, Trippel EA, Robinson JAB (2010) Genetic, maternal and environmental variance components for body weight and length of Atlantic cod at 2 points in life. Journal of Animal Science 88:3513-3521.

Winkelman AM, Peterson RG (1994) Heritabilities, dominance variation, common environmental effects and genotype by environment interactions for weight and length in Chinook salmon. Aquaculture 125:17-30.

Chapter 4

Likelihood-based genetic mark–recapture estimates when genotype samples are incomplete and contain typing errors.

Macbeth G.M., Broderick D, Ovenden J.R., Buckworth R.C. (2011) Likelihood-based genetic mark–recapture estimates when genotype samples are incomplete and contain typing errors. Theoretical Population Biology 80:185-196.

4.1 ABSTRACT

Genotypes produced from samples collected non-invasively in harsh field conditions often lack the full complement of data from the selected microsatellite loci. The application to genetic mark-recapture methodology in wildlife species can therefore be prone to misidentifications leading to both ‘true non-recaptures’ being falsely accepted as recaptures (Type I errors) and ‘true recaptures’ being undetected (Type II errors). A new likelihood method is presented that allows every pairwise genotype comparison to be evaluated independently. This method was applied to determine the total number of recaptures by estimating and optimising the balance between Type I errors and Type II errors. It was demonstrated through simulation that the standard error of recapture estimates can be minimised through new algorithms. Interestingly, the precision of recapture estimates actually improved when individuals with missing genotypes were included, as this increased the number of pairwise comparisons potentially uncovering more recaptures. Simulations suggest the method is tolerant of per locus error rates of up to 5% and can theoretically work in data sets with as little as 60% of loci genotyped. The methods can be implemented in data sets where standard mismatch analyses fail to distinguish recaptures. Finally, it was also demonstrated that by assigning a low Type I error rate to the matching algorithms a dataset of individuals of known capture histories could be generated that is suitable for downstream analysis with traditional mark recapture methods.

4.2. Introduction

The objective of DNA-based mark-recapture studies is to accurately determine if two samples are either from the same individual or different individuals (Paetkau and Strobeck, 1994). Even well designed studies that theoretically achieve objectives may in practice have genotype profiles that are not always correct or complete (Taberlet et al., 1996; Gagneux et al., 1997) which result in biased recapture estimates. While it is laudable to avoid such problems in the first place, the reality is that they exist in all datasets to varying degrees.

An *a posteriori* approach to handle missing loci and genotype errors based on likelihood methods of Kalinowski et al. (2007) is evaluated. When assessing the hypothesis that a recapture exists between two genotype samples there are two sources of misidentification. The first source of misidentification is a Type I error where two samples from different individuals are incorrectly identified as a recapture. This is a known problem in genetic mark-recapture studies and when two individuals have identical genotypes they are called ‘shadows’ (Mills et al., 2000). In the case when genotyping errors are present in the data, Type I errors can also occur. Type II errors are manifested by genotyping errors when two observed genotypes from the same individual are concluded to be from two individuals.

The presence of Type I and Type II errors are an important challenge to genetic mark-recapture studies and are more likely in datasets that comprise of individuals with incomplete or ‘partial’ genotypes. This lack of genotype information is further exacerbated when comparisons are attempted to be made among individuals with partial genotypes that have few loci or even no loci in common. In studies where partial genotypes are rare, individuals or loci can be culled (Paetkau, 2003) until pairwise comparisons can be made at enough loci for individual identification (Rudnick et al., 2008). However this approach may not be workable if few individuals are fully genotyped (e.g. Chu et al., 2006; Chaline et al., 2004). When partial genotypes are common, the removal of too many individuals reduces the chances of detecting recaptures and inflates the variance in population size estimates, whereas the removal of too many loci increases the prevalence of miss-identification.

Type I errors cause underestimates of population size from mark-recapture studies if too many individuals are falsely assessed as recaptures (Waits and Leberg, 2000). The prevalence of Type I errors in a mark-recapture study is determined by:

(i) the allelic diversity of the microsatellite loci used for genotyping,

(ii) the relatedness of individuals in the population as Type I errors are also more likely to occur between pairs of siblings or other first order relatives (Evett and Weir, 1998) and

(iii) sample size of the study. While a genetic mark-recapture study may be designed in the first instance to minimise miss-identifications, unforeseen increases in sample size (e.g. augmentation of historical data) can scuttle best laid plans. Progressively adding more samples will exponentially increase the number of pairwise comparisons and will increase the incidence of Type I errors.

Type II errors cause inflated estimates of population size from mark-recapture studies when genotyping errors among recaptures are inadvertently identified as new individuals (Creel et al., 2003). Common genotyping errors include allelic drop out, scoring of artefact peaks, misinterpretation of allele banding patterns and those from DNA contamination (Fernando et al., 2003; Hoffman and Amos, 2005; Wright et al., 2009). Some loci are more susceptible to genotyping errors than others (Hoffman and Amos, 2005) and poor DNA quality is more error prone (Creel et al., 2003; Taberlet ,1996). It should still be born in mind that even high quality DNA can have surprisingly high error rates with allelic dropouts of 21-57% (Soulsbury et al., 2007). Repetitive PCR amplification of error-prone loci

(Taberlet et al., 1996) will go a long way to detecting errors but it will not eliminate them completely. Close scrutineering of the genotyped data is necessary and suspect genotypes can be detected using mismatch techniques as potential recaptures match at all or almost all loci (Paetkau, 2003; McKelvey and Schwartz, 2005; Kalinowski et al*.,* 2006). However these methods are generally not suitable in data where missing loci are common.

To obtain reliable biological information from mark-recapture studies using genotype data, the prevalence of both Type I and Type II errors should be quantified and their effect on downstream analyses minimised. For example, Creel et al. (2003) described a threshold based on probability of identity (*PID*) which effectively accepts a given level of Type I errors for population size estimation, but as the authors note, the difficulty with this approach is in determining the appropriate *PID* threshold. Knapp et al. (2009) used *PID* to estimate population size while accounting for genotyping errors and Type I errors, but it is not clear how robust these methods are when applied to data with missing loci as each non-missing locus combination will have a different Type I error distribution based on *PID*. Wright et al. (2009) explicitly modelled allelic dropout errors from repeated genotyping to adjust estimates of population parameters and while individuals with missing loci were evaluated the problem of Type I errors was not addressed. An alternative approach using likelihood ratios based on the methods of Kalinowski et al. (2007) offer the most promise for addressing the issue of misidentifications when a proportion of genotypes in a dataset have missing data.

The objective in this paper was to evaluate the robustness of a new approach that uses likelihood ratios to maximize the accuracy of estimating the number of recaptures in the presence of partial genotypes and genotype errors. The new method achieves this by simultaneously accounting for Type I and Type II errors in a stochastic optimisation procedure which minimises the estimated variance of recapture estimates. The mathematical theory of the approach is described and tested through stochastic simulation by examining the effect of partial genotypes, genotype errors and sample size in both bi-allelic and multi-allelic genotypes. Also investigated was how the methods could be applied in related populations such as those with full-sibs.

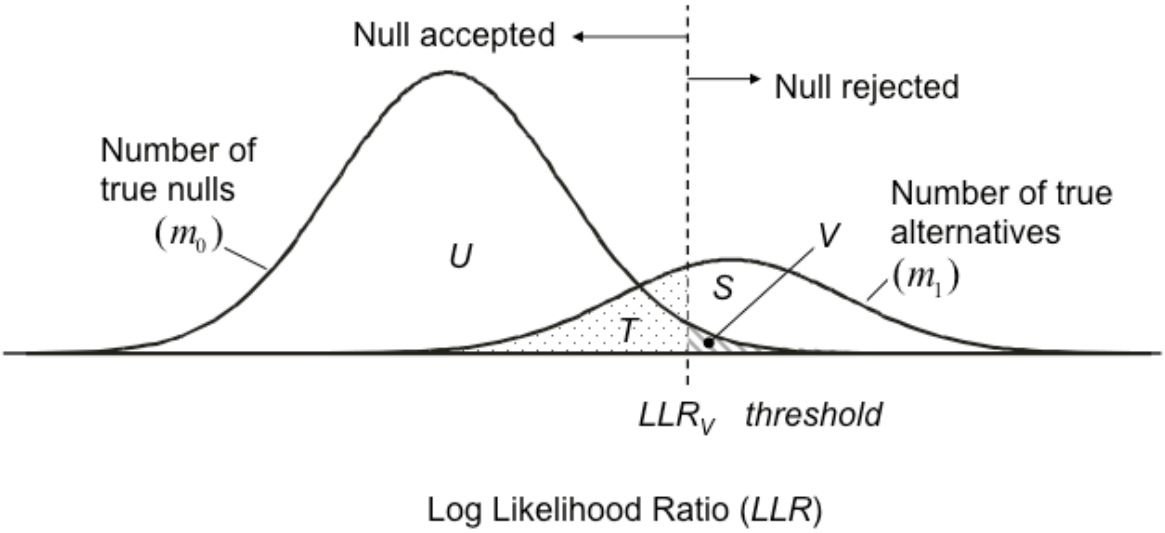
4.3. Materials and methods

4.3.1 Overview of mathematical theory:

The model is based on the classical multiple hypothesis testing model (Figure 4.1).

Figure. 4.1. True null and true alternative multiple hypothesis distributions. To the right of the Log Likelihood Ratio threshold (*LLR*$_V$) are *S* the number of true positives (recaptures) and *V* the number of false positives (Type I errors) and to the left of *LLR*$_V$ are *U* the number of true negatives and *T* the number of false negatives (Type II errors). The total number of true nulls $m_0 = U + V$ and the total number of true alternatives $m_1 = S + T$ with the Type I error rate $\alpha = V / m_0$ and the Type II error rate $\beta = T / m_1$.

> Note: The height of the null curve divided by m_0 forms a density distribution with the total area under the curve equal to unity. In this case the area to the right of *LLR*$_V$ equals α. Similarly the height of the alternative curve divided by m_1 forms a density distribution with the area to the left of *LLR*$_V$ equal to β as shown in Figure 1.3.

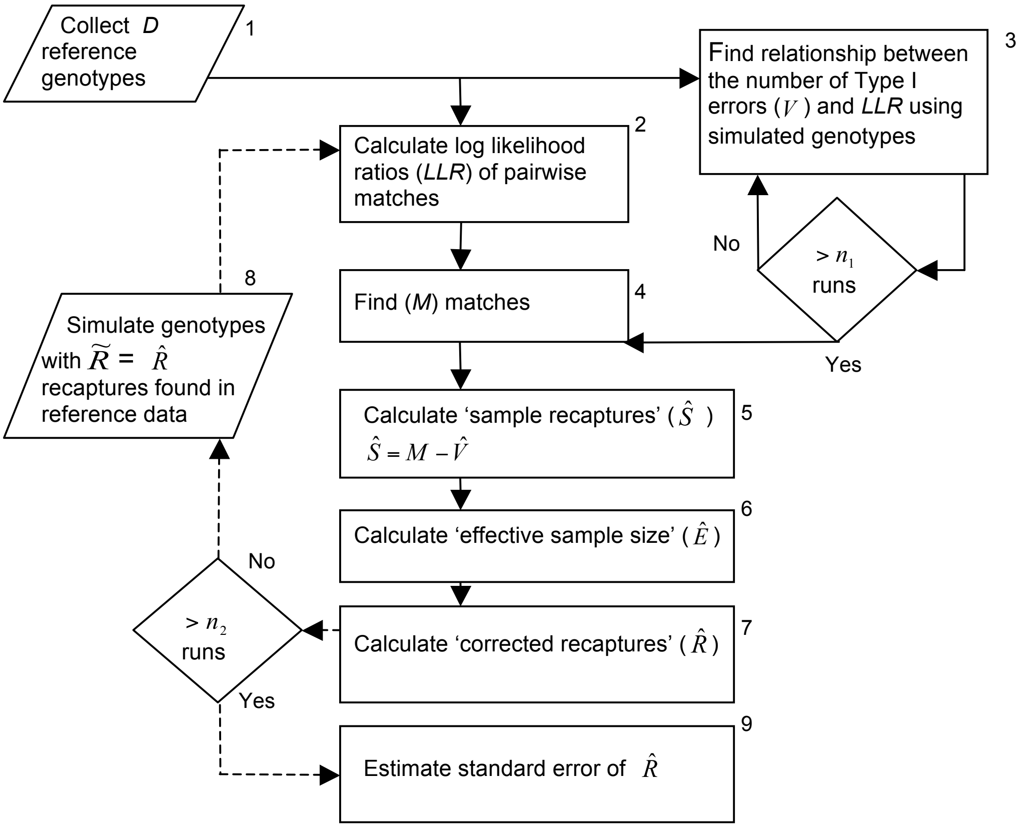


The Log Likelihood Ratio (*LLR*) distribution of all the true nulls (m_0) are from individuals randomly selected from the population. The number of pairwise matches found to the right of the Log Likelihood Ratio threshold (*LLR*$_V$) in Figure 4.1 is defined as M. For a given value of *LLR*$_V$, *M* is determined directly from *D* samples in the reference data with the total estimated number of Type I errors ($\hat{V}$) determined by simulation. The estimated number of ‘sample recaptures’ found to the right of LLR_V is then estimated as: $\hat{S} = M - \hat{V}$. The estimate of the total number of true alternatives (m_1) is the number of ‘corrected recaptures’ ($\hat{R}$) calculated from *D*, $\hat{S}$ and a new estimated term $\hat{E}$ called the ‘effective sample size’.

4.3.2 Processes detailing mathematical theory:

Using the nomenclature (Section 4.10, Appendix I) the general strategy for estimating $\hat{R}$ is portrayed in ten different processes of which processes one to nine are shown in Figure 4.2.

Figure. 4.**2.** Flow chart showing the process of estimating corrected matches ($\hat{R}$) when accepting *V* Type I errors. Each process is numbered and referenced in the text. Dashed lines represent the additional steps to estimate standard error ($\hat{R}$) for a single genotype file.


Collect *D* reference genotypes
1
Find relationship between the number of Type I errors (V) and *LLR* using simulated genotypes
3
Calculate log likelihood ratios (*LLR*) of pairwise matches
2
No
$> n_1$ runs
Yes
8
Simulate genotypes with $\widetilde{R} = \hat{R}$ recaptures found in reference data
Find (*M*) matches
4
Calculate ‘sample recaptures’ ($\hat{S}$)
$\hat{S} = M - \hat{V}$
5
Calculate ‘effective sample size’ ($\hat{E}$)
6
No
$> n_2$ runs
Yes
Calculate ‘corrected recaptures’ ($\hat{R}$)
7
Estimate standard error of $\hat{R}$
9

After collecting the genotype data the relationship between log likelihood ratios (*LLR*) and the total number of Type I errors (*V*) is determined so that the number of genotype matches (*M*) can be calculated at given fixed levels of *V* (*process* 1-4). Sample recaptures ($\hat{S}$) were then estimated (*process* 5). Using an estimate of effective sample size ($\hat{E}$) corrected recaptures ($\hat{R}$) were then estimated (*process* 6-7). Simulations were then run to determine the standard error around the recapture estimate (*process* 8-9). The final *process* 10 describes how a converged estimate of $\hat{R}$ is determined from a range of *V* priors. A detailed description of all the processes is described below.

Process 1: *Collect D reference genotypes*

The reference data is a collection of multi-locus diploid genotypes, each representing a single sample. The genotypes can be collected from field tissue samples or generated through simulation (see below: *Testing theory through stochastic simulation*).

Process 2: *Calculate log likelihood ratios (LLR) to rank genotype matches*

Likelihood ratios to identify genotype matches between pairs of genotype samples *a* and *b* are developed below using similar methodology to that used by Marshall et al. (1998) and Kalinowski et al. (2007). A single locus can be evaluated for the likelihood of the null hypothesis (H_0), that an alleged match is not a true match such that both samples are from individuals randomly selected from the population. This is then tested against the alternate hypothesis (H_1) that an alleged match is a true match such that a sample with genotype g_a is from the same individual as the sample with genotype g_b using:

$P(g_a,g_b|H_0) = P(g_a)P(g_b) = L(H_0|g_a,g_b)$ = likelihood of H_0 given the observed genotypes,
$P(g_a,g_b|H_1) = I(g_a|g_b)P(g_b) = L(H_1|g_a,g_b)$ = likelihood of H_1 given the observed genotypes,
where $I(g_a|g_b)$= 1 if $g_a = g_b$ and 0 otherwise.

The likelihood of a true match divided by the likelihood of a random selection is

$$L(H_1|g_a,g_b) / L(H_0|g_a,g_b) = I(g_a|g_b) / \mathrm{P}(g_a) \qquad (4.1)$$

The genotype probabilities, and hence the likelihood ratios, can be calculated from the population allele frequencies which may be known or estimated from the reference data (Table 4.1). With no genotyping errors, the likelihood ratio is zero if any allele is different between two genotypes, making the log likelihood ratio undefined.

Table 4.1 Likelihood ratios of match pairs. *X* represents any allele that is not *B*, *Y* represents any allele that is not *C* and *Z* represents any allele that is neither *B* nor *C*. The frequency of alleles *B* and *C* is denoted *p* and *q* respectively, where *p*+*q* is less than one when there are more than two alleles.

Reference genotype (g_a)	Alleged match genotype (g_b)	$I(g_a\|g_b)$	$P(g_a)$	$L(H_1\|g_a,g_b)$ / $L(H_0\|g_a,g_b)$
BB	*BB*	1	p^2	$1/(\ p^2)$
BC	*BC*	1	$2pq$	$1/(2pq)$
BB	*XX* or *BX*	0	p^2	0
BC	*BY* or *CX* or *ZZ*	0	$2pq$	0

Building on equation 4.1, the random genotype replacement model of Marshall et al. (1998) and Kalinowski et al. (2007) was modified to model genotype error rates for genotype matches. At any given locus the probability of observing genotype *g* is equal to (1-ε)*P*(*g*) + ε *P*(*g*) where ε is the genotype error rate per locus. The first term is the probability that the locus has genotype *g* and is not in error, while the second term is the probability that the locus has genotype *g* and is in error. Table 4.2 lists the components of the likelihood equations when none, one or two genotype errors exist between a pair of genotypes being compared. The likelihood ratio (*LR*) for each locus simplifies to:

$$LR = \frac{(1-\varepsilon)^2[I(g_a \mid g_b)P(g_b)] + 2\varepsilon(1-\varepsilon)[P(g_a)P(g_b)] + \varepsilon^2[P(g_a)P(g_b)]}{(1-\varepsilon)^2[P(g_a)P(g_b)] + 2\varepsilon(1-\varepsilon)[P(g_a)P(g_b)] + \varepsilon^2[P(g_a)P(g_b)]}$$
$$= (1-\varepsilon)^2 I(g_a \mid g_b)/P(g_a) + \varepsilon(2-\varepsilon) \qquad (4.2)$$

When there is a genotype error rate per locus with ε greater than zero, *LR* estimates are determined from equation 4.2. Log likelihood ratios (*LLR*) for multiple loci assume independence between loci and are calculated from the natural logarithm of the product of those *LR* estimates using equation 4.3 where L^* is the number of locus pairs present with alleles in both the g_a and g_b genotypes. Note that L^* may be less than the number of loci in the genotyping panel (*L*) as missing loci may be present within each genotype forming partial genotypes. All samples in the reference data are compared with each other to give many *LLR* values. The pairwise matches between genotypes with the highest *LLR* values are those most likely to be a true match pertaining to a recapture in wildlife studies.

$$LLR = \ln \prod_{i=1}^{L^*} LR_i \tag{4.3}$$

Table 4.2 The likelihood equations when the error rate per loci is ε for: *L*(*H₁*|*g_a*,*g_b*) the hypothesis that a true match exists between the sample *a* and sample *b*, and for *L*(*H₀*|*g_a*,*g_b*) the alternative hypothesis that a match occurs by chance, with terms shown in columns two and three respectively. Genotype probabilities *P*(*g*) resulting in errors are underlined with the number of genotype errors within each additive term indicated in column one.

Genotype errors	$L(H_1 \vert g_a, g_b)$	$L(H_0 \vert g_a, g_b)$
0	$= (1-\varepsilon)^2 I(g_a \vert g_b) P(g_b)$	$= (1-\varepsilon)^2 P(g_a) P(g_b)$
1	$+ \varepsilon(1-\varepsilon) P(g_a) \underline{P(g_b)}$	$+ \varepsilon(1-\varepsilon) P(g_a) \underline{P(g_b)}$
1	$+ \varepsilon(1-\varepsilon) \underline{P(g_a)} P(g_b)$	$+ \varepsilon(1-\varepsilon) \underline{P(g_a)} P(g_b)$
2	$+ \varepsilon^2 \underline{P(g_a)}\,\underline{P(g_b)}$	$+ \varepsilon^2 \underline{P(g_a)}\,\underline{P(g_b)}$

Process 3: *Determine log likelihood threshold values (LLR$_V$)*

The quantitative relationship between the number of Type I errors (*V*) and *LLR* were determined by simulating *D* genotypes without 'simulated recaptures' ($\widetilde{R} = 0$) following steps (i) to (iv) of *process* 8 described below. The *D* genotypes simulated in this process corresponds to *D*(*D*-1)/2 pairwise comparisons which was used as an estimator of m_0 as $m_0/(m_0 + m_1)$ is assumed to be close to unity. Genotype errors were not modelled in this process as the error model assumes the replacement of alleles drawn randomly from estimated population allele frequencies and would not contribute to differences in *LLR$_V$* estimates. Partial genotypes in the simulated data were created by selecting missing locus combinations chosen by a random draw with replacement of samples in the reference data. The simulated data files were replicated n_1 times (e.g. $n_1 = 100$) with the pairwise comparisons within each file emulating true nulls. Equation 4.3 was used to determine the *LLR* values of the pairwise comparisons from the n_1 files which were pooled and then sorted. The *LLR* that yields *V* Type I errors was determined by choosing the $(n_1 V)^{th}$ highest ranked *LLR* value from the sorted *LLR* values. Using this method, likelihood threshold values (LLR_V) were determined corresponding to fixed threshold values (*V*) of 0.01, 0.05, 0.125, 0.25, 0.50, 0.75, 1.00, 1.25 and from 1.50 up to 50.0 using increments of 0.5.

Process 4: *Find the number of matches (M) in the reference data.*

Pairwise comparisons in the reference data were made within the lower diagonal of the *D* by *D* pairwise matrix. Matching genotypes were defined as those sample pairs with their *LLR* greater than the LLR_V threshold value (Figure 4.1). A threshold of $LLR_{0.75}$ is expected to give 0.75 Type I errors within *M* matches.

Process 5: *Estimate 'sample recaptures' ($\hat{S}$) by correcting for Type I errors.*

Pairwise comparisons with *LLR* values above LLR_V are assumed to have $\hat{V} = \mathrm{E}(V)$ Type I errors with the number of matches *M* corrected for the number of Type I errors using: $\hat{S} = M - \hat{V}$. There may be many other recaptures within the data which could not be found as their *LLR* values were either too low, or perhaps too many genotype errors occurred between the two samples from the same individual. The total number of Type II errors (*T*) need to be estimated to obtain the number of 'corrected recaptures' ($\hat{R}$). Prior to estimating $\hat{R}$ the effective sample size needs to be determined.

Process 6: *Estimate effective sample size (E)*

The effective sample size (*E*) can be thought of as the maximum theoretical size of the dataset satisfying the constraint that the total number of Type I errors does not exceed V (Figure 4.1).

The logic behind estimating *E* is intuitive. *E* is deduced by emulating true positives from the data and counting the number of pairwise comparisons that did not exceed *V* Type I errors. The *LLR* of pairwise comparisons that are above the LLR_V threshold are called 'enabled' comparisons with their sum from all pairwise comparisons in the data equal to *K.* With one sample group $K = k_{1,1}$ and is calculated from the lower diagonal of all pairwise comparisons (Figure 4.3). The method to estimate *E* starts by initialising *K* to zero followed by five steps:

(i) Label the two genotypes being compared as g_c and g_d. Mimic pairwise comparisons between the two samples by only comparing loci that have alleles recorded in both samples g_c and g_d. Count the number of loci in which both genotypes g_c and g_d have no missing data to define L^*. Delete all loci in g_c that were missing in g_d to create genotype g_c*. Delete all loci in g_d that were missing in g_c to create genotype g_d*.

(ii) Mimic genotype errors between pairwise comparisons. The intuition behind this step is that higher error rates will reduce *E* which in turn will allow an additional correction for the total number of Type II errors caused by genotype errors. In this step copy g_c* twice to generate two new samples g_a and g_b, then for both genotypes g_a and g_b randomly select each locus using the per locus error rate (ε') and randomly select one of the two alleles which is then replaced from an allele drawn randomly from the allele frequencies estimated in the population. In this process all random sampling applied a uniform distribution.

(iii) Calculate log likelihood ratios. Using L^* loci calculate the *LLR* between the new genotypes g_a and g_b using equation 4.2 and 4.3 to give LLR_1^*. Repeat step (ii) above by replacing g_c* for genotype g_d* to give LLR_2^*.

(iv) Sum *K* to estimate total effective size. As two log likelihood ratios are determined from a single pairwise comparison in the data *K* is incremented by 0.5 to count the 'enabled' pairwise comparisons within the lower diagonal matrix. If LLR_1^* were greater than LLR_V then add 0.5 to *K*, also if LLR_2^* were greater than LLR_V then also add 0.5 to *K*, thus for each pairwise comparison *K* is incremented by either 0, 0.5 or 1. Repeat the above steps (i) to (iv) for every pairwise comparisons in the lower diagonal matrix to find the sum of *K*.

(v) Estimate the effective sample size (*E*) using the quadratic solution of $\hat{E}$ in $2K = \hat{E}(\hat{E} - 1)$ as:

$$\hat{E} = [(8K + 1)^{0.5} + 1]/2 \qquad (4.4)$$

which is a measure of the size of a square matrix holding the *K* lower diagonal comparisons.

Process 7: *Estimate the number of corrected recaptures ($\hat{R}$)*

If sample size (*D*) is equal to the effective size ($\hat{E}$) then the number of 'sample recaptures' ($\hat{S}$) is equal to the number of 'corrected recaptures' ($\hat{R}$). If not an additional correction was applied to estimate $\hat{R}$. Using $\hat{E}$ and $\hat{S}$, the probability of a recapture by drawing two samples at random (*PIR*) was derived by solving the binomial equation: $1 - (1 - PIR)^{\hat{E}} = \hat{S}/\hat{E}$. The corrected number of recaptures ($\hat{R}$) was then estimated for the total sample size *D* using equation 4.5 where $[1 - (1 - PIR)^D]$ is the probability of a match with *D* samples.

$$\begin{aligned}\hat{R} &= D[1-(1-PIR)^{D}] \\ &= D[1-(1-(1-\exp[\ln(1-\hat{S}/\hat{E})/\hat{E}]))^{D}]\end{aligned} \quad (4.5)$$

$\hat{R}$ includes both the sample recaptures ($\hat{S}$) and an estimate of the sum of all Type II errors ($\hat{T}$) with the power of the genotype data to identify individual recaptures estimated as:

$$1-(\hat{R}-\hat{S})/\hat{R} = 1-\hat{T}/\hat{R} = 1-\hat{\beta} \quad (4.6)$$

Process 8: *Simulation of genotypes*

As shown in Figure 4.2 the genotype data is simulated with n_2 runs (e.g. n_2=100) so that the standard error of $\hat{R}$ can be evaluated. Simulated recaptures ($\widetilde{R}$) can only be in whole numbers whereas the estimate of corrected recaptures ($\hat{R}$) can contain a fraction. The process of determining the number of simulated recaptures ($\widetilde{R}$) in each of n_2 runs is described in *process 9*. The simulated genotype data is created by random sampling from a uniform distribution using the following steps:

(i) generate (*D* - $\widetilde{R}$) genotypes of *L* loci by random sampling allele frequencies within each locus,

(ii) add $\widetilde{R}$ duplicate genotypes by copying them from step (i) to give *D* total genotypes,

(iii) add genotype errors in all *D* samples by randomly selecting each locus using the per locus error rate (ε') then randomly select one of the two alleles which is then replaced from an allele drawn randomly from the allele frequencies in the data, and

(iv) for every *D* genotype create missing loci by deleting all loci found missing in a randomly selected reference genotype by sampling with replacement.

Process 9: *Estimation of the standard error of corrected recaptures ($\hat{R}$)*

Using simulated data generated in *process* 8, there were n_2 estimates of corrected recaptures from simulated data ($\hat{R}_{sim}$) generated by performing *processes* 2 to 8 (Figure 4.2). Where the estimate of recaptures in the reference data ($\hat{R}$) was not a whole number, a random choice was used to determine the recaptures simulated ($\widetilde{R}$) *e.g.* if $\hat{R}$=2.2 then on average 80% (p_1) of simulations will have $\widetilde{R}_1$ =2 and 20% (p_2) will have $\widetilde{R}_2$=3. The standard deviation of $\hat{R}_{sim}$ estimates ($\sigma_{\hat{R}_{sim}}$) from n_2 simulated datasets were used to estimate the standard error of $\hat{R}$ from the reference data, by removing the variance attributed to rounding $\widetilde{R}$ to whole numbers (σ_w^2):

Standard error of $\hat{R}$ $= \sqrt{\sigma_{\hat{R}_{sim}}^2 - \sigma_w^2}$ (4.7)

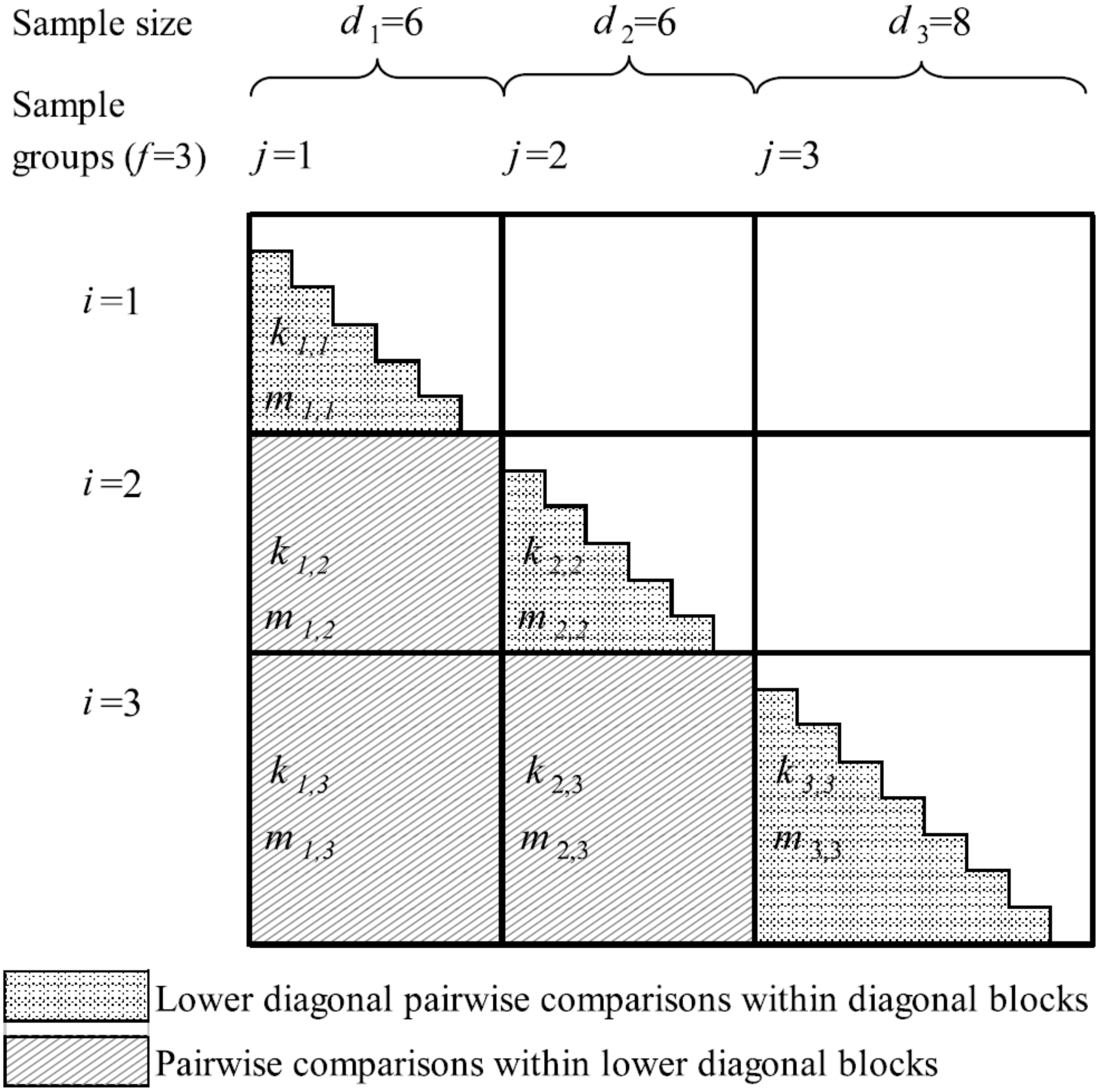
where $\sigma_w^2 = \left\{ n_2\left(p_1\widetilde{R}_1^2 + p_2\widetilde{R}_2^2\right) - \left[n_2\left(p_1\widetilde{R}_1 + p_2\widetilde{R}_2\right)\right]^2 / n_2 \right\} / (n_2 - 1) = n_2 p_1 p_2 / (n_2 - 1)$ given $\widetilde{R}_1 = \widetilde{R}_2 - 1$ with p_1 and p_2 determined from those proportions occurring within simulation runs.

Process 10: *Convergence of $\hat{R}$ with the smallest standard error.*

Following *processes* 2 to 7 (Figure 4.2) the reference data yielded a range of $\hat{R}$ recaptures, called $\hat{R}_V$ estimates each from a different fixed threshold value (*V*) of 0.01, 0.05, 0.125, 0.25, 0.50, 0.75, 1.00, 1.25 and from 1.50 up to 50.0 using increments of 0.5. In this final process the $\hat{R}_V$ estimate that gives the smallest standard error is called the 'converged solution' of $\hat{R}$.

Using *V*=0.75 Type I errors as a prior, $\widetilde{R} = \hat{R}_{0.75}$ recaptures were estimated from *D* genotypes sampled repeatedly from n_2 simulated datasets (*process* 8). From the range of fixed threshold values the *V* value yielding the smallest standard error of corrected recaptures (estimated from $\sigma_{\hat{R}_{sim}}$ the standard deviation of n_2 simulated datasets) is selected as the next *V* prior with $\widetilde{R} = \hat{R}_V$. The process is repeated with convergence defined when a new *V* prior is the same as one of the previous set of *V* priors used. The converged solution of $\hat{R}$ is set equal to the $\hat{R}_V$ solution which from the set of *V* priors gave the smallest $\sigma_{\hat{R}_{sim}}$ value. The converged solutions of *M,* $\hat{V}$, $\hat{E}$ and $\hat{T}$ are also those solutions corresponding to the *V* prior selected with the smallest $\sigma_{\hat{R}_{sim}}$.

Figure. 4.3. Example of 'block' structure for a mark-recapture study through space or time containing three groups of samples each of size 6, 6 and 8. Total sample size *D*=*d*$_1$+d$_2$+d$_3$. Vectors *m* and *k* are determined from estimates within the lower diagonal *D* matrix. The total number of matches $M = m_{1,1} + m_{1,2} + m_{1,3} + m_{2,1} + m_{2,3} + m_{3,3}$ and total number of enabled comparisons $K = k_{1,1} + k_{1,2} + k_{1,3} + k_{2,1} + k_{2,3} + k_{3,3}$.



4.3.3 Extension of theory to multiple sampling events

The method described so far can only be applied to finding recaptures within a single sampling event while an application to traditional mark-recapture studies requires recaptures between two or more groups of samples to be determined. Consequently, the mathematical theory needs to be extended to the analysis of multiple sampling events. *Processes* 1 to 3 are the same but the remaining processes were modified to determine recaptures within and between different groups.

The number of Type I errors was determined between each *i*th sampling group (d_i) to account for possible differences in genotype quality between them. Samples can be split into *f* spatial or temporal groups for analysis with pairwise comparisons between the groups forming a two dimensional block of individual comparisons (Figure 4.3) with total sample size $D = \sum_{i=1}^{f} d_i$ and total number of enabled comparisons in the lower diagonal matrix $K = \sum_{i=1}^{f} \sum_{j=i}^{f} k_{ij}$. The total number of pairwise comparisons that were not enabled is therefore $D(D-1)/2 - K$ with the number not enabled between the *i*th and *j*th group equal to $(d_i - 1)d_i / 2 - k_{ii}$ when *i=j,* and $d_i d_j - k_{ij}$ when $i \neq j$. It was assumed that the *V* Type I errors, pooled from all pairwise comparisons in the lower diagonal *D*x*D* matrix, are more likely to occur between the d_i and d_j blocks which had the most number of pairwise comparisons that were not enabled. The number of Type I errors within each block (v_{ij}) was therefore estimated by multiplying $\hat{V}$ to the proportional contribution of comparisons that were not enabled within the *i*th and *j*th group, to the total number that were not enabled as $\hat{v}_{ij} = \frac{\hat{V}\left[d_i(d_i - 1)/2 - k_{ii}\right]}{D(D-1)/2 - K}$ when $i = j$, and $\hat{v}_{ij} = \frac{\hat{V}\left[d_i d_j - k_{ij}\right]}{D(D-1)/2 - K}$ when $i \neq j$. However if by chance all pairwise comparisons were enabled with $K = D(D-1)/2$ then v_{ij} was partitioned using the proportional contribution of enabled comparisons within each block as $\hat{v}_{ij} = \frac{\hat{V}\left[d_i(d_i - 1)/2\right]}{D(D-1)/2}$ when $i = j$, and $\hat{v}_{ij} = \frac{\hat{V}\left[d_i d_j\right]}{D(D-1)/2}$ when $i \neq j$.

Table 4.3. Analysis of multiple sampling events (*i.e.* groups of spatial or temporal data). Given there are $\hat{s}_{ij}$ sample recaptures and k_{ij} enabled comparisons within the *i*th and *j*th groups; the effective block size (b_{ij}), effective sample size ($\hat{e}_{ij}$), probability of an individual resample (pir_{ij}) and corrected number of recaptures ($\hat{r}_{ij}$) can be determined from the number of samples d_i and d_j. The Lincoln-Petersen population estimate ($\hat{N}$) is also determined where d_i is the number captured on the first visit, d_j is the number captured on the second visit and $\hat{r}_{ij}$ is the number captured on the second visit that were captured on the first visit.

	Diagonal block (*i*=*j*)	Off-diagonal block (*i*≠*j*)
b_{ij} =	d_i	$(d_i d_j)^{0.5}$
$\hat{e}_{ij}$ =	$[(8k_{ij}+1)^{0.5}+1]/2$	$k_{ij}^{0.5}$
pir_{ij} =	$1-\exp[\ln(1-\hat{s}_{ij}/\hat{e}_{ij})/\hat{e}_{ij}]$	$1-\exp[\ln(1-\hat{s}_{ij}/\hat{e}_{ij})/\hat{e}_{ij}]$
$\hat{r}_{ij}$ =	$b_{ij}[1-(1-pir_{ij})^{b_{ij}}]$	$b_{ij}[1-(1-pir_{ij})^{b_{ij}}]$
$\hat{N}$ =		$\dfrac{(d_i+1)(d_j+1)}{\hat{r}_{ij}+1}-1$

The number of matches between the *ij*th groups (m_{ij}) were determined using the *LLR*$_V$ threshold and therefore the total number of matches found in the entire dataset (*M*) remains the same even if the data is partitioned into *f* groups with $M = \sum_{i=1}^{f}\sum_{j=i}^{f} m_{ij}$. The number of 'sample recaptures' within a block ($\hat{s}_{ij}$) was then determined as $\hat{s}_{ij} = m_{ij} - \hat{v}_{ij}$. The estimate of $\hat{s}_{ij}$ is used to determine the probability of an individual recapture between the *ij*th groups (pir_{ij}) which is then used to estimate the number of corrected recaptures between the same two groups ($\hat{r}_{ij}$) (Table 4.3). The equation for the Lincoln-Petersen (Serber, 1982) population estimate ($\hat{N}$) is also provided in Table 4.3 which assumes equal catchability of individuals between the *i*th and *j*th groups.

With multiple sampling events the method to estimate the standard error of $\hat{r}_{ij}$ was similar to that described in *processes* 8 and 9. The differences are that missing loci combinations were randomly sampled with replacement from those observed within each of the reference genotype groups. This allowed Type II errors to be more accurately assessed given potential genotype quality differences amongst the groups or sampling sessions.

4.3.4 Testing theory through stochastic simulation

Reference data genotypes were simulated to test the robustness of estimating recaptures using the methods described here. Genotypes were generated consisting of *L* loci with either (a) multi-allelic frequencies of 0.25, 0.25, 0.2, 0.15, 0.05, 0.05, 0.02, 0.01, 0.01, 0.005 and 0.005 (Kalinowski et al., 2007) which are thought to be typical for empirical microsatellite allele frequencies or (b) with bi-allelic frequencies of 0.45 and 0.55. Loci were assumed to be unlinked and individuals unrelated. The reference data files were generated using steps (i) to (iii) as described in *process 8* with $\widetilde{R}$ equal to the simulated number of recaptures. Step (iv) was performed using uniform random sampling of missing loci from a given proportion of loci successfully genotyped (*Y*).

The simulation parameters used in this study are listed in Table 4.4. These are selected combinations of parameters used to demonstrate the theory with a focus on testing the robustness in extremely challenging data of low quality. When estimating total recaptures (*R*) the solutions presented are those that converged using *process 10*. The mathematical theory was tested by repetitive sampling of *D* reference genotypes to create n_3 reference files for a given set of simulation parameters. The converged solutions of *M, V,* $\hat{E}$*,* $\hat{T}$ and $\hat{R}$ from the n_3 reference files were collated with means and standard deviations determined. Within each of the n_3 datasets standard errors were estimated using n_2 simulation runs (Figure 4.2.). The bias was measured as the difference between the mean converged recapture estimates ($\overline{\hat{R}}$) and the theoretical expectation of simulated recaptures, $\mathrm{E}(\widetilde{R})$, using the formulae: $Bias(\overline{\hat{R}},\widetilde{R}) = \overline{\hat{R}} - \mathrm{E}(\widetilde{R})$.

The effect of family relationships on recapture estimates was estimated by populating the simulated data with five sets of sire and dam pairs each with five full-sib offspring. As described above genotype errors were added to the parent and offspring genotypes with missing loci created from a given proportion of loci successfully genotyped (*Y*).

4.4. Results

4.4.1 Effect of the presence of partial genotypes on recapture estimation

The simulation parameters for each results section are listed in Table 4.4. Simulations using multi-allelic loci were used to examine the relationship between the proportion of loci successfully genotyped and its effect on the accuracy of estimating recapture numbers

Table 4.4 Summary of simulation parameters used to demonstrate theory with the location of each shown within the results section. Parameters include: data type, number of simulated datasets to determine the *LLR* threshold (n_1), number of simulated datasets to determine the standard error of $\hat{R}$ (n_2), number of replicated reference data files (n_3), number of simulated recaptures ($\widetilde{R}$), genotype sample size (*D*), number of loci in a genotyping panel (*L*), random proportion of loci successfully genotyped (*Y*), genotype error rate per locus used in equation 4.2 (ε) and the genotype error rate simulated in the data (ε').

(Results section) Location	Data type	n_1	n_2	n_3	$\widetilde{R}$	*D*	*L*	*Y* (%)	ε (%)	ε' (%)
(4.4.1) Figure 4.4	Multi-allelic	100	100*	1	20	520	8	60	1	1
		100	100*	1	20	520	8	70	1	1
		100	100*	1	20	520	8	80	1	1
		100	100*	1	20	520	8	90	1	1
(4.4.1) In text	Bi-allelic	100	100*	1	20	520	20	60	1	1
(4.4.2) Figure 4.5	Bi-allelic	100	100*	1	20	520	14	80	1	1
		100	100*	1	20	520	16	80	1	1
		100	100*	1	20	520	18	80	1	1
		100	100*	1	20	520	20	80	1	1
(4.4.2) In text	Multi-allelic	100	100*	1	20	520	4	80	1	1
		100	100*	1	20	520	6	80	1	1
(4.4.3) Table 4.5	Multi-allelic	100	100	200	20	270	8	80	1	1
		100	100	200	20	1020	8	80	1	1
		100	100	200	20	4020	8	80	1	1
(4.4.3) Table 4.5	Bi-allelic	100	100	200	20	270	20	80	1	1
		100	100	200	20	1020	20	80	1	1
		100	100	200	20	4020	20	80	1	1
(4.4.4) Table 4.6	Multi-allelic	100	100	200	20	520	8	80	0.1	0.1
		100	100	200	20	520	8	80	1	1
		100	100	200	20	520	8	80	5	5
		100	100	200	20	520	8	80	0.1	5
		100	100	200	20	520	8	80	5	0.1
(4.4.4) Table 4.6	Bi-allelic	100	100	200	20	520	20	80	0.1	0.1
		100	100	200	20	520	20	80	1	1
		100	100	200	20	520	20	80	5	5
		100	100	200	20	520	20	80	0.1	5
		100	100	200	20	520	20	80	5	0.1
(4.4.5) In text	Multi-allelic	100	100	100	100	500	7	75	1	1
(4.4.6) In text	Multi-allelic	100	100	100	20	200	7	100	5	5
(4.4.7) In text	Multi-allelic	1000	100	100	20	520	8	80	1	1

* Constant value of $\widetilde{R}$=20 used in step 8 of Figure 4.2 with the standard deviation of $\hat{R}$ from n_2 runs used as an estimate of the standard error of $\hat{R}$ given a single data set.

(Figure 4.4). Reducing the expected number of Type I errors used as a threshold to define a match (*V*) did not lead to a reduced number of estimated recaptures (Figure 4.4a) as this was offset by an increased number of Type II errors ($\hat{T}$). Figure 4.4a shows that in genotype files with as little as 60% of loci genotyped, the number of estimated recaptures ($\hat{R}$) were similar to the number of simulated recaptures ($\widetilde{R}$=20). The standard error of $\hat{R}$ was smallest when 90% of loci were genotyped with *V*=0.125 (Figure 4.4b). Data that had 70% fully genotyped had the smallest standard errors when *V* was between 1.5 and 2.0 (Figure 4.4b). As expected, as the number of Type I errors (*V*) increased, the number of Type II errors ($\hat{T}$) decreased (Figure 4.4c). The number of Type II errors was largest when only 60% of loci were genotyped (Figure 4.4c).

Similar trends were observed when loci were simulated to be bi-allelic with a higher accuracy in the number of recaptures estimated as the proportion of loci successfully genotyped increased (see simulation parameters: Table 4.4). To minimise the standard error for bi-allelic data when only 60% of loci were genotyped, it was necessary to set *V*=5.5 which resulted in $\hat{T}$=12.5 and $\hat{R}$=20.2 ($\underline{+}$7.9 standard error).

Figure. 4.4. Configuration with 60% (o), 70% (Δ), 80% ($\times$) and 90% ($\bullet$) loci randomly genotyped. The relationship between the expected number of Type I errors (*V*) used as a threshold to define a match on (a) average number recaptured ($\hat{R}$), (b) standard error of $\hat{R}$ from a single estimate and (*c*) average number of Type II errors ($\hat{T}$); determined from $n_2 = 100$ reference files simulated using eight multi-allelic loci, 1% random genotype errors per locus and *D*=520 genotype samples inclusive of $\widetilde{R}$=20 recaptures.

Figure 4.4 (a)

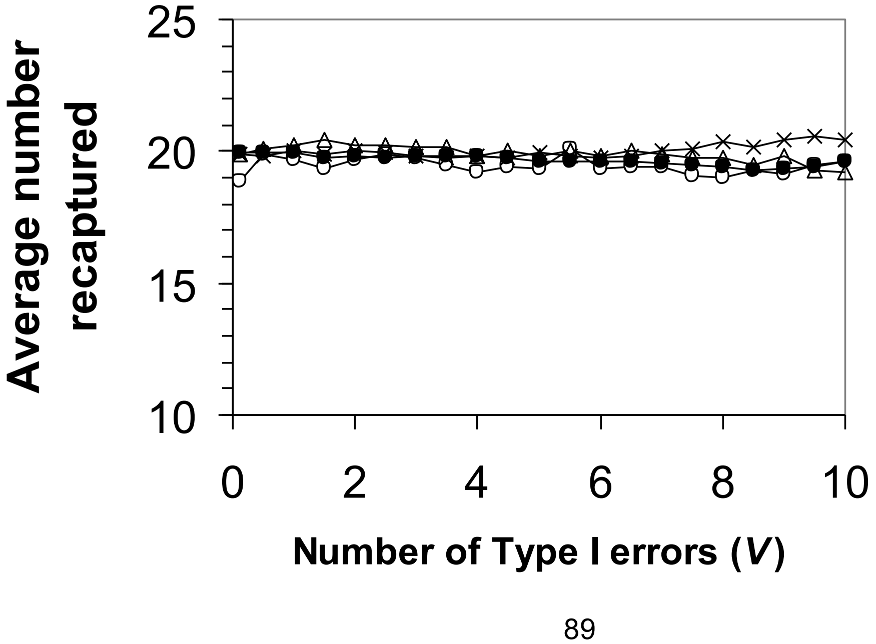


Figure 4.4 (b)

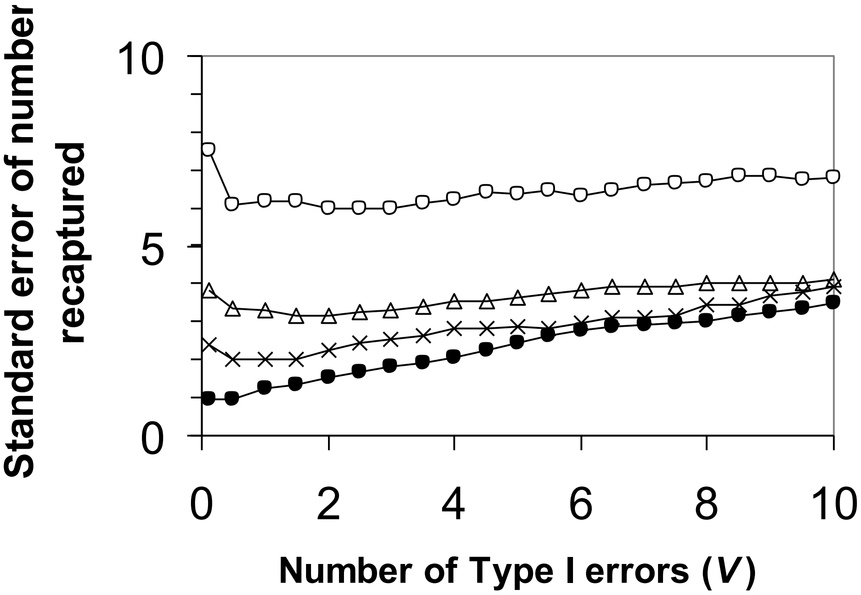


Figure 4.4 (c)

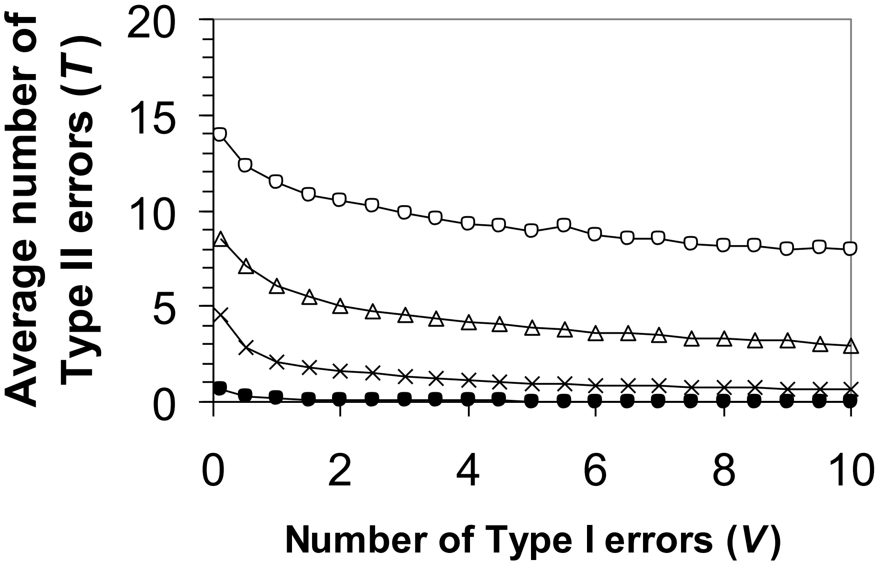


Figure. 4.5. Configuration with 14 bi-allelic loci (o), 16 bi-allelic loci (Δ), 18 bi-allelic loci ($\times$) and 20 bi-allelic loci ($\bullet$). The relationship between the expected number of Type I errors (*V*) on (a) average number recaptured (*R*), (b) standard error of $\hat{R}$ from a single estimate, (*c*) average number of Type II errors ($\hat{T}$); determined from $n_2 = 200$ reference files simulated using 1% random genotype errors per locus, 80% loci randomly genotyped, *D*=520 genotype samples inclusive of $\tilde{R}$=20 recaptures. Allelic frequencies were 45% and 55% across all loci.

Figure. 4.5 (a)

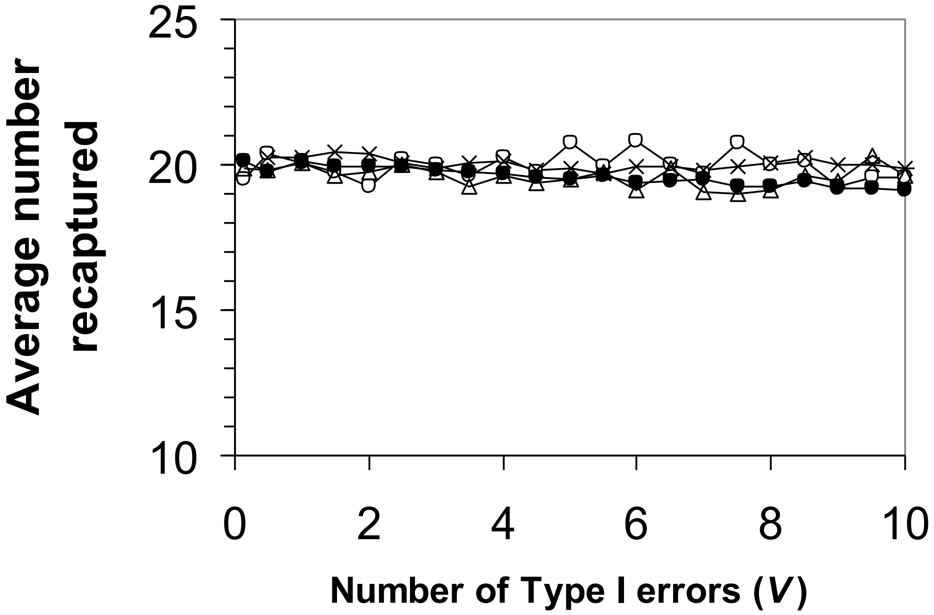


Figure. 5 (b)

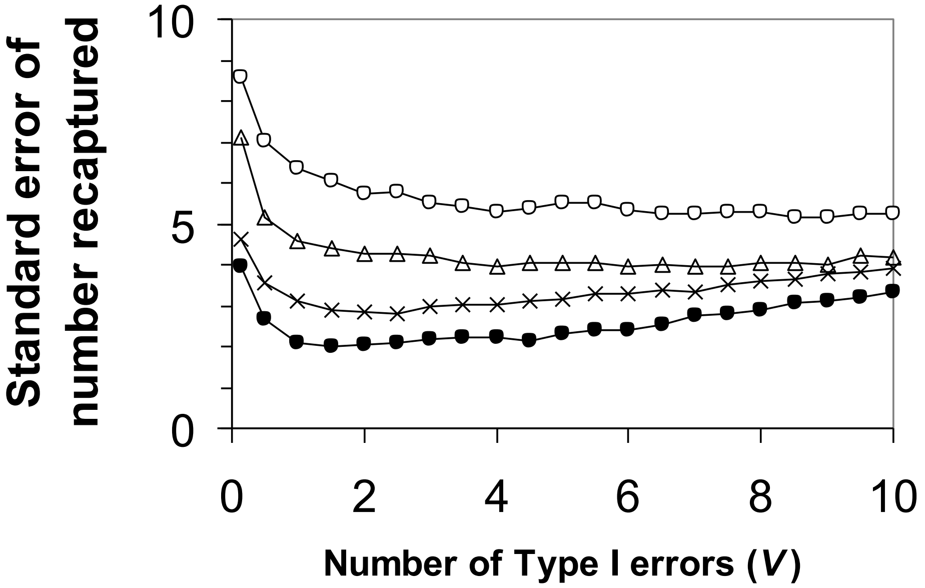
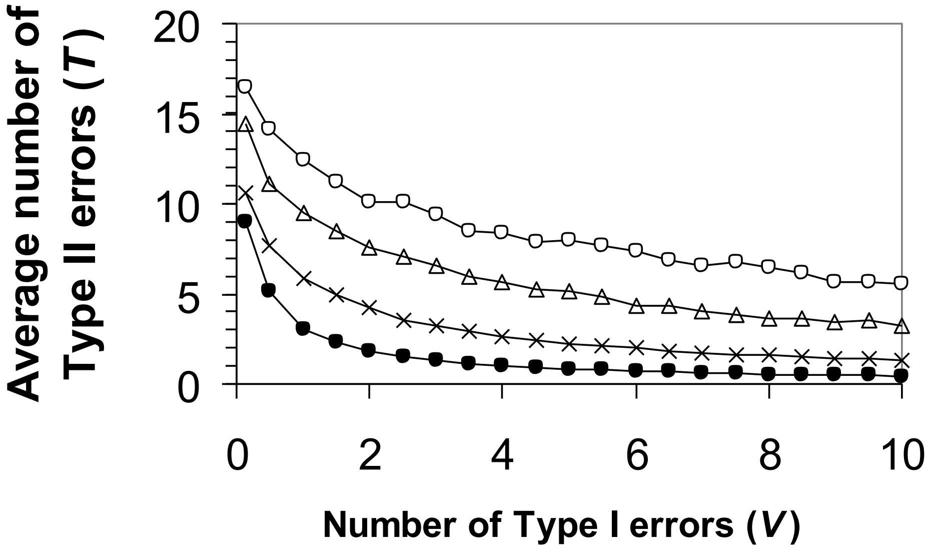


Figure. 4.5 (c)



4.4.2 Effect of number of loci in the genotyping panel on recapture estimation

Simulations were used to examine the relationship between the number of bi-allelic loci and the accuracy of estimating recapture numbers when the probability of being genotyped was kept constant at 80%. For panels of 14, 16, 18 or 20 loci, the number of estimated recaptures ($\hat{R}$) were similar to the number of simulated recaptures ($\tilde{R}$ =20) across a range of *V* values from 0.125 to 10 (Figure 4.5a). For each of the panels there was a definite range of *V* values where the standard error of $\hat{R}$ was smallest (Figure 4.5b). Genotypes consisting of 20 loci gave the smallest $\hat{R}$ standard errors of between 2.0 and 2.1 for *V* values ranging from 1.0 to 2.5 (Figure 4.5b). As expected, when there were fewer loci per genotyping panel, the standard error of $\hat{R}$ was larger. For genotypes consisting of 14 to 20 loci, the number of Type II errors ($\hat{T}$) was at least five when *V*=0.5 (Figure 4.5c). However, as *V* increased, $\hat{T}$ decreased with $\hat{T}$ being smaller when the number of loci per genotype was 20 compared to 18, 16 or 14.

Similar trends were observed when loci were simulated to be multi-allelic. When recaptures were determined with a six locus panel (see simulation parameters: Table 4.4), the smallest standard error of estimated recaptures occurred when *V* was 1.5 with $\hat{T}$ equal to 6.3 and estimated recaptures $\hat{R}$ =20.2+(3.1 standard error). With a panel of four multi-allelic loci, the smallest standard error of estimated recaptures occurred when *V* was 3.5 with $\hat{T}$ equal to 11.1 and estimated recaptures $\hat{R}$ =19.7+(7.0 standard error).

4.4.3 Effect of sample size on recapture estimation

The effect of multi-allelic or bi-allelic loci using different sample sizes was evaluated on the accuracy of estimating recapture numbers again assuming 80% of loci per genotype were successfully typed (see simulation parameters: Table 4.4). In this example there were n_3=200 reference genotype files, simulated for each sample size 270, 1020 and 2020, and within each data type. Parameters that converged with the smallest standard error of $\hat{R}$ are summarised in Table 4.5. As sample size increased it became more difficult to identify individual recaptures with the number of Type II errors increasing together with the larger standard deviation of $\hat{R}$ estimates. No significant bias in estimating the number recaptured compared to the number of simulated recaptures ($\tilde{R}$=20) was evident with the number of Type II errors increasing with sample size.

The power to identify individual recaptures (1-$\hat{\beta}$) decreased from 0.95, 0.80 and 0.70 in the biallelic data, and 0.93, 0.83 and 0.75 in the multi-allelic data, with *D* equal to 270, 1020 and 2020 respectively.

Table 4.5. The effect of the number of genotype samples (*D*) on recapture estimation with bi-allelic or multi-allelic loci. Converged results of $\hat{E}$, *M*, $\hat{V}$, $\hat{T}$ and $\hat{R}$ (mean ± standard deviation) from n_3=200 simulated genotype files each containing $\tilde{R}$=20 recaptures. Other fixed parameters included percentage genotyped per locus *T*=80% and genotype error rate per locus $\varepsilon = \varepsilon'$ =1%.

Number of samples	Effective sample size	Matches found	Type I errors	Type II errors	Corrected recaptures
D	$\hat{E}$	*M*	$\hat{V}$	$\hat{T}$	$\hat{R}$
Bi-allelic data using 20 loci					
270	261±2	20.0±1.5	1.1±0.3	1.1+0.3	20.0±1.4
1020	909+24	18.4±3.6	2.5±1.1	4.0±0.9	19.9±3.1
2020	1672±70	17.8±4.5	4.0±1.7	6.1±1.4	19.9±4.0
Multi-allelic data using 8 loci					
270	261±2	19.3±1.5	0.7±0.2	1.4±0.3	20.0±1.5
1020	931±14	17.7±2.2	1.2±0.5	3.3±0.6	19.8±2.2
2020	1754±39	16.4±2.9	1.4±0.7	4.8±0.9	19.8±3.0

4.4.4 Effect of genotype errors on recapture estimation

Using parameters listed in Table 4.4, the sensitivity in estimating $\hat{R}$ was evaluated when the error rate in the data (ε') was different from the assumed error rate (ε) used in equation 4.2. As expected an increase in error rates from 0.1% to 5.0% ($\varepsilon = \varepsilon'$) caused an increase in the standard deviation of $\hat{R}$ in both bi-allelic and multi-allelic data sets Table 4.6a.

Table 4.**6** The effect of genotype errors per locus on recapture estimation when estimating *LLR* using ε when the true error rate in the reference data is ε' and assuming bi-allelic or multi-allelic loci. Converged results of $\hat{E}$, *M*, $\hat{V}$, $\hat{T}$ and $\hat{R}$ (mean ± standard deviation) from n_3=200 simulated genotype files each containing $\tilde{R}$=20 recaptures. Other fixed parameters included, percentage genotyped per locus *Y*=80% and number of samples *D*=520.

	Error rate per locus		Effective	Matches	Type II	Type II	Recaptures
	(used)	(true)	sample size	found	errors	errors	
	ε (%)	ε' (%)	$\hat{E}$	*M*	$\hat{V}$	$\hat{T}$	$\hat{R}$
Bi-allelic data using 20 loci							
(a)	0.1	0.1	506±4	20.3±1.8	1.4±0.3	1.4±0.3	19.9±1.6
	1.0	1.0	490±6	19.2±2.2	1.6±0.5	2.2±0.5	19.8±2.1
	5.0	5.0	436±15	16.6±3.7	2.7±1.3	3.9±0.6	19.6±3.6
(b)	0.1	5.0	423±25	16.4±5.3	3.9±2.0	5.2±0.9	18.6±4.6*
	5.0	0.1	494±6	19.5±2.1	1.6±0.5	1.8±0.3	19.8±2.1
Multi-allelic data using 8 loci							
(a)	0.1	0.1	501±5	19.6±1.7	1.0±0.3	1.0±0.3	20.0±1.6
	1.0	1.0	493±5	19.1±1.9	1.1±0.4	3.9±0.6	20.1±1.9
	5.0	5.0	465±7	17.4±2.7	1.5±0.6	5.7±1.3	19.8±2.8
(b)	0.1	5.0	444±8	15.5±2.6	1.4±0.5	6.1±1.7	19.4±3.2*
	5.0	0.1	495±5	19.1±1.9	1.1±0.4	1.9±0.5	19.8±1.8

* Average of $\hat{R}$ from 200 simulated genotypes is lower than $\mathrm{E}(\tilde{R}) = 20$ (P< 0.01).

When incorrectly assigning a wrong prior for genotype error ($\varepsilon \neq \varepsilon'$) the standard deviation of $\hat{R}$ increased when either over-estimating (ε =5%, ε' =0.1%), or under-estimating (ε =0.1%, ε' =5.0%), the genotype error rate (Table 4.6b) compared to simulations when the true error rate was used (Table 4.6a). Over the 200 datasets bias in the theoretical estimates was not significant (P>0.05) when over-estimating true genotype error rates but a significant bias (P<0.01) was detected when under-estimating them. This is a theoretical bias and was not significant when compared to the variation expected in estimating $\hat{R}$ from a single genotype file.

4.4.5 Effect of recapture numbers on recapture estimation

To test the theory with a larger number of recaptures n_3=100 reference files were simulated each with $\tilde{R} = 100$ recaptures and *D*=500 samples using the multi-allelic frequencies with only 75% genotyped and error rates $\varepsilon' = \varepsilon$ =1% (see simulation parameters: Table 4.4).

To find individual matches a small expected level of Type I errors were tolerated using *V*=0.25. With this restriction over half the matches with *M*=57.1±5.1 (s.d.) were detected, effective size $\hat{E}$ =379±9 (s.d.) and corrected recaptures $\hat{R}$ =94.4±7.4 (s.d.). In the n_3=100 datasets there were 0, 1 and 2 Type I errors at respective frequencies of 77, 17 and 6 and averaged 0.29 which was close to the expected *V*=0.25 value.

To find the best population estimate of $\hat{R}$ the converged estimates of $\hat{R}$ having the smallest variance via *process* 10 were determined. Over the 100 datasets there was improved precision (P<0.01) and accuracy (P<0.01) in estimating the number of recaptures with $\hat{R}$ =98.6±4.9 (s.d.) compared to the first estimate with *V* fixed at 0.25. In this case the effective size increased to *E*=449±10 (s.d.) and *V* averaged 18.0±3.3(s.d.).

4.4.6 Effect of genotype errors on recapture estimation with $\hat{E}/D$ close to unity

Using the multiallelic frequencies of seven loci *D*=200 genotype samples were generated containing 20 recaptures in each simulated dataset. Each dataset had all genotypes amplified and a high error rate per locus of 5%. In n_3=100 datasets there were in total 2000 simulated recaptures of which 1968 were correctly identified when assigning matches from the converged solutions of $\hat{R}$ with the smallest variance (*process 10*). In each of the 100 data sets there were 0, 1 and 2 Type I errors at frequencies of 90, 8, and 2 respectively with $\hat{E}/D = 0.99$ and $(1-\hat{\beta}) = 0.98$.

The correctly identified recaptures in each of the 100 data sets had 0, 1, 2, 3 and 4 genotype errors at average counts of 11.4, 6.7, 1.5, 0.06 and 0.01 respectively. A mismatch analysis (Paetkau, 2003; McKelvey and Schwartz, 2005) revealed that there was a continuous distribution of mismatches with no zero counts between them in 92% of these simulations with average mismatches at 0, 1, 2, 3, 4, 5, 6 and 7 loci of 11, 7, 2.1, 9, 145, 1243, 6024 and 12457 respectively. Therefore, if a mismatch analysis were used in this data, it would be possible to infer recaptures only 8% of the time compared to the new method which could infer matches 100% of the time with approximately 0.12 Type I errors in each sample. This demonstrates that the new methods used to identify recaptures can be applied in datasets where other methods fail.

4.4.7 Effect of full-sibs on recapture estimation

In addition to the usual simulation parameters, $\widetilde{R} = 20$, $D = 520$, $Y = 80\%$ and $\varepsilon = \varepsilon' = 1\%$ (see simulation parameters: Table 4.4), the reference data was populated with five sets of sire and dam pairs each with five full-sib offspring. In this structure there were $5(5\times4)/2 = 50$ pairwise comparisons between full-sibs and $5(2\times5) = 50$ pairwise comparisons between parent and offspring giving a total 100 pairwise comparisons each with a 50% coefficient of relationship.

In 100 reference datasets simulated using multi-allelic datatypes, with full-sibs and parents as described above, *process 10* was used to converge to a recapture estimate with the smallest variance. This resulted in higher than expected recaptures ($\mathrm{E}(\widetilde{R})$=20.0) with mean and standard deviation of $\hat{R}$ =22.5+2.7 with the bias being significant over the 100 datasets (P<0.01) but not from a single dataset (P>0.05). Over the 100 datasets there were 0, 1, 2, 3 and 4 Type I errors, at frequencies of 97, 70, 23, 9 and 2 from full-sib matches, and parent-offspring matches with $\hat{E} = 491 \pm 5$.

Another 100 reference datasets were simulated with full-sibs and parents as described above. This time recapture numbers were determined from a small fixed level of Type I errors by constraining *V* to a value of 0.01 with the relationship between *V* and *LLR* determined using $n_1 = 1000$ (Figure 4.2). The total number of Type I errors over the $n_3 = 100$ reference datasets from full-sib and parent-offspring matches fell to 26 averaging

0.26 per dataset. There was no significant bias (P>0.05) in estimating recapture numbers at the expense of an increased standard deviation giving $\hat{R}$ =20.1$\underline{+}$4.1 and $\hat{E}$ =402$\underline{+}$11.

The standard deviation of $\hat{R}$ from n_3 reference files was similar to the average of n_3 estimated standard errors of $\hat{R}$ determined from n_2 simulated datasets. This result was consistent with other simulations reported.

In an applied extension to this research the number of full-sibs in genotype data produce a signature number of false positives which can be used to estimate the proportion of full-sibs (Section 4.11, Appendix II – percentage sibship estimation).

4.5. Discussion

Previous studies could not be identified that defined likelihood equations to rank genotype matches with missing data (as described in this chapter), although likelihood equations have been previously applied to parent-offspring pedigree relationships (Marshall et al., 1998; Kalinowski et al., 2007). While much attention in the molecular literature has been placed on genotyping errors (Wang, 2004; Waits and Paetkau, 2005; Beja-Pereira et al., 2009) and to a lesser extent shadows (Mills et al., 2000; Paetkau, 2003; Hoyle et al., 2005), no other papers addressing the importance of missing data, effective size, Type I errors and Type II errors in a mark-recapture framework were found. This may be partly due to genetic mark-recapture studies often having small sample sizes where the problem of Type I errors is less acute. Nonetheless, a survey of the literature shows that while most studies are aware of Type I errors, few studies formally take sample size into account such as Paetkau (2003) and Hunter et al. (2010).

Unknown is the number of studies that were abandoned because of a perceived lack of statistical power to distinguish between the null and alternative hypothesis. The $\hat{E}/D$ statistic can be used to determine if all individuals in a dataset are comparable at a given number of Type I errors. Even if all individuals cannot be uniquely identified, then it is possible to estimate the number of recaptures in a dataset using the method presented here. The key to arriving at an accurate estimate of recaptures is appreciating the trade-off that exists between Type I and Type II errors and how they influence the effective size of the dataset. The *a posteriori* approach is one option in dealing with missing data. A preferable option would be to use sampling methods that would decrease the amount of missing data followed by careful laboratory methods to minimise genotyping errors (Taberlet et al., 1996; Morin et al., 2001; Paetkau, 2003).

The robustness of this method was evaluated against a range of simulated variables. It was not practical to test all combinations of these variables with a cross-sectional approach used by examining the change in one parameter by holding other parameters constant. Adding more loci in the multi-allelic and bi-allelic data types will improve the ability to find correct recaptures.

Datasets of the highest quality, where $\hat{E}/D$ is close to unity at low Type I error levels, will be suitable for all analyses including those traditional mark-recapture approaches that require unambiguous presence/absence data for each individual. In lower quality datasets (*E*/*D* < 1) not all pairwise comparisons can be assessed and Type II errors are expected to occur. When *E*/*D* is less than one the methodology can estimate the total number of recaptures more accurately by accounting for Type II errors leading to the estimation of biologically important parameters *e.g.* recapture rate, migration rate and population size based on the Lincoln-Petersen (Flagstad et al., 2004; Leigh et al., 2006; Seber, 1982) and other methods (e.g. those in MARK; White and Burnham, 1999).

It is important to realise that even in cases where *E/D* < 1, standard mark-recapture methods such as CAPWIRE (Miller et al., 2005) that require individuals of known re-capture histories can be used. In this case a reduced set of pairwise comparisons within the effective size can be utilised where all specific genotype pairs are individually identified. Operationally this is achieved by setting a low Type I error threshold, such as *V*=0.25, and using the detected matches (m_{ij}) and corresponding effective size ($\hat{e}_{ij}$) instead of the true sample size (d_i) as input into the mark-recapture analysis. This approach is similar to culling samples with low number of amplified loci (Paetkau, 2003; Rudnick et al., 2008). In contrast, the method culls on individual pairwise comparisons thereby removing less data and maximising the chances of finding recaptures.

It is desirable to estimate the error rate per locus to obtain more accurate and precise estimates of recaptures. If there were sufficient numbers of recaptures in a genetic mark-recapture study, then the methods used here could be used to estimate the genotype error rate among recaptures without the need for replication. Finding errors from matches using *a posteriori* approach is attractive especially in large datasets where the multi-tube method (Taberlet et al., 1996; Bellemain et al., 2005) can exhaust finite samples and be prohibitive in time and cost. Genotying errors can be estimated directly by repeatedly genotyping the same sample, but this is an expensive process (Johnson and Haydon, 2007). Error rates per locus of between 0.5% and 1% are usual in many laboratories but higher error rates are known to occur especially in studies that involve limited amounts of poor quality DNA (Pompanon et al., 2005; Soulsbury et al., 2007).

The results of this study (Table 4.6) show that estimated recapture numbers were not biased when overestimating genotype error rates, however, underestimating prior genotype error rates could bias results. Interestingly, there is no requirement using the new approach to arrive at an accurate prior for genotyping error rate (ε), only that the value chosen in equation 4.2 is higher than the actual error rate in the data (ε'). The reason for this is that true matches are more likely to be refused when $\varepsilon' > \varepsilon$ resulting in smaller than expected recapture estimates. Conversely when $\varepsilon' < \varepsilon$, Type I errors are inadvertently estimated as being more prevalent for a given *LLR* threshold. This reduces the effective size with the number of Type II errors adjusted to yield unbiased recapture estimates at the expense of increasing the standard error.

The theory to estimate recaptures in multiple sampling events was provided, however a single sampling event could be divided and analysed as a multiple sampling event. This may be desirable when missing loci among genotypes and mistypings are thought not to be independent, for example when a higher mistyping rate is expected among partial genotypes compared to complete genotypes. In this case the genotype samples could be split into partial and complete genotype datasets and applying a higher error rate to the partial dataset. Alternatively the higher error rate could be used to analyse all the data as it is not expected to bias the result as discussed above. If available, different error rates for each loci could also be easily implemented in the new methods.

Kalinowski et al. (2007) reported that the value of increasingly realistic models of genotyping error has not been tested against simpler models. The estimation of likelihood ratios using equation 4.2 is simple and also computationally efficient. A more complex error model could be developed by partitioning genotyping errors into allelic dropouts and other false alleles (Wang, 2004). As discussed above, the method is tolerant of the estimated frequency of genotyping errors and it is probably less likely that a more complex error model could yield a significant improvement in accuracy. This is supported by Wang (2004) who concluded that accurate sibship inference can be obtained using wildly guessed typing errors provided sufficient information is provided in the data.

In practice, the methods can identify allelic dropouts in putative recaptures. Not only can the electropherogram be re-examined in putative matches to see if there have been any miss-callings, but the methods can also list rejected matches with high likelihood ratios that were almost accepted as being a match. These rejected matches, especially those with one allelic dropout, gives the molecular geneticist a handy tool that could be used to shortlist the number of samples to re-examine.

Mismatch techniques used to identify potential genotyping errors and determine whether loci used will discriminate among individuals (Paetkau, 2003; McKelvey and Schwartz, 2005) are powerful *a posteriori* tools because they make few assumptions and are not affected by relatedness in the population. They only require the bimodality of the mismatch distribution to separate true recaptures from other pairwise combinations. Unless an exorbitant number of loci are used this bimodality breaks down quickly when partial genotypes are considered, with the methods presenting a favourable means of estimating recapture numbers.

One of the assumptions in the method is that the genotype samples were collected from an outbred population. As demonstrated over-estimates of recapture numbers can occur in data sets containing samples related by kinship as this introduces a source of Type I error that is not identified, nor corrected, by the methods. It was also demonstrated that the bias in recapture numbers can be reduced by fixing the number of Type I errors to a conservatively low level (e.g. V=0.01). Alternatively, if kinship structures of the population are known they could be modelled in the simulated population when determining the relationship between the number of Type I errors and LLR so that the additional Type I errors caused by kinship can be identified with unbiased estimates of recapture numbers determined.

4.6. Software

A program called ‘SHAZA’ (SHAdow Zone Analysis) was specifically developed to conduct genotype match analysis using the equations described in this paper. The SHAZA user manual and software (including ANSI C source code) are available via the internet from the authors on http://molecularfisherieslaboratory.com.au/shadow-zone-analysis-software-shaza

SHAZA is a versatile software package that allows users to generate and analyse mark-recapture datasets that have missing loci profiles and typing errors derived from either pilot or simulated data. Under Hardy-Weinberg assumptions this program provides researchers with a realistic framework within which to evaluate the efficacy of a program under a range of dataset sizes and marker resolution.

4.7. Conclusions

For wildlife studies, where statistical power to distinguish between the null and alternative hypothesis is often compromised for a variety of reasons, the methods presented here offer a superior means of robustly estimating recapture numbers. This has been demonstrated by the application of the methodology to simulated datasets. Bearing in mind of the assumptions in the model, the method described here (as implemented in the SHAZA software) is recommended for finding and estimating recaptures in a range of scenarios where genotype data has one or a combination of the following features:

(i) has genotypes with missing locus data,

(ii) has genotype errors prevalent and

(iii) if the likelihood match/mismatch distributions overlap making recapture inferences difficult if not impossible to achieve.

It is also recommend that the use of SHAZA should be used to assist in the experimental design of non-invasive wildlife studies.

4.8 Acknowledgements

This work was supported with grants from the Australian Fisheries Research Development Corporation. Simon Hoyle made valuable early contributions to the work on shadows. Professor Lorenz Hauser provided useful discussions in the development of this method which led to the use of the likelihood approach. Thanks to Dr Warwick Nash, Dr Hock Seng Lee, Dr George Leigh, Dr Tony Swain, Dr Alison Kelly and Dr Jessica Morgan who reviewed this paper prior to submission, and to anonymous reviewers for their contributions.

4.9 References

Beja-Pereira A, Oliveira R, Alves PC, Schwartz MK, Luikart G (2009) Advancing ecological understandings through technological transformations in noninvasive genetics. Molecular Ecology 9:1279-1301.

Bellemain E, Swenson JE, Tallmon D, Brunberg S, Taberlet P (2005) Estimating population size of elusive animals with DNA from hunter-collected faeces: four methods for brown bears. Conservation Biology 19:150-161.

Chaline N, Ratnieks FLW, Raine NE, Badcock NS, Burke T (2004) Non-lethal sampling of honey bee, *Apis mellifera*, DNA using wing tips. Apidologie 3:11-318.

Chu J-H, Lin Y-S, Wu H-Y (2006) Applicability of non-invasive sampling in population genetic study of Taiwanese Macaques (*Macaca cyclopis*). Taiwania 51:258-256.

Creel S, Spong G, Sands JL, Rotella J, Zeigle J, Joe L, Murphy KM, Smith D (2003) Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. Molecular Ecology 12:2003-2009.

Evett IW, Weir BS (1998) Interpreting DNA evidence: Statistical genetics for forensic scientists. Sinauer, Sunderland.

Fernando P, Vidya TNC, Rajapakse C, Dangolla A, Melnick DJ (2003) Reliable non-invasive genotyping: fantasy or reality? Journal of Heredity 94:115-123.

Flagstad O, Hedmark E, Landa A*,* Broseth H, Persson J, Andersen R, Segerstrom P, Ellegren H (2004) Colonization history and noninvasive monitoring of a reestablished wolverine population. Conservation Biology 18:676-688.

Gagneux P, Boesch C, Woodruff DS (1997) Microsatellite scoring errors associated with non-invasive genotyping based on nuclear DNA amplified from shed hair. Molecular Ecology 6:861-868.

Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Molecular Ecology 14:599-612.

Hoyle S, Peel D, Ovenden JR, Broderick D (2005) ShadowBoxer and LocusEater: programs to optimise experimental design and multiplexing strategies for harvest rate estimates of fish populations using genetic tagging. Molecular Ecology Notes 5:974-976.

Hunter M, Broderick D, Ovenden JR, Tucker K, Bonde RK, McGuire PM, Lanyon J (2010) Characterization of highly informative cross-species microsatellite panels for the Australian dugong (*Dugong dugon*) and Florida manatee (*Trichechus manatus latirostris*) including five novel primers. Molecular Ecology Research 10:368-377.

Johnson PCD, Haydon DT (2007) Maximum-likelihood estimation of allelic dropout and false allele error rates from microsatellite genotypes in the absence of reference data. Genetics 175:827-842.

Kalinowski ST, Sawaya MA, Taper ML (2006) Individual identification and distribution of genotypic differences between individuals. Journal of Wildlife Management 70:1148-1150.

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Molecular Ecology 16:1099-1106.

Knapp SM, Craig BA, Waits LP (2009) Incorporating genotyping error into non-invasive DNA-based mark-recapture population estimates. Journal of Wildlife Management 73:598-604.

Leigh GM, Hearn WS, Pollock KH (2006) Time-dependent instantaneous mortality rates from multiple tagging experiments with exact times of release and recovery. Environmental and Ecological Statistics 13:89-108.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. Molecular Ecology 7:639-655.

McKelvey KS, Schwartz MK. (2005) DROPOUT: a program to identify problem loci and samples for noninvasive genetic samples in a capture-mark-recapture framework. Molecular Ecology Notes 5:716-718.

Miller CR, Joyce P, Waits LP (2005) A new method for estimating the size of small populations from genetic mark–recapture data. Molecular Ecology 14:1991-2005.

Mills L, Citta J, Lair K, Schwartz M, Tallmon D (2000) Estimating animal abundance using non-invasive DNA sampling: promise and pitfalls. Ecological Applications 10:283–294.

Morin P, Chambers AKE, Boesch C, Vigilant I (2001) Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytesverus*). Molecular Ecology 10:1835-1844.

Paetkau D (2003) An empirical exploration of data quality in DNA-based population inventories. Molecular Ecology 12:1375-1387.

Paetkau D, Strobeck C (1994) Microsatellite analysis of genetic variation in black bear populations. Molecular Ecology 3:489-495.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. Nature Reviews Genetics 6:847-859.

Rudnick JA, Katzner TE, Bragin EA, DeWoody JA (2008) A noninvasive genetic evaluation of population size, natal philopatry, and roosting behavior of non-breeding eastern imperial eagles (Aquila heliaca) in central Asia. Conservation Genetics 9:667-676.

Seber GAF (1982) The estimation of animal abundance and related parameters, 2nd ed. Griffin, London.

Soulsbury CD, Graziella I, Edwards KJ, Baker PJ, Harris S (2007) Allelic dropout from a high-quality DNA source, Conservation Genetics 87:733-738.

Taberlet S, Griffin B, Goossens S, Questiau S, Manceau V, Escaravage N, Waits LP, Bouvet J (1996) Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Research 26:3189-3194.

Waits LP, Leberg PL (2000) Biases associated with population estimation using molecular tagging. Animal Conservation 3:191-199.

Waits LP, Paetkau D (2005) Noninvasive genetic sampling tools for wildlife biologists: A review of applications and recommendations for accurate data collection. Journal of Wildlife Management 69:1419-1433.

Wang J (2004) Sibship reconstruction from data with typing errors. Genetics 166:1963-1979.

White GC, Burnham KP (1999) Program MARK: Survival estimation from populations of marked animals. Bird Study 46 Supplement:120-138.

Wright JA, Barker RJ, Schofield MR, Frantz AC, Byrom AE, Gleeson DM (2009) Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples. Biometrics 65:833-840.

4.10 Appendix I - Nomenclature

α Type I error rate.

β Type II error rate.

b_{ij} Effective block size between the i^{th} and j^{th} sampling events.

D Total number of genotype samples from all sampling events.

d_i Total number of genotype samples within the i^{th} sampling event.

ε Genotype error rate per locus used in equation 2.

ε' Genotype error rate per locus simulated in data.

E Effective sample size.

e_{ij} Effective sample size between the i^{th} and j^{th} sampling events.

g_a Genotype at a single locus from sample *a* .

g_b Genotype at a single locus from sample *b.*

$I(g_s|g_m)$ Probability of having a genotype identical to the sample genotype (g_s) given the alleged match genotype (g_m).

K Total number of enabled comparisons in *D* by *D* matrix.

k_{ij} Number of enabled comparisons between the i^{th} and j^{th} sampling events.

L Number of loci in genotyping panel.

L^* Number of loci present between two genotypes.

LR Likelihood ratio.

LLR Log likelihood ratio.

LLR_V Log likelihood ratio threshold corresponding to *V* Type I errors.

M Number of pairwise matches in the data with their *LLR* greater than LLR_V.

m_{ij} Number of matches between the i^{th} and j^{th} sampling events with less than v_{ij} Type I errors.

m_0 Number of pairwise comparisons in the reference data that are true negatives.

m_1 Number of pairwise comparisons in the reference data that are true recaptures.

n_1 Number of simulated datasets used to estimate LLR_V.

n_2 Number of simulated datasets used to estimate the standard error of *R*.

n_3 Number of replicated reference data files used to test theory.

N Lincoln-Petersen population estimate.

PIR Probability of individual recapture.

pir_{ij} Probability of individual recapture between the i^{th} and j^{th} sampling events.

$P(g)$ Frequency of genotype (g) in population.

$\hat{R}$ Corrected number of recaptures based on *D* samples (can be a fraction).

r_{ij} Corrected number of recaptures between the *i*th and *j*th sampling events.

$\widetilde{R}$ Number of recaptures used in simulation – measured in whole numbers.

s_{ij} Number of true positives between the *i*th and *j*th sampling events.

S Total number of true positives (‘sample recaptures’).

T Total number of false negatives (Type II errors).

U Total number of true negatives.

V Total number of false positives (Type I errors).

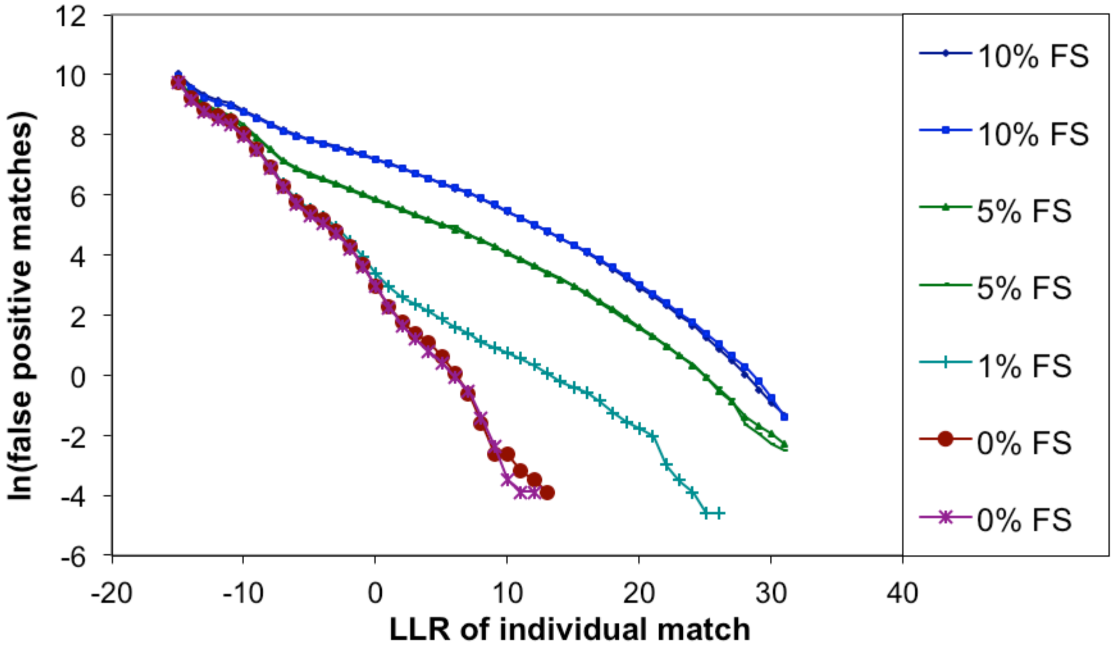
v_{ij} Number of Type I errors between the *i*th and *j*th sampling events.

Y Simulated proportion of loci genotyped.

4.11 Appendix II – Percentage sibship estimation

Extending on the likelihood theory and the estimation of false positive matches some preliminary work to estimate the proportion full sibs in datasets showed promise (Figure 4.6) The method appeared to be a very sensitive method of detecting a low percentage of full-sibs, even in genotype data that was insufficient to determine pedigree assignments of individuals (Wang, 2004). The replicate simulations with 0%, 5% and 10% full-sibs demonstrated the low variability of the plots. This area of research is worthy of further investigation. The results could be implemented within a new version of SHAZA that simultaneously estimates recaptures and sibship relationships.

Figure. 4.6 Log likelihood ratio of an individual match (equation 4.3) plotted against the natural logarithm of the cumulative number of false positive genotype matches when the number of full-sibs (FS) in the data vary between 0% and 10%. Data simulated using 1200 genotypes with 7 loci using Spanish mackerel allele frequencies* with no missing loci.



*For more information on the Spanish mackerel genotype data see, Macbeth et al (2013), http://ncbi.nlm.nih.gov/pubmed/23550119.

Chapter 5

How many fish under the boat? Estimating abundance of narrow-barred Spanish mackerel (*Scomberomorus commerson*) using a genetic mark-recapture approach.

Macbeth G.M., Broderick D., Buckworth R.C., Ovenden J.R., Wang Y-G., (2014) A mark-recapture design using tissue genotypes for estimating abundance of narrow-barred Spanish mackerel (*Scomberomorus commerson*).

5.1 ABSTRACT

Abundance is a key ecological monitoring parameter, and also a highly relevant fisheries management parameter, which is difficult to measure in marine fish populations. To demonstrate a new technique to estimate abundance Narrow-barred Spanish mackerel *Scomberomorus commerson* was used as an example species. Lures with a specially designed hook were used for non-invasive sampling of tissues. Once struck the hook tip contained a small sample of DNA tissue that was genotyped and compared to genotypes of fish landed during the same fishing trip. This simultaneous mark-recapture design was used to estimate the average number of active feeding fish encountered per fishing day which was estimated to be 281 fish with a 95% confidence interval ranging from 187 to 312 fish. The 95% confidence interval for the percentage of fish caught ranged from 11% to 19% with a mean of 17%. Genetic sampling combined with random sampling of fishing transects may be useful in monitoring changes to abundance over time.

5.2 Introduction

Changes to abundance over time is a key measure used in ecological monitoring (Pratchett et al., 2006; Latimore, 2007; Simin et al., 2012) yet abundance estimation of fisheries populations is a difficult task. A fundamental equation used in fisheries modelling to estimate fish abundance (*N*) is given by the relationship *C/E*=*qN* where *C* is the catch, *E* is a unit of fishing effort, *q* is the catchability coefficient defined as the probability of an individual fish being caught by a unit of effort (Pinhorn, 1988; Arreguín-Sánchez, 1996; Maunder et al., 2006). The equation has been applied to tag recovery (Hilborn, 1990) but is more generally used to infer changes to *N*, assuming *C*/*E* is proportional to *N* but this requires *q* to be determined. Perhaps the biggest concern to inferring changes to *N* with *C*/*E* statistics is that the catchability coefficient (*q*) may not be independent of *N* (Harley et al., 2001; Ellis and Wang, 2007; Maunder et al., 2006) so that *C/E* is not proportional to *N*. The problem of estimating abundance (*N*) in fishery populations is that *q* is not known with precision and will not be constant due to various factors including time of day, season of year, La Nina and El Nino events, latitude and longitude, lunar cycles, technology creep, aggregation behaviour of fish and targeting behaviour of fishers (Li et al., 2013; Marchal et al., 2006, Ellis and Wang, 2007). Linear equations have been used to correct *q* to account for different efficiencies of fishing gear (Arreguín-Sánchez, 1996) while generalised linear models (Campbell, 2004) and maximum likelihood methods (Wang, 1999) have also been used estimate *q*. The main point here is that *q* is required to estimate stock abundance but it is a very difficult parameter to determine with any degree of accuracy as it potentially requires many years of catch statistics for assessment (Arreguín-Sánchez, 1996).

This study demonstrated an alternative method of measuring relative changes in fish abundance without the need to directly estimate *q*. Here genetic tagging was used (Buckworth 2004) to sub-sample within a fishing ground an area called a 'fishing zone' in which fish are landed. Sampling along the fishing zone consisted of *in situ* sampling of wild fish using 'genetag' lures (Buckworth, 2004) and genotype sampling of landed fish. By comparing the genotypes collected from the 'genetag' lures and landed fish (Macbeth et al., 2011) within 'fishing zones' it was possible to estimate the abundance of active feeding fish encountered per day of fishing effort (N^*). It is only after estimating N^* that an estimate of catchability within the fishing zone using $q^*=C^*/N^*$ was determined where C^* is the number of fish landed per day of fishing effort. The importance of N^* is that it provides a new measure of abundance which is directly related to *N* by a constant factor (a/A) giving $N^*= Na/A$ where *a* is the area covered by the 'fishing zone' per day and *A* is the total area of the fishery (Arreguín-Sánchez 1996; Ellis and Wang, 2007).

To demonstrate the feasibility of estimating abundance using *in situ* genetic sampling narrow-barred Spanish mackerel, *Scomberomous commerson*, was chosen. It is a large fast-swimming pelagic predator, found throughout tropical and sub-tropical neritic waters of the Indo-West Pacific (Collette and Nauen, 1983). This species within Northern Territory waters appear to be from a single genetic population (Macbeth et al., 2013). The Northern Territory Spanish Mackerel Fishery uses trolled lures or baited lines. Usually taken in depths less than 100 m (McPherson, 1988), *S. commerson* is often associated with reefs and islands, and is targeted in commercial, artisanal and recreational fisheries throughout its range. This fish species attains at a caudal fork length of around 80 cm and age of 2 years (Devaraj, 1993; McPherson, 1993; Mackie et al., 2003).

This large species seems to be suited to genetic tagging methods. The challenges and feasibility of estimating the abundance of active feeding fish encountered per day of fishing effort (N^*) and the proportion of these fish caught within the fishing zone (q^*) are the focus of this study. The hypothesis tested is that N^* and q^* can be obtained with finite confidence limits from real fishing data.

5.3 Methods

5.3.1 Nomenclature

List of terms used:

B	number of days of fishing effort,
C	number of fish caught within the fishing zone,
C^*	number of fish caught per day of fishing effort,
E	unit of fishing effort,
F	number of fin samples genotyped from C subsample,
k	the k^{th} fishing trip,
K	total number of fishing trips,
L	number of lure samples genotyped,
M	number of putative genotype matches,
N	total abundance of population,
N^*	abundance of active feeding fish encountered per day,
q	population catchability coefficient,
q^*	proportion of fish caught within the fishing zone,
$\hat{R}$	corrected number of recaptures,
S	total number of loci in sample,
S^*	number of loci with a putative error,
t	the t^{th} time period,
T	total number of time periods used to solve equation 3,
V	expected number of false positives genotype matches,
W	number of wild feeding fish prior to fishing,
Y	individual estimate of lure to fin recaptures,
Y^*	vector of Y estimates,
Z	individual estimate of lure to lure recaptures,
Z^*	vector of Z estimates,
ε	genotype error rate of lure samples,
ε'	genotype error rate of fin samples,
μ_{LLk}	mean lure to lure matches,
μ_{LFk}	mean lure to fin matches,
σ_{LLk}	standard error of lure to lure matches and
σ_{LFk}	standard error of lure to fin matches.

5.3.2 Sample collection from fins

Fin samples from landed fish that were also tissue-sampled by a gene-tag lure represent ‘recaptures’ in this mark-recapture study. Both lure and fin samples were collected from the same boat and fishing session. During the fishing period (e.g. morning, day etc.) fins were accumulated in a plastic bag on ice and frozen at the end of the period. Some paired fins within bags were split into two separate fins creating two samples from the same individual. The number of fin samples (*F*) was taken from a subsample of the total number of fish caught (*C*).

5.3.3 Sample collection from Lures

Genotypes of lure samples are equivalent to ‘marks’ in a mark-recapture study, which were subsequently ‘recaptured’ when the fish was landed, sampled and found to have a genotype matching a lure. Wild feeding *S. commerson* were genetically tagged non-invasively during commercial fishing operations in northern tropical Australia adjacent to the Northern Territory, from 2003 to 2006 (Buckworth et al., 2012). Two fishing lines were leased from a commercial fishing operator leaving four to six lines for normal operations. The leased lines deployed lures that were each mounted with two ‘genetag’ hooks (Buckworth, 2004). This hook was specially designed with a J-shaped copper tube forming the functional hook. The tip of each hook has a hollow stainless steel tip with an interior barb. When struck, the hook shaft straightens, releasing the fish but leaving a small tissue sample inside the tip. The lure is retrieved as soon as possible after the strike occurs. As the hooks have been straightened, there is little chance of tissue being sampled during additional strikes from different individuals. As the lures were retrieved, hook-tips were removed and placed in separate 80% ethanol vials at 4^{o} C regardless of the visible presence of tissue. The number of lure samples that were genotyped is denoted by (*L*).

5.3.4 Genotype loci and quality control

Detailed genotyping methods are provided in supporting information, *File S1,* of Macbeth et al. (2013). Briefly, DNA was extracted from lure tips using Qiagen DNeasy® tissue kit. The tip was placed in a 1.5ml Eppendorf tube and proteinase K solution was added. The tip was removed after incubation at 50 ^{o}C prior to the column purification of the DNA. Fin samples (~3-5mm^2) were defrosted and washed in Milli-Q water and air-dried. DNA was extracted using Chelex-100 (Walsh et al., 1991) or salting out (Sambrook et al., 1989) methods. A panel of seven polymorphic microsatellite loci were used to genotype tissue from lure tips and fin clips (Macbeth et al., 2013). Partial genotypes scoring at four or fewer loci did not pass the quality control and were discarded because their genotypes had limited statistical power and were likely to be more error prone than those individuals with more complete genotypes (Paetkau, 2003).

5.3.5 Molecular genetic analysis

Allele frequencies of both lure and fin samples were pooled for analysis. Genetic mark-recapture analysis was performed using a specialised program for detecting putative matches in genotype data, SHAZA version 1.00 (Macbeth et al., 2011) with non-target species removed using correspondence analysis (Macbeth et al., 2013). SHAZA estimates the corrected number of genotype matches ($\hat{R}$) by accounting for genotype error rates and missing data. The SHAZA user manual and software (including ANSI C source code) are available from the authors on: http://molecularfisherieslaboratory.com.au/shadow-zone-analysis-software-shaza/.

5.3.6 Determination of genotyping error rate

Genotype error rates per locus were determined separately for lures (ε) and fins (ε') as the small tissue samples from lures may be more prone to DNA degradation than the larger fin tissue samples collected. Two genotypes from the same lure were collected when tissue samples were lodged in the tips of both hooks on the lure. These duplicate samples were used to estimate the error rate per locus from the lure tissue samples assuming the two samples were from a single individual. Error rates per locus in fin samples were determined from the fin pairs that split in two within the same sample bag giving two samples from the same individual. Genotypes of fin pairs matching between sample bags cannot be from the same individual and were known false positives.

Using program SHAZA (Macbeth et al., 2011), putative genotype matches were determined using a prior estimate of the error rate per locus of 1% and by defining a threshold of the number of cumulative false positives in the set of putative matches V=1.0 (default value). The sum of genotype differences between pairs of loci in the list of putative matches (S*) were used to estimate a new error rate per locus using $1-(1-S^*/S)^{0.5}$ where S is the sum of loci pairs from all genotypes within the list of putative matches. The process of estimating error rates was performed iteratively until no new putative matches were found. As described above, error rates in fin and lure samples were determined separately. The putative matches (duplicates) within the fin and lure datasets were removed prior to pooling the genotype data to determine lure to fin matches. The likelihood ratio (*LR*) of a match (Macbeth et al., 2011, equation 2) was modified to account for the different error rate per locus within lure (ε) and fin (ε') samples as

$$LR_{lure\ to\ fin} = (1-\varepsilon)(1-\varepsilon')I(g_a \mid g_b)/P(g_a) + \varepsilon - \varepsilon\varepsilon' + \varepsilon' \qquad (5.1)$$

where $P(g_a)$ is the probability of genotype g_a and $I(g_a \mid g_b)$ equals 1 if genotype g_a is the same as genotype g_b and equals 0 otherwise.

5.3.7 Estimation of putative recaptures

Eight fishing trips retrieved 10 or more struck lures. Using equation 1, the number of lure to fin recaptures was determined using SHAZA (Macbeth et al., 2011). This program minimised the variance of estimated putative recaptures by accounting for false negative and false positive recaptures. The estimated recaptures were simulated 100 times using SHAZA which was set to bootstrap missing locus combinations observed within the fin genotype data and within the lure genotype data (Macbeth et al., 2011). For each fishing trip *k* this procedure gave a mean (μ_{LFk}) and a standard error (σ_{LFk}) of lure to fin recaptures. The same procedure was used to get a mean (μ_{LLk}) and a standard error (σ_{LLk}) of lure to lure recaptures in each trip *t*. The standard errors σ_{LFk} and σ_{LLk} represent the accuracy of detecting recaptures from the available genotype data and did not include random sampling components of error which would be added when pooling recaptures from multiple fishing trips.

5.3.8 Pooling recaptures from multiple fishing trips

The general formula for calculating the probability of having *m* recaptures across a total of *K* fishing trips is given by:

$$P(m) = \sum_{i=1}^{I} \prod_{k=1}^{K} dbinom(x_{ik}, size_k, prob_k) \qquad (5.2)$$

where *I* is the number of combinations of x_{ik} satisfying $m = \sum_{k=1}^{K} x_{ik}$, *dbinom* is the binomial probability function in R (R Development Core Team, 2008), x_{ik} is the number of recaptures within combination *i* and trip *k*, *size*$_k$ is the sample size within trip *k* and *prob*$_k$ is the probability of a recapture within trip *k*. For example if *m*=1 and *K*=2 then there are only two possible combinations (*I*=2) with {x_{11}=0, x_{12}=1} and {x_{21}=1, x_{22}=0}.

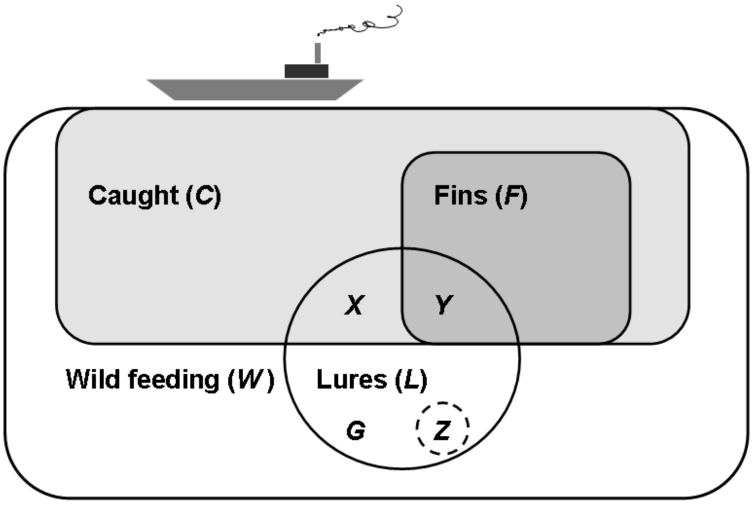
The binomial distribution measured the sampling error, and to include the error in genotype assignment of lure to fin recaptures, *prob*$_k$ was sampled from the normal distribution as $prob_k \sim N(\mu_{LFk}, \sigma_{LFk}) / size_k$ where the number of pairwise comparisons $size_{k=}n_{Lk}.n_{Fk}$ with n_{Fk} equal to the number of fin genotypes and (n_{Lk}) number of lure genotypes within trip *k*. Each *P*(*m*) value was therefore not unique and average values were taken from 1000 iterations. Over all *m* values, the area under the average *P*(*m*) curve approximated unity. By drawing a random number between zero and one, the cumulative distribution under the average lure to fin *P*(*m*) curve was used to sample 1000 estimates of the number of lure to fin recaptures and stored in vector *Y**.

Similarly, when including the error in the genotype assignment of lure to lure recaptures, 1000 iterations of *P*(*m*) values were determined by sampling $prob_k \sim N(\mu_{LLk}, \sigma_{LLk}) / size_k$ where $size_k$ = n_{Lk} (n_{Lk} -1) as lures were not compared to themselves. By drawing a random number between zero and one, the cumulative distribution under the average lure to lure *P*(*m*) curve was used to sample 1000 estimates of the number of lure to lure recaptures and stored in vector *Z**. The distribution of estimates in vectors *Y** and *Z** contain both the random sampling error and genotype error components.

5.3.9 Feeding and capture model

The fishing model in the form of a subset diagram is shown in Figure 5.1. The number of fish caught (*C*), number of caught fish with fins genotyped (*F*) and the number of lures genotyped (*L*) were determined directly from sample collection numbers. The number of recaptures *Y* and *Z* were sampled in sequence from vectors *Y** and *Z**.

Figure 5.1. Feeding and capture model. Relationships shown are total feeding (*W*) within the fishing zone, number caught (*C*), captured fish with fins genotyped (*F*) and the number of lures genotyped (*L*). Genetag fish are partitioned between those caught without a fin sample (*X*), those lure to fin recaptures (*Y*) and those lures remaining at large (*G*) with (*Z*) equal to the number of lure to lure recaptures.



The number of wild feeding fish within the 'fishing zone' prior to being caught (*W*) is of most interest. Assuming lures are deployed prior to fins being sampled, then *W* could be estimated using: *W=L.F/Y.* This equation is equivalent to mark-recapture using the Petersen Method (Seber 1982) where the number marked is equivalent to *L*, the number captured is *F* and the number of animals with a mark that were captured is *Y*. The Petersen Method was modified to account for simultaneous lure deployment and captures. The strike rate of 'genetag' lures was assumed to be the same as commercial lures used to land fish on board. Using this assumption, *W* was estimated using equation (3) which uses information from both lure to fin recaptures as well as lure to lure recaptures. Solutions for *W* for each fishing trip were determined by iteration by arbitrarily subdividing the fishing trips into *T*=100 time periods which was sufficiently large to achieve convergence. The number of active feeders encountered per day of fishing effort (*N**) was determined by the sum of *W* over all fishing trips divided by the total number of days of fishing effort. The derivation of equation (3) is detailed in the appendix with solutions for *C*/*W* estimating the proportion of active feeding fish landed within the 'fishing zone' (*q**= *C*/*W*=*C**/*N**) where *C** is the average catch per day of fishing effort.

$$\frac{Y+Z}{F+L} \approx \frac{1}{T}\sum_{t=1}^{T}\left[\left(\frac{t.L}{T} - \sum_{t'=0}^{t-1}\left(\frac{C}{T}P_{t'}\right)\right) \Big/ \left(W - \frac{tC}{T}\right)\right] \qquad (5.3)$$

5.4 Results

5.4.1 Genotype error rate from fin samples

Genotype error rates were determined from 7831 fin genotypes having more than four of the possible seven loci. Among these, a total of 113 putative matches were detected by SHAZA when setting two parameters (i) a cumulative false positive rate of less than *V*=1.0 pairwise matches amongst them, and (ii) a prior error rate per locus of ε'=1.0%. Of the 113 putative matches 70, 17, 18 and 8 were from genotypes with 7, 6, 5 and 4 loci respectively giving *S*=(70x7+17x6+18x5+8x4)=714 loci in total. There were *S**=19 loci having at least one allele different between the loci pairs of each fin to fin match. The probability of an error in fin to fin genotype matches was determined as $\hat{\varepsilon}'$=1-(1-*S**/*S*)$^{0.5}$=(1-(1-19/714)$^{0.5}$=1.3%. The SHAZA analysis were re-run with $\varepsilon' = \hat{\varepsilon}'$ with no difference in putative matches detected. There is a high degree of confidence that the error rate was detected from paired genotypes from the same individual, as all of the 113 putative paired matches detected by SHAZA were within the same bag, with the first known false positive (between bag match) occurring at the 128th highest likelihood ranked match.

5.4.2 Genotype error rate from lure samples

Genotype error rates were determined from 664 lure samples that were genotyped with more than 4 loci. There were 59 putative matches detected by SHAZA when setting the false positive threshold level to *V*= 0.01 match and using a prior error rate per locus of ε=1.0%. When increasing the total false positive threshold to *V*= 1.0, twenty more putative matches were discovered giving a total of 79 matches. By chance, as defined by *V*=1.0, one of these 79 matches was expected to be a false positive although all the 79 matches were used to estimate the error rate. From these matches there were 31 differences between the 482 loci pairs giving an error rate per locus of $\hat{\varepsilon}$=3.3%. The analysis was re-run with ε=3.3% ('-e 0.033' option in SHAZA) resulting in two additional putative matches found giving a new error rate ε=3.8% per locus. An additional SHAZA run with '-e 0.038' revealed no difference in matches. A chi-squared analysis indicated that the lure error rate was significantly higher than the fin sample rate (P<0.01).

Table 5.1. Recapture estimates from eight fishing trips. Number caught (*C*), number of lures sampled (*L*) and number of fins sampled (*F*) during fishing trips lasting more than one day together with recapture estimates ($\hat{R}$) between lure and fin samples with standard error of genotype assignment.

Fishing trip	Days of fishing effort (B)	Sample numbers			Recaptures ± genotype error	
		Caught (*C*)	Lures (*L*)	Fins (*F*)	Lure to Fin (*Y*)	Lure to Lure (*Z*)
1	4	105	44	74	0.0	2.4±0.7
2	7	297	87	206	3.2±0.6	1.0±0.4
3	6	136	30	88	1.1±0.4	0.0
4	7	452	29	276	5.2±0.6	0.0
5	4	281	68	244	4.1±0.5	1.0±0.2
6	10	248	38	208	1.1±0.2	0.0
7	7	566	11	279	1.1±0.4	0.0
8	6	306	18	21	0.0	0.0
Total	51	2391	325	1396	15.8	4.4

5.4.3 Increased pairwise comparisons using SHAZA

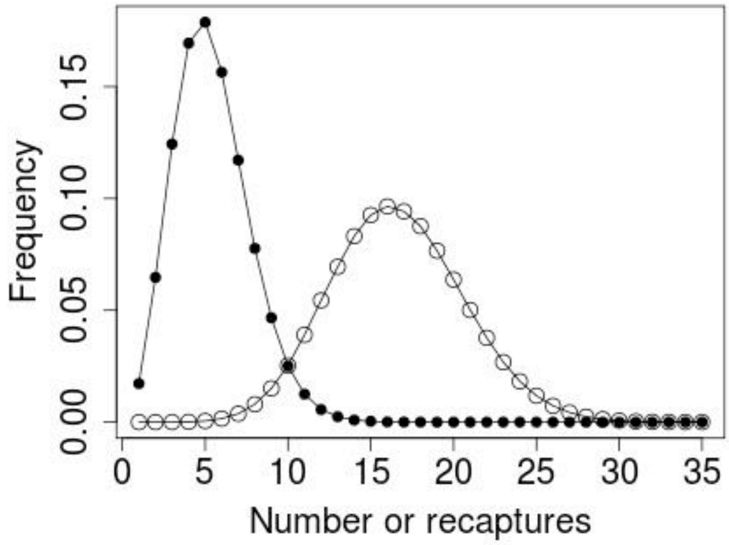
There were 5560 fin genotypes and 382 lure genotypes that had all 7 loci detected. Instead of discarding partial genotypes, the program SHAZA was capable of selecting individual pairwise comparisons while managing a constraint on false positives. The use of SHAZA in this way increased the total number of paired genotype comparisons by 80% compared to discarding lure and fin genotypes with missing loci.

5.4.4 Estimation of capture rates from fishing trips

Recapture rates were estimated from tagged fish from lure genotypes and from those fish subsequently caught and genotyped from fin samples. There were eight fishing trips in which at least ten lures were struck. No recaptures were observed between fishing trips. Listed in Table 5.1 for each fishing trip are recapture estimates ($\hat{R}$) together with the number of days fishing effort (*B*), number of fish caught (*C*), number of lures sampled (*L*) and number of fins sampled (*F*). Six of these trips had fin to lure recaptures. Three trips had lure to lure recaptures indicating the same fish had struck more than one lure. In fishing trip eight there were 306 fish caught but no recaptures detected. The number of recaptures were not determined with 100% accuracy with the standard error of these estimates, up to ±0.7, reflecting the accuracy with which genotype matches could be estimated. The estimates in Table 5.1 are a fraction higher than whole numbers due to the net correction made by SHAZA for false positives and false negatives.

The estimated recaptures from Table 5.1 were pooled across all trips using equation 5.2 which included the variance associated with both (i) the accuracy of genotype matching and (ii) the binomial random sampling error of lure to fin and lure to lure match estimates (Figure 5.2).

Figure 5.2. Pooled trip estimates of recaptures. The two distributions show fin to lure recaptures (open circle) and lure to lure recaptures (solid dot) from vectors *Y** and *Z**.



The feeding and capture model was applied using equation 5.3 to estimate the size of the pooled feeding aggregates (*W*) using total captured (*C*), total number of lures sampled (*L*) and total number of fins sampled (*F*) tabulated for each trip in Table 5.1. Lure to fin recaptures (*Y*) and lure to lure recaptures (*Z*) were determined by iterative random sampling. Their frequency distribution is shown in Figure 5.2. The average number of active feeding fish divided by the total number of days fishing effort is illustrated in Figure 5.3. The upper and lower 95% confidence interval for active feeding fish per day of fishing effort ranged from 187 fish to 312 fish with a mean of 281 fish. The average percentage of active feeders caught was 17% with the 95% confidence intervals between 11% and 19% (Figure 5.4).

Figure 5.3. Number of active feeders encountered per day of fishing effort (*N**). Distribution of *N** determined from equation 4 by sampling fin to lure recaptures (*Y*) and lure to lure recaptures (*Z*) from 1000 estimates in vectors *Y** and *Z**.

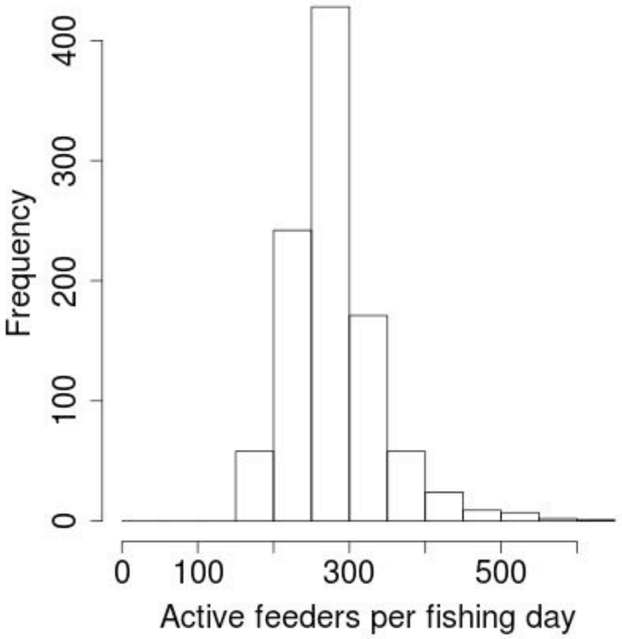
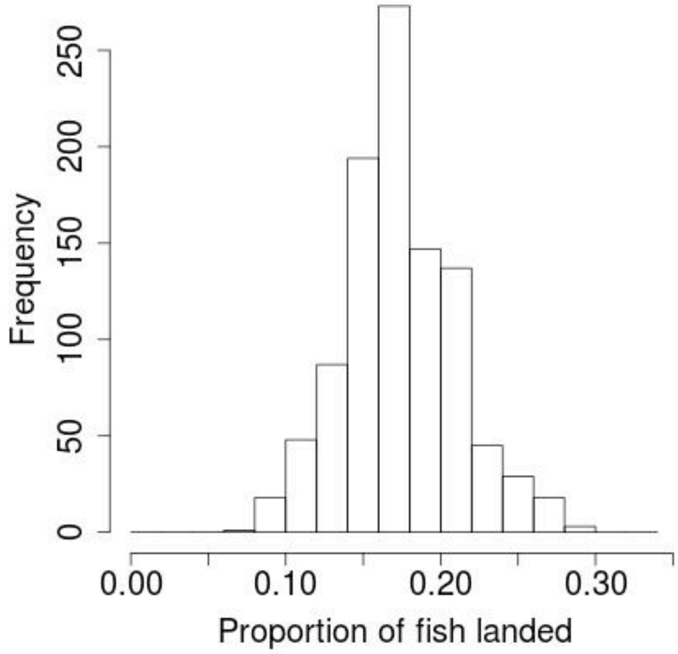


Figure 5.4. Proportion of fish landed (*q**) that were actively feeding. The distribution showing the precision of estimation by accounting for both genotype and random sampling errors from 1000 iterations.



5.5 Discussion

This study demonstrated that it was possible to obtain finite estimates of the number of active feeding fish encountered per day of fishing effort (*N**) and the proportion of fish that were landed (*q**).

In practice there is an instantaneous fishing zone which is constantly changing with movement of the fishing vessel. It is assumed mixing of tagged fish within the instantaneous fishing zone is sufficient so that the proportion of tagged fish landed is a representative random sample of both tagged and untagged fish within the fishing zone. The model can give a consistent estimate for *W* even if the fish are not in equal densities throughout the day. It does not matter what proportion of fish are caught in each period as the pooled estimate of *W* will be the same. It is also assumed that marking does not affect catchability and there is some evidence from duplicate lure samples that fish are not hesitant to strike again after being tagged. Genetic tags are fixed and not lost between sampling occasions however the accuracy of identifying genetic matches was not 100% leading to an increase in the standard error of parameter estimates.

Compared to estimates of changing abundance inferred from catch per unit of effort data (C/E), the abundance measure (*N**) is a more reliable and responsive estimate to changes in stock abundance. For example, C/E is determined using catch numbers which will increase with improvements in fishing technology causing an upward bias in abundance estimates (Marchal et al., 2006). Improved fishing methods will not increase the number of active feeders within the 'fishing zone' prior to harvest, which *N** measures, and is therefore not susceptible to many forms of technology creep. Without random sampling of fishing zones the method would be prone to technology creep that targets higher fishing density and, as a consequence will increase the abundance estimate.

The ability to detect the instantaneous proportion of stock removed (*q**) within 'fishing zones' is also an advantage over traditional C/E methods and may give early indications of depleted abundance over time. For example, it is feasible that C/E estimates remain stable over recording periods while *q** increases over the same periods. An upward trend in *q** may in this case provide an early warning signal prior to fishing grounds experiencing lower C/E values. If the fishery experiences lower C/E, fishers may move to new grounds to maintain C/E and profitability (Rose and Kulka, 1999) and in doing so mask true population decline.

Recent assessments of the Northern Territory stock of narrow-barred Spanish mackerel, based on C/E data, indicate that the fishery has recovered from heavy fishing (drift netting) by a distant water fleet during the 1970s and 1980s, and is now fished at sustainable levels (Buckworth, 2004; Grubert et al., 2013). The current fishery harvest rate on the Northern Territory fishery grounds is estimated as 21% (Grubert et al., 2013). This reported harvest rate seems high given the upper 95% confidence limit of *q** is 19% and that the total fishing zones covered by the fishing fleet each year are likely to be a small fraction of the total fishery ground. A possible explanation for the difference would be that the concentration of harvesting occurs in fish schools, which congregate in close proximity to reefs and islands (McPherson, 1988), and that they are targeted multiple times per year.

Given that the annual catch of *S. commerson* in the Northern Territory may be as few as 30 000 to 60 000 fish (Buckworth et al., 2012) and the estimate of *q** was 17%, then a very rough estimate of the wild population over all the ‘fishing zones’ covered annually by the fishing fleet would be (30 000+ 60 000)/(2*0.17)= 265 000. The population density (*N*/a*) could also be estimated with *a* equal to the width of the ‘fishing zone’ multiplied by the distance travelled during the eight fishing trips. The challenge in this case would be to estimate the width of the fishing zone.

As a result of this pilot study, it is recognised that there are a number of changes that could improve the efficiency of estimating *N** and *q**:

(i) genotyping a higher proportion of fins from landed fish, which would increase the chance of finding more lure to fin recaptures,

(ii) implementing a larger number of fishing lines with 'genetag' lures per fishing vessel, which would increase the number of lure to lure and lure to fin recaptures,

(iii) record the number and time of commercial fishing lines deployed perhaps using video recording to refine the time-series of recaptures,

(iv) improve the estimate of the percentage of fish caught by correcting for the number of lines deployed for 'genetag' lures that replaced the number of commercial lines available,

(v) record more loci to the panel to increase the ability to recover all pairwise comparisons and improve on the 80% recovered in this study using SHAZA (Macbeth et al., 2011),

(vi) monitor the speed of the fishing vessel during active fishing periods which may be useful in estimating fish density,

(vii) use random sampling of ‘fishing zones’ (Ellis and Wang, 2007) using a dedicated monitoring vessel and

(viii) estimate *N** and *q** at annual intervals.

In this study the three years of fishing trips were combined to estimate *N** and *q**. However, with the efficiencies listed above it should be feasible to determine these measures annually to be able to detect and respond in a timely manner to any significant time series change that may occur. This will in turn improve the ability to manage maximum sustainable yield.

The increased error rates in lure genotypes compared to fins is thought to be related to the smaller amounts of tissue sampled from lures; they are likely to degrade more quickly than the larger fin samples. Poor DNA quality is generally more error prone than that of high quality DNA (Creel et al., 2003; Taberlet et al.,1996). However, the genotype error rates of up to 3.8% in this study compared favourably to error rates as high as 21-57% in Soulsbury et al. (2007) and 31% in Gagneux et al. (1997). As indicated with simulated data (Macbeth et al., 2011) the estimates of the numbers recaptured are somewhat robust to a deviation from the true error rate. As in this study it is not uncommon for non-invasive sampling of genotypes to result in a large proportion of genotypes with missing loci. It was shown that the use of SHAZA was critical for the success of this study. No longer is it required to discard genotypes with missing loci as has been suggested (Creel et al., 2003).

In summary, the number of active feeding fish encountered per day of fishing effort within the fishing zone (*N**) and the percentage of those active feeding fish landed (*q**) provide a baseline for future ecological comparisons measuring changes due to naturally occurring events or due to exploitation pressure. The methods deployed are unique and introduce a new tool for monitoring line-caught commercial fish populations.

5.6 Acknowledgements

Raewyn Street and Marcus McHale provided technical assistance in the genetics laboratory. Charles Bryce, Adrian Donati and other Northern Territory Fisheries staff assisted in sample collection. Simon Hoyle, Paige Everson, Ann Preece and Dr George Leigh provided valuable input at various stages throughout this study. The support of the Northern Territory Spanish Mackerel Fishermen’s Association members, especially Norm Hedditch, is gratefully acknowledged and Prof. Carl J. Walters (Univ. British Columbia) is thanked for his encouragement of this work. Funding was provided by the Fisheries Research and Development Corporation (Project No. 2002/011), the Queensland and Northern Territory Governments and the University of Queensland.

5.7 References

Arreguín-Sánchez F (1996) Catchability: a key parameter for fish stock assessment. Reviews in fFsh Biology and Fisheries 6:221-242.

Buckworth RC (2004) Effects of Spatial Stock Structure and Effort Dynamics on the Performance of Alternative Assessment Procedures for the Fisheries of Northern Australia. PhD Thesis, Univ. of British Columbia, 226 p.

Buckworth RC, Ovenden JR, Broderick D, Macbeth GM, McPherson GR, Phelan MJ (2012) GENETAG: Genetic mark-recapture for real-time rate monitoring: Pilot studies in northern Australia Spanish Mackerel fisheries. Northern Territory Government, Australia. Fishery Report No. 107.

Campbell RA (2004) CPUE standardisation and the construction of indices of stock abundance in a spatially varying fishery using general linear models. Fisheries Research 70:209-227.

Collette BB, Nauen CE (1983) FAO Species Catalogue. Vol.2. Scombrids of the world. An annotated and illustrated catalogue of tunas, mackerels, bonitos and related species known to date. FAO Fisheries Synopsis 125, 137 p.

Creel S, Spong G, Sands JL, Rotella J, Zeigle J, Joe L, Murphy KM, Smith D (2003) Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. Molecular Ecology 12:2003-2009. doi:10.1046/j.1365-294X.2003.01868.x

Devaraj M (1993) Contributions to the taxonomy, biology and production of certain Indian fish stocks. DSc. Thesis, Madurai Kamaraj University.

Ellis N, Wang Y-G (2007) Effects of fish density distribution and effort distribution on catchability. ICES Journal of Marine Science 64:178-191.

Gagneux P, Boesch C, Woodruff DS (1997) Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. Molecular Ecology 6:861-868.

Grubert MA, Saunders TM, Martin JM, Lee HS, Walters CJ (2013) Stock assessments of selected Northern Territory fishes. Department of Primary Industries and Fisheries, Northern Territory, Fishery Report No 110. ISBN: 978-0-7245-475-8.

Harley SJ, Myers RA, Dunn A (2001) Is catch-per-unit-effort proportional to abundance? Canadian Journal of Fisheries and Aquatic Sciences 58:1760-1772.

Hilborn R (1990) Determination of fish movement patterns from tag recoveries using maximum likelihood estimators. Canadian Journal of Fisheries and Aquatic Sciences 47:635-643.

Latimore J (2007) The vigor-organisation-resilience concept of ecological health: lessons from temperate warmwater stream fish communities. American Fisheries Society Symposium 49:587-593.

Li G, Zou X, Chen X, Zhou Y, Zhang M (2013) Standardisation of CPUE for Chilean jack mackerel (*Trachurus murphyi*) from Chinese trawl fleets in the high seas of the Southeast Pacific Ocean. Journal of Ocean University of China 12:441-451.

Macbeth GM, Broderick D, Buckworth RC, Ovenden JR (2013) Linkage disequilibrium estimation of effective population size with immigrants from divergent populations: A case study on Spanish mackerel (Scomberomous commerson) Genes Genomics Genetics http://ncbi.nlm.nih.gov/pubmed/23550119

Macbeth GM, Broderick D, Ovenden JR, Buckworth RC (2011) Likelihood-based genetic mark–recapture estimates when genotype samples are incomplete and contain typing errors. Theoretical Population Biology 80:185-196. doi:10.1016/j.tpb.2011.06.006

Mackie M, Lewis PD, Gaughan DJ, Buckworth RC (2003) Stock assessment of Spanish mackerel (*Scomberomorus commerson*) in Western Australia. Final Report, Fisheries Research and Development Corporation Project 1999/151. Fisheries Department of Western, Australia. Fishery Report No. 88.

Maunder MN, Sibert JR, Fonteneau A, Hampton J, Kleiber P, Harley SJ (2006) Interpreting catch per unit effort to assess the status of individual stocks and communities. ICES Journal of Marine Science 63:1373-1385.

Marchal P, Andersen B, Caillart B, Eigaard GO, Hovgaard H, Iriondo A, Le Fur F, Sacchi J, Santurtun M (2006) Impacts of technological creep on fishing effort and fishing mortality, for a selection of European fleets. ICES Journal of Marine Science 64:192-209.

McPherson GR (1988) A review of large coastal pelagic fishes in the south pacific region, with special reference to *Scomberomorus commerson* in north-east Australian waters. Workshop in Pacific inshore fishery resources, Noumea, New Caledonia, 14-25 March 1988.

McPherson GR (1993) Reproductive biology of the narrow barred spanish mackerel (*Scomberomorus commerson* Lacepède, 1800) in Queensland waters. Asian Fisheries Science 6:169-182

Paetkau D (2003) An empirical exploration of data quality in DNA-based population inventories. Molecular Ecology 12:1375–1387.

Pinhorn AT (1988) Catchability in some of the major groundfish fisheries off the east coast of Canada. Journal of Northwest Atlantic Fishery Science 8:15-23.

Pratchett MS, Wilson SK, Baird AH (2006) Declines in the abundance of Chaetodon butterfly fishes following extensive coral depletion. Journal Fish Biology 69:1269-1280.

R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rose GA, Kulka DW (1999) Hyperaggregation of fish and fisheries: how catch-per-unit-effort increased as the northern cod (*Gadus norhua*) declined. Canadian Journal of Fisheries and Aquatic Sciences 56:118-127.

Simin D-M, Bagher NSM, Jasem G-M, Najmeh J, Emad K (2012) Application of abundance biomass curve in ecological health assessment of khure-mussa (Northwest of the Persian Gulf). Marine Science 3:1-9.

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular cloning: A laboratory manual*, 2nd edn. Cold Spring Harbour Laboratory Press, Cold Spring, New York.

Seber GAF (1982) The estimation of animal abundance and related parameters, 2nd ed. Griffin, London.

Soulsbury CD, Graziella I, Edwards KJ, Baker PJ, Harris S (2007) Allelic dropout from a high-quality DNA source, Conservation Genetics 8:733-738.

Taberlet P, Griffin S, Goossens B, Questiau S, Manceau V, Escaravage N, Waits LP, Bouvet J (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, 24:3189–3194. doi:10.1093/nar/24.16.3189

Walsh P, Metzger D, Higuchi R (1991) Chelex 100 as a medium for simple extraction of DNA for PCR-based typing forensic material. BioTechniques 10:506-513.

Wang Y-G (1999) A maximum likelihood method for estimating natural mortality and cachability coefficient from catch-and-effort data. Marine and Freshwater Research 50:307-311.

5.8 Author contributions

Michael Macbeth developed the mathematical theory and wrote the manuscript, all authors contributed to editing the manuscript, Damien Broderick laboratory genotyping, Dr Rik Buckworth sample collection, Professor You-Gan Wang statistical advice and Dr Rik Buckworth, Associate Professor Jenny Ovenden, Michael Macbeth and Damien Broderick designed the experiment and supervised the collection of the dataset.

5.9 Data accessibility

Raw data in the form of multi-loci genotypes to individual fish are available at "http://era.deedi.qld.gov.au/3574".

5.10 Appendix – estimating wild feeding school

Estimating wild feeding school (*W*) from fin samples and 'genetag' lures.

The proportion of lures captured can be expressed in terms of total number captured or as a subsample from fins DNA sampled:

$$(X+Y)/C = Y/F \tag{i}$$

where during a fishing session of duration *T* there are: *C* caught fish with a subsample of *F* fish with fins genotyped*, Y* captured fish with both a fin genotype and a lure genotype and, *X* captured fish with a lure genotype but not a fin genotype sampled. Given the number of lures below the boat is changing over time the ratio in (i) can be expressed as:

$$\frac{Y}{F} \approx \frac{1}{T}\sum_{t=1}^{T} P_t \tag{ii}$$

where P_t is the proportion of active feeding fish with a 'genetag' lure sampled at time *t.* The number of active feeders at time *t*, $(W - C_t)$, is determined from the number of wild feeders before , (*W*), less the number of captued fish at time *t*, (C_t), giving:

P_t = [(lures struck up to time *t*)-(fish with lures caught up to time *t*-1)]/(feeding fish at time *t*)

$$= \left(L_t - \sum_{t^*=0}^{t-1}\left(\frac{C}{T}P_{t^*}\right)\right)\Bigg/(W - C_t) = \left(\frac{t.L}{T} - \sum_{t^*=0}^{t-1}\left(\frac{C}{T}P_{t^*}\right)\right)\Bigg/(W - \frac{tC}{T}) \tag{iii}$$

where $C_t = t.C/T$ is the number of fish caught up to time *t*, $L_t = t.L/T$ is the number of lures struck up to time *t* and *L* is the total number of lures struck during a fishing period of *T* time units. By substituting equation (iii) in equation (ii) it follows that:

$$\frac{Y}{F} \approx \frac{1}{T}\sum_{t=1}^{T}\left[\left(\frac{t.L}{T} - \sum_{t^*=0}^{t-1}\left(\frac{C}{T}P_{t^*}\right)\right)\Big/(W - \frac{tC}{T})\right] \qquad \text{(iv)}$$

Using *C*, *L* and the ratio *Y/F* the size of the feeding school (*W*) can then be determined from equation (iv) which is solved by iteration using *T* increments in time.

Equation (iv) was developed further to include additional information from the number of lure by lure recaptures during a fishing session (*Z*). Assuming random sampling the expectation is that $Z/L \approx Y/F$ with a combined estimate of these ratios giving $(Y+Z)/(F+L)$ which from equation (iv) leads to:

$$\frac{Y+Z}{F+L} \approx \frac{1}{T}\sum_{t=1}^{T}\left[\left(\frac{t.L}{T} - \sum_{t'=0}^{t-1}\left(\frac{C}{T}P_{t'}\right)\right)\Big/(W - \frac{tC}{T})\right] \qquad \text{(v)}$$

The size of the feeding school (*W*) can then be determined as above by iteration using *T* increments in time.

Proof that $Z/L \approx Y/F$ is as follows: The expectation of *Z* is the product of lures sampled at time *t*, $(L_t - L_{t-1})$, by the proportion of lures in active feeders sampled at time *t*, (P_t), summed over time *T* giving:

$$Z = \sum_{t=1}^{T}(L_t - L_{t-1}).P_t = \sum_{t=1}^{T}\left(\frac{t.L}{T} - \frac{(t-1)L}{T}\right).P_t$$ which divided by *L* becomes

$$\frac{Z}{L} \approx \sum_{t=1}^{T}\left(\frac{t}{T} - \frac{(t-1)}{T}\right).P_t = \frac{1}{T}\sum_{t=1}^{T}P_t \approx \frac{Y}{F}$$ from equation (ii).

Chapter 6

Linkage disequilibrium estimation of effective population size in Spanish mackerel (*Scomberomorus commerson*) with immigrants from divergent populations.

Macbeth G.M., D Broderick, Buckworth R.C, Ovenden J.R. (2012) Linkage disequilibrium estimation of effective population size in Spanish mackerel (*Scomberomorus commerson*) with immigrants from divergent populations. Genes Genomes and Genetics http://ncbi.nlm.nih.gov/pubmed/23550119

6.1 ABSTRACT

Estimates of genetic effective population size (*Ne*) using molecular markers are a potentially useful tool for the management of endangered through to commercial species. But pitfalls are predicted when the effective size is large as estimates require large numbers of samples from wild populations for statistical validity. Simulations showed that linkage disequilibrium estimates of *Ne* up to 10,000 with finite confidence limits can be achieved with sample sizes around 5000. This was deduced from empirical allele frequencies of seven polymorphic microsatellite loci in a commercially harvested fisheries species, the narrow barred Spanish mackerel (*Scomberomorus commerson*). As expected, the smallest standard deviation of *Ne* estimates occurred when low frequency alleles were excluded. Additional simulations indicated that the linkage disequilibrium method was sensitive to small numbers of genotypes from cryptic species or conspecific immigrants. A correspondence analysis algorithm was developed to detect and remove outlier genotypes that could possibly be inadvertently sampled from cryptic species or non-breeding immigrants from genetically separate populations. Simulations demonstrated the value of this approach in Spanish mackerel data. When putative immigrants were removed from the empirical data, 95% of the *Ne* estimates from jackknife resampling were above 24,000.

6.2 Introduction

The effective number in a breeding stock was defined by Wright (1930) as an idealised number of parents in a population that cause a given level of inbreeding, or given change in allele frequencies. This effective number “is much smaller as a rule than the actual number of adult individuals” (Wright, 1930) but is an important parameter in ecological studies as any change over time indicates underlying changes in population structure. The mean squared correlation of alleles at different loci is a measure of linkage disequilibrium which can be used to estimate genetic effective population size (*Ne*) of diploid individuals. In small populations there is a higher correlation of alleles between loci compared to larger populations (Hedgecock et al., 2007; Pudovkin et al., 1996; Zhdanova and Pudovkin, 2008) and hence there is a relationship with genetic effective population size (Waples, 2006). It was suggested by Waples and Do (2010) that strong assortative mating would lead to biases in $\hat{Ne}$. Later, Waples and England (2011) investigated migration between populations and concluded that the linkage disequilibrium method was robust to equilibrium migration with $\hat{N}e$ reflecting that of the local subpopulation. Waples and England (2011) also showed that pulse migration of strongly divergent individuals was found to depress estimates of local *Ne*.

The effect of pulse migration is an important finding, as related factors could also lead to depressed *Ne* estimates. These factors could include inadvertent sampling of non-target species and sampling of the same species but from populations that have become genetically divergent over many generations. Some fish species are known to exhibit natal philopatry where individuals have home spawning grounds, but later disperse. Examples include herring, cod, sharks, swordfish and anadromous salmonids (Bekkevold et al., 2007; Svedang et al., 2007; Jorgensen et al., 2009; Smith and Alvarado-Bremer, 2010; Beacham et al., 2005). Under this model, samples from a single location taken when the species was in the dispersed phase could represent several genetically distinct (i.e. mixed) stocks. The samples would not represent a panmictic population causing deviations from the expected linkage disequilibria and a bias in the linkage disequilibrium estimation of *Ne*. For example, a downward bias in *Ne* estimates ($\hat{N}e$) was simulated by Palstra and Ruzzante (2011) when divergent populations were pooled.

The frequency of natal philopatry is poorly known across marine species and virtually unknown in Australian fisheries species (Tillet et al., 2012; Blower et al., 2012). A species of considerable fisheries interest in Australia, and much of the Indo-West Pacific, is the narrow-barred Spanish mackerel, *Scomberomorus commerson*. It is a large, fast-swimming pelagic predator found throughout tropical and sub-tropical neritic waters of the Indo-West Pacific (Collette and Nauen, 1983). If *S. commerson* exhibit natal philopatry, the mixing of genetically distinct populations within the sample collection area could depress $\hat{N}e$ in a similar manner suggested by pulse migration (Waples and England, 2011). Seasonal aggregation for spawning followed by dispersal is supported by several lines of evidence;

(a) seasonal variations in the availability of *S. commerson* (Buckworth et al., 2007),

(b) a tag release study in northern Australia showing dispersal of recaptured fish with 12% over 600 nautical miles away (Buckworth et al., 2007),

(c) movement of fish on the eastern Australian coast southwards in summer presumably for feeding (McPherson, 1988) and

(d) multiple genetically distinct stocks in south-east Asia (Fauvelot and Borsa, 2011).

The species is under active management throughout its range in Australia and accurate estimates of effective population size have the potential to assist (Hare et al., 2011; Luikart et al., 2010; Ovenden et al., 2007; Palstra and Ruzzante, 2008).

This paper documents a case study of the pitfalls associated with the estimation of *Ne* in *S. commerson* when large samples of genotypes (*S*>5000) were taken from a single location in northern Australia. The estimated *Ne* determined from empirical data was compared from simulated populations. The estimates of *Ne* were critically reviewed by testing hypotheses that the sampled population is a mixed stock. In addition a method of screening and removing individuals likely to be from non-target populations or species was developed.

6.3 Methods

6.3.1 Linkage disequilibrium estimation of effective population size ($\hat{Ne}$)

Linkage disequilibrium estimation of effective population size is derived from the correlation of alleles between loci. The correlation is determined from allele frequencies and has the general form of the phi correlation coefficient

$$\hat{r}_{A_jB_k} = \frac{\hat{\Delta}_{A_jB_k}}{\sqrt{[\hat{p}_{A_j}(1-\hat{p}_{A_j})+\hat{D}_{A_j}][\hat{p}_{B_k}(1-\hat{p}_{B_k})+\hat{D}_{B_k}]}} \quad \text{(Weir, 1996; p137)}$$

where $\hat{r}_{A_jB_k}$ is the estimated correlation between the *j*th allele in locus *A* and *k*th allele in locus *B* given $\hat{p}_{A_j}$ is the empirical frequency estimation of allele *j* in locus *A*, $\hat{p}_{B_k}$ is the empirical frequency estimation of allele *k* in locus *B,* $\hat{D}_{A_j} = f(A_jA_j) - \hat{p}_{A_j}{}^2$ and $\hat{D}_{B_k} = f(B_kB_k) - \hat{p}_{B_k}{}^2$ represent the additional variance in allele frequencies due to deviations in Hardy Weinberg equilibrium where $f()$ in the above formulae denote the observed homozygote frequencies. When diploid genotypes are sampled the gametic phase is unknown with linkage disequilibrium determined by the Burrows estimate $\hat{\Delta}_{A_jB_k} = \hat{p}(A_jB_k) - 2\hat{p}_{A_j}\hat{p}_{B_k}$ (Schaid, 2004). In this equation $\hat{\Delta}_{A_jB_k}$ is the deviation from the estimated probability of A_jB_k gametes, $\hat{p}(A_jB_k)$, from their expected probability $2\hat{p}_{A_j}\hat{p}_{B_k}$. The value $\hat{p}(A_jB_k)$ had to be determined indirectly from the count of A_jB_k combinations within biallelic genotypes (Table 6.1) as the gamete frequencies A_jB_k were unknown. In Table 6.1 the '#' indicated that there were no A_jB_k gametes present within the genotype thus the expected number of A_jB_k gametes given the genotype $A_jA_{j*}B_kB_{k*}$ is equal to $X_{A_j,A_{j*},B_k,B_{k*}}/2$ where $X_{A_j,A_{j*},B_k,B_{k*}}$ is the number of observed $A_jA_{j*}B_kB_{k*}$ genotypes. The ‘estimated observed’ frequency of A_jB_k gametes summed from both intra and inter gametic loci is $p(A_jB_k) = \lfloor 2X_{A_j,A_j,B_k,B_k} + X_{A_j,A_j,B_k,B_{k*}} + X_{A_j,A_{j*},B_k,B_k} + X_{A_j,A_{j*},B_k,B_{k*}} / 2 \rfloor / G$ with X being the count of each genotype and G is the total number of gametes (Schaid, 2004).

Table 6.1 Count of *A_jB_k* pairs within genotypes created from parental gametes at locus *A* and *B* where *j** (or *k**) is not allele *j* (or *k*).

		Female gametes			
		A_jB_k	A_jB_{k*}	$A_{j*}B_k$	$A_{j*}B_{k*}$
Male gametes	A_jB_k	2	1	1	1
	A_jB_{k*}	1	0	1#	0
	$A_{j*}B_k$	1	1#	0	0
	$A_{j*}B_{k*}$	1	0	0	0

The ‘#’ indicates where *A_jB_k* combinations occur in genotypes but not gametes.

Under the assumption of unlinked and neutral loci effective population size was estimated using linkage disequilibrium by correcting second order terms for sampling error

$$\hat{Ne} = \frac{1/3 + \sqrt{1/9 + 2.76\hat{r}^{2}{}'}}{2\hat{r}^{2}{}'} \tag{6.1}$$

where $\hat{r}^2{}' = \hat{r}^2 - E(\hat{r}^2_{sample})$ given $\hat{r}^2$ is the observed *r*-squared component calculated as the mean $\hat{r}_{A_jB_k}{}^2$ between all alleles over *L*(*L*-1)/2 pairwise comparisons of *L* loci, and $E(\hat{r}^2_{sample}) = \left[\frac{1}{S} + \frac{3.19}{S^2}\right]$ is the term correcting upward bias due to sampling *S* individuals (Waples, 2006). The derivation of these equations were the subject of a full paper (Waples, 2006). Briefly $\hat{Ne}$ is a quadratic solution (equation 6.1) for *Ne* formed by equating $\hat{r}^2{}'$ to $\frac{1}{3Ne} - \frac{0.69}{Ne^2}$ where $\frac{1}{3Ne}$ is the drift term assuming loci are unlinked in a random mating population and $-\frac{0.69}{Ne^2}$ is a second order correction determined by Waples (2006) using simulations.

The jackknife method was used to estimate the upper and lower 95% confidence intervals of $\hat{Ne}$ (Waples and Do, 2008). Large undefined Ne estimates occur when the correction due to finite sample size $\hat{r}^2_{sample}$ is greater than $\hat{r}^2$ resulting in a negative Ne estimate. Negative estimates are plausible and indicate that the sample size *S* is too small with the correction for sample size being larger than the $\hat{r}^2$ value determined from the data. Ne estimates were determined using program LDNE where the lower 95% confidence intervals of $\hat{Ne}$ were determined by the jackknife method (Waples and Do, 2008).

Built into the program of Waples and Do (2008) is a threshold called P_{crit}, which is used to exclude $\hat{r}_{A_jB_k}{}^2$ from the average $\hat{r}^2$ if $\hat{p}_{A_j}$ or $\hat{p}_{B_k}$ are below the P_{crit} threshold. Allele frequencies close to zero can bias $\hat{r}_{A_jB_k}{}^2$ (Waples, 2006). The study investigated $\hat{Ne}$ across a range of P_{crit} values as low frequency alleles are more common in large datasets. While the theory of Waples (2006) was tested using diallelic loci it applies equally well in polymorophic data sets (Waples and Do, 2010).

6.3.2 Collection of empirical data

Effective population size was estimated from genotypes of *S. commerson* individuals collected from a defined area, largely within 500km north-west of Darwin, Northern Territory (for more details see: Supplementary genotype methods).

6.3.3 Simulations with different effective population sizes

Ten thousand replicate linkage disequilibrium *Ne* estimates were determined each for a range of population sizes *N* from 3000 to 60000. The genotypes in each simulated population were generated using program SHAZA http://molecularfisherieslaboratory.com.au/shadow-zone-analysis-software-shaza/ (Macbeth et al*.,* 2011). This program simulated *N* first generation diploid genotypes by random sampling alleles within loci from the empirical allele frequencies of *S. commerson* in the Darwin population. The first *N*/2 genotypes were defined as females and the remainder males. Each individual in the next generation was simulated by random selection of a male and female with replacement. For each parental genotype and for all seven loci a single allele was randomly selected to create an individual diploid genotype. Following this process a total number of *N* individuals was created in four discrete generations.

In this design *N* is approximately equal to *Ne* (Waples, 2006). In each replicate, *Ne* was estimated from 5413 generation four genotypes using a plan 2 sampling procedure (Waples 1989). Generation four was used to estimate *Ne* as this was sufficient for $\hat{r}^2$ to approach an asymptotic value (Sved, 1971; Waples, 2006). For example, the expectation of $\hat{r}^2$ in the first generation of simulated genotypes will be zero resulting in upwardly biased estimates of Ne. Simulated genotypes had no missing loci therefore prior to estimating Ne missing loci were introduced to emulate the empirical data structure which had missing loci. The missing loci were introduced for each and every genotype in the simulated data by randomly drawing with replacement a genotype in the empirical data and deleting all loci in the simulated genotype that were found to be missing in the empirical genotype sampled.

6.3.4 Ne estimates from empirical data with outlier genotypes removed

Putative ‘outlier’ genotypes, defined as genotypes not originating from the focal population under investigation, were identified and removed from the empirical data using a correspondence analysis (CA). The CA algorithm used here was developed in a pilot study by visual assessment of simulated outliers from plots of the first two principal components of a singular value decomposition. Up to ten CA iterations were performed with iterations continuing until no further outliers were found. In each iteration, outlier genotypes were defined when principal components *PC1* and *PC2* (Appendix A) satisfied a threshold $\sqrt{(PC1^2 + PC2^2)} > 2$ which removed outliers furthest from the central cluster.

6.3.5 Ne estimates from empirical data with outlier genotypes removed and genotypes from non-target species added

To test the sensitivity of *Ne* estimates in genotype samples containing non-target species, a test was conducted by adding one hundred genotypes of a non-target species (grey mackerel, *Scomberomorus semifasciatus*) to the ‘cleaned’ *S. commerson* data. It was anticipated that adding foreign genotypes will increase $\hat{Ne}$ bias and indirectly show that cleaning the data could reduce bias in empicical data estimates. *Scomberomorus semifasciatus* genotypes amplified at five of the seven *S. commerson* loci with alleles at loci SCA47 and SCA49 marked as missing.

6.3.6 Simulation of genetically divergent populations

To further test the efficiency of the CA algorithm for detecting outlier genotypes, ten simulated populations were considered that diverged from a founding population across numerous generations. The allele frequencies of the founding population matched those from empirical *S. commerson* samples. Population size was set at *N*=10000 and after 100, 200, 500, 1000 or 2000 generations the population was sampled (sample size of 5413 genotypes). As described above, program SHAZA was used to generate *N* genotypes of the founding population. This was followed by creating *N* genotypes each successive generation from random sampling of parental alleles as described previously using an equal sex ratio. Pairwise F_{ST} values were determined between divergent simulated populations using Genetix 4.05 software (Belkhir et al., 1996-2004). For each of the ten populations, 100 samples were randomly removed and replaced by 100 random genotypes selected from one of the other nine populations. Following this procedure there were *n*=90 populations with 100 immigrant genotypes from non-target populations and *n*=10 populations with no immigrants. *Ne* was estimated before and after the data was cleaned using correspondence analysis.

The ability of the CA algorithm to identify immigrants was compared to the Bayesian clustering approach of STRUCTURE version 2.3.3 (Pritchard et al., 2000). STRUCTURE analysis was applied to the 90 populations that contained 100 immigrants after diverging 2000 generations. Runs were performed by specifying: *k*=2 clusters, an admixture ancestry model with allele frequencies correlated and a burn in length of 100 000 iterations followed by 100 000 MCMC iterations. One sample location was assumed with no location prior possible.

6.4 Results

6.4.1 Empirical data

The majority of the 5413 *S. commerson* samples were genotyped with all seven loci (71%), but some samples were genotyped with either six (12%), five (10%) and four (7%) polymorphic microsatellite loci. The numbers of alleles per microsatellite locus varied from 24 (90RTE) to 38 (SCA8) with 65% of alleles across all loci having frequencies less than or equal to 0.01 (Table 6.2). These low frequency alleles were selectively removed from data used to estimate *Ne* by the LDNe software depending on the chosen P_{crit} thresholds (for more details see: Supplementary genotype results).

Against expectations, LDNE estimates ($\hat{Ne}$) from empirical data varied systematically across P_{crit} values (Table 6.3). As the P_{crit} threshold decreased in magnitude so too did the magnitude of non-negative estimates of *Ne*. This covariance raised doubts about setting P_{crit} to 1/(2*S*)= 1/(2x5413)~0.0001, where all singleton alleles would be removed, and the general effectiveness of removing low frequency alleles for the estimation of *Ne*. The lower confidence interval of $\hat{Ne}$ was more stable than the mean estimates, but still varied widely from 406 to 24728 and as such provided no informative value of the lower bound of $\hat{Ne}$.

Table 6.2 Locus and allele frequency summary. Sample size at each locus (S_L) and number of alleles (*Na*) for microsatellite loci used to genotype *S. commerson* with the maximum frequency and number of alleles within loci having frequencies less than or greater than the range shown.

Locus	S_L	*Na*	Maximum allele frequency	Number of alleles with frequencies:		
				greater than 0.10	between 0.01 and 0.001	less than 0.001
SCA30	5210	36	0.178	2	17	8
SM3	5206	32	0.183	4	8	13
SM37	4611	37	0.127	2	16	9
SCA47	4781	27	0.486	3	4	14
SCA49	4829	25	0.248	5	5	8
90RTE	5266	24	0.735	1	6	11
SCA8	5139	38	0.216	4	12	11

Table 6.3 Estimates of LDNE effective population size ($\hat{Ne}$) in *S. commerson.* $\hat{Ne}$ at different P_{crit} thresholds with the upper and lower 95% confidence intervals.

	P_{crit}						
	0.05	0.02	0.01	0.001	0.0005	0.0001	0.0000
$\hat{Ne}$	-40163[a]	-799447	79842	17503	3584	503	418
$\hat{Ne}_{lower}$	19595	24728	22209	12759	3290	489	406
$\hat{Ne}_{upper}$	Infinite	Infinite	Infinite	27158	3921	517	428

[a] Negative $\hat{Ne}$ estimates indicate a large undefined Ne.

6.4.2 Simulations with different effective population sizes

Simulations indicated that 5413 genotype samples should be sufficient to estimate effective population size if the true size was 3000 and 10000 (Section 6.10 – Supplementary information: Figure 6.1 and Figure 6.2). Simulations with *N*=3000 (Figure 6.1) had no extreme estimates of Ne, whereas simulations with *N*=10000 (Figure 6.2) had a small number of outlier estimates that were greater than 40000 or less than minus 20000. In Figures 6.1 and 6.2, P_{crit} values between 0.01 and 0.001 gave the smallest standard deviation of $\hat{Ne}$, illustrating the importance of removing the majority of low frequency alleles.

As expected, simulations with *N*=100 and *N*=1000 (Figure 6.S1 and Figure 6.S2) gave more precise estimates of Ne than with *N*=3000 (Figure 6.1). Increasing *N* from 10000 to 30000 and 60000 (Figures 6.S3, 6.S4 and 6.S5) resulted in a lower precision of $\hat{Ne}$ with a greater number of negative and extremely large estimates of Ne. An interesting finding was that, at large *N* values such as 60000, the lower 95% confidence interval (Figure 6.S5) was more precise than the expected mean value (Figure 6.S4) particularly at P_{crit} values around 0.01. The results indicate that there was sufficiency in the data to detect the lower 95% confidence interval if *N* was equal to 60000 with the mean lower confidence interval being 22,188 (P_{crit}=0.01).

It is important to note that the smallest 1% of $\hat{Ne}$ using P_{crit} =0.0000 determined from the 100th ranked positive value was 4134, 5308 and 5846 when *N* was 10000, 30000 and 60000 respectively, which revealed an anomaly between the simulation results and empirical data estimates of $\hat{Ne}$. If the true Ne was larger than 10000 then the smallest $\hat{Ne}$ estimate expected at P_{crit} =0.0000 would be greater than 4134 (P< 0.01) which differs from the empirical estimate of 418. Conversely, if the true Ne was smaller than or equal to 10000 then simulations indicated that no negative estimates of $\hat{Ne}$ would be expected at P_{crit} =0.02 (P< 0.0001) which was contrary to that observed from empirical data with $\hat{Ne}$=-799447 (Table 6.3). This highlighted that there was a significant difference between the empirical and simulated data, which was subsequently investigated by examining ‘outlier’ genotypes.

Figure 6.1 Frequency of 10000 *Ne* estimates when simulating a population size of *N*=3000 at different P_{crit} values.

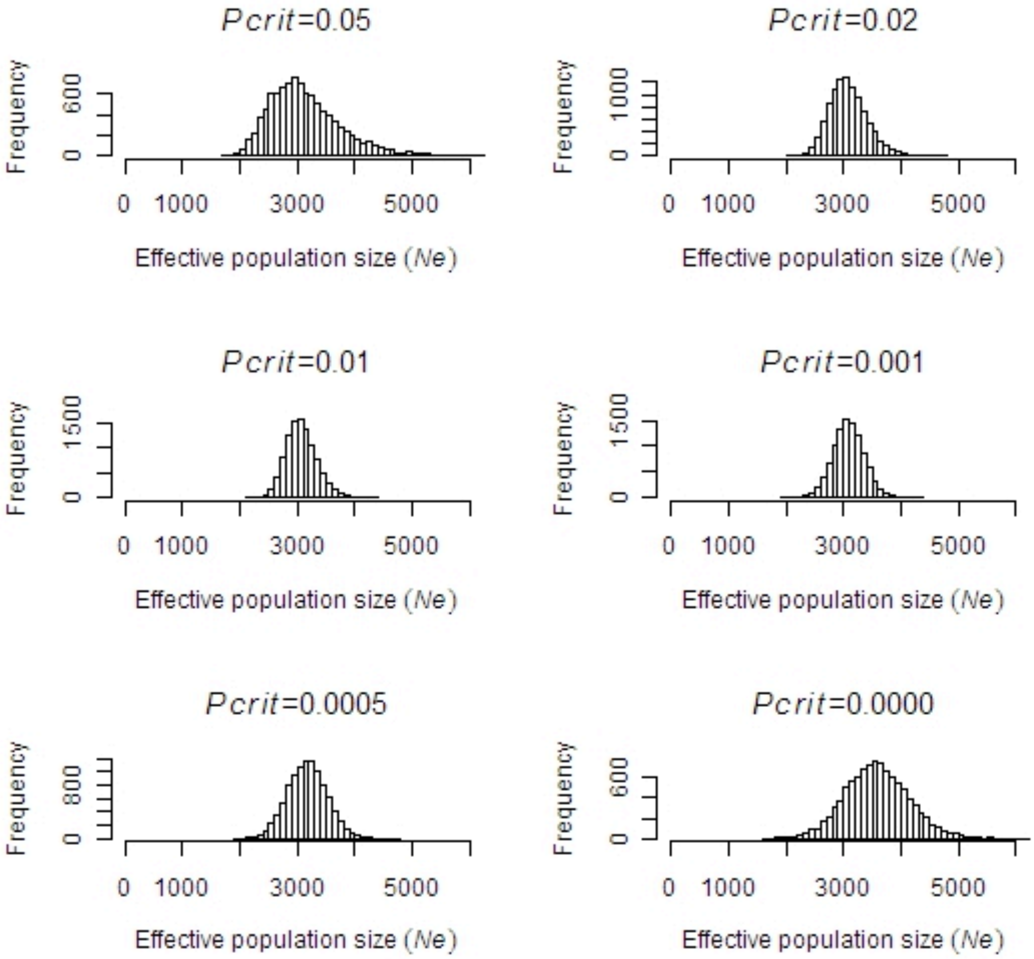
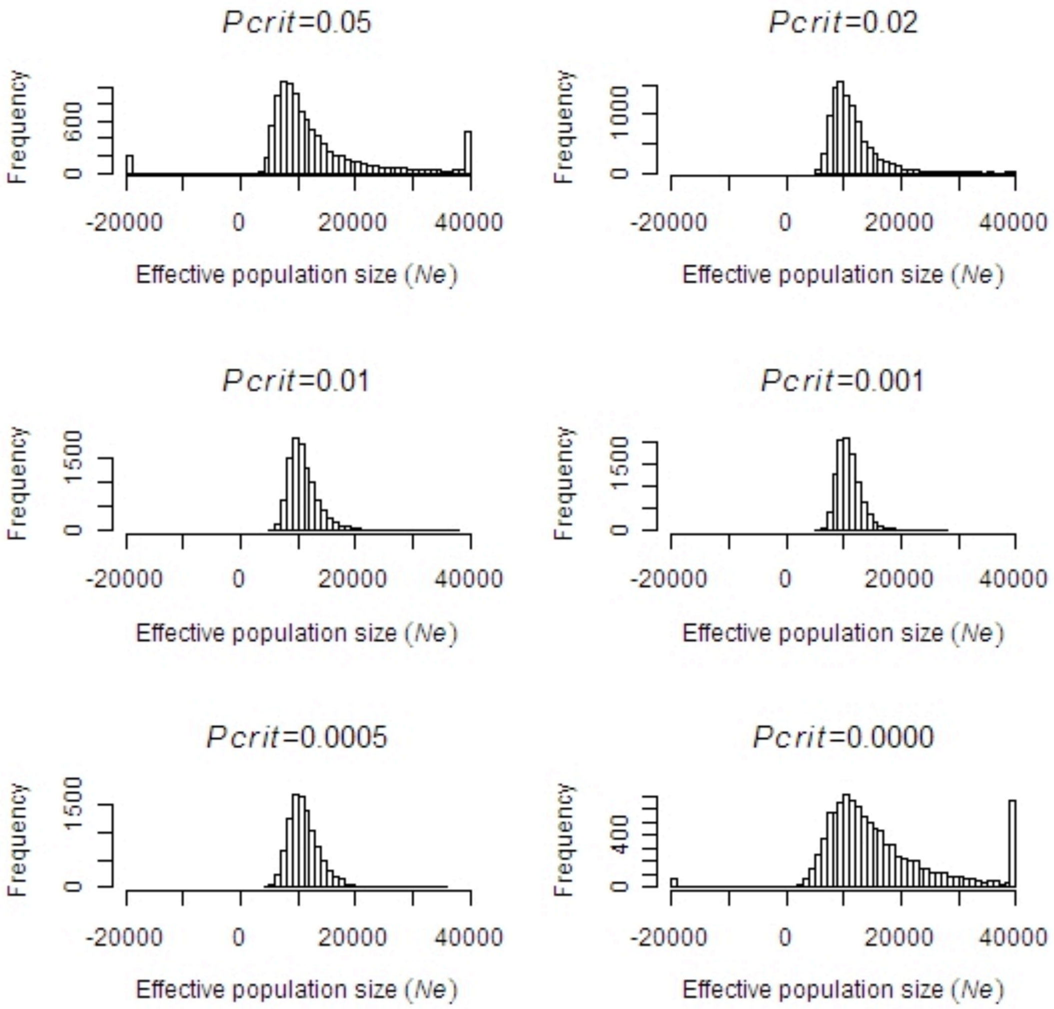


Figure 6.2 Frequency of 10000 *Ne* estimates when simulating a population size of *N*=10000 at different P_{crit} values. The frequency of all *Ne* estimates less than 20000 and greater than 40000 were pooled and are indicated on the *x*-axis limits of each graph.



6.4.3 Ne estimates from empirical data with outlier genotypes removed

The removal of putative outlier genotypes from empirical *S. commerson* data took nine *CA* iterations before there were no genotypes exceeding the $\sqrt{(PC1^2 + PC2^2)} > 2$ threshold (Figure 6.S6). An order of magnitude increase in $\hat{Ne}$ (Table 6.4) was observed after the first iteration, which removed just 33 outliers (0.6% of total number of genotypes). This indicated that putative outlier genotypes can significantly bias *Ne* estimates in empirical data.

Table 6.4 Estimates of *Ne* in *S. commerson* after CA iterations. The removal of putative outliers from nine sequential correspondence analysis (CA) iterations with the cumulative number of genotypes removed indicated in brackets and the following estimates of *Ne* at different P_{crit} thresholds.

CA	P_{crit}						
iteration (removed)	0.05	0.02	0.01	0.001	0.0005	0.0001	0.0000
0 (0)	-40163[a]	-799447	79842	17503	3584	503	418
1 (33)	-32062	-117650	90318	112421	55074	4968	5051
2 (38)	-33926	-114426	91549	104569	53546	8082	7947
3 (51)	-34571	-104127	93996	105937	48611	8838	9495
4 (60)	-37447	-99305	86818	113630	51105	133636	171370
5 (90)	-38487	-86051	89982	302878	-448815	-51226	-36471
6 (119)	-35678	-76242	120453	302946	-146528	-38189	-30685
7 (153)	-38909	-75672	101714	610512	-69972	-16082	-16082
8 (170)	-32038	-65015	296541	-795394	-58191	-14132	-14132
9 (174)	-32371	-67105	550582	-420513	-48637	-14059	-14059

[a] Negative $\hat{Ne}$ estimates indicate a large undefined Ne.

After the nine *CA* iterations, 3.2% of samples were removed. Subsequent *Ne* estimates on the cleaned data were negative at all P_{crit} thresholds, except when P_{crit} was 0.01 ($\hat{Ne}$=550582). This indicated that a P_{crit} of 0.01 provided the highest accuracy as it had the smallest confidence interval assessed by the fact that it was the only P_{crit} value where the correlation of alleles between loci was greater than that expected from sampling error. At this P_{crit} value $\hat{Ne}$ was relatively stable at around 80000 to 100000 until the last two iterations with $\hat{Ne}$ increasing to 550582. When P_{crit} was 0.01, the harmonic mean of $\hat{Ne}$ across all nine iterations was 110000.

The lower 95% confidence interval of the Ne estimates ($\hat{Ne}_{lower}$) from Table 6.4 is reported in Table 6.5. The lower confidence intervals appeared to be more stable than the estimates provided in Table 6.4 when the P_{crit} values were equal to or greater than 0.001. The range of $\hat{Ne}_{lower}$ estimates when P_{crit} = 0.01 were within 21% of each other with a harmonic mean of 24000.

Table 6.5 Lower 95% confidence interval of Ne from *S. commerson* genotypes. The removal of putative outliers from nine correspondence analysis (CA) iterations with the cumulative number of genotypes removed indicated in brackets and the following estimates of the lower 95% confidence interval ($\hat{N}e_{lower}$) at different P_{crit} thresholds.

CA iteration (removed)	P_{crit} 0.05	0.02	0.01	0.001	0.0005	0.0001	0.0000
0 (0)	19595	24728	22209	12759	3290	489	406
1 (33)	22540	30509	22943	26461	17594	1988	2046
2 (38)	21571	30713	23011	26119	17498	2849	2913
3 (51)	21232	31541	23144	33737	25337	7606	8131
4 (60)	20110	31970	22720	26879	16904	16696	14799
5 (90)	19615	33487	22809	42238	60094	-271390 [a]	-83353
6 (119)	20379	35118	24305	29804	53307	-98902	-59311
7 (153)	19174	34947	23471	31098	80748	-35452	-35453
8 (170)	21646	37832	27703	36446	151392	-23066	-23066
9 (174)	21445	37064	28922	35858	-615338	-23260	-23260

[a] Negative $\hat{N}e$ estimates indicate a large undefined Ne.

6.4.4 Ne estimates from empirical data with outlier genotypes removed and genotypes from non-target species added

Adding non-target species (grey mackerel, *S. semifasciatus*) to the ‘cleaned’ *S. commerson* data significantly reduced *Ne* estimates (Table 6.6). Considering the total sample size was 5413, the results clearly show that only a small proportion of non-target species can have a large impact on linkage disequilibrium estimates of *Ne*. For example, adding as few as eight (0.15%) *S. semifasciatus* genotypes resulted in a 5.7 fold reduction in $\hat{N}e$ when P_{crit} =0.01. All of the 200 non-target grey mackerel genotypes were identified and removed by the first iteration of CA analysis compared to the nine iterations that were required with the empirical data (Table 6.4). This suggests that the putative outliers in the empirical data were more similar to *S. commerson* than *S. semifasciatus.*

The *S. semifasciatus* samples did not amplify at loci SCA47 and SCA49. Removing all genotypes in the empirical data that did not amplify at these two loci produced a similar $\hat{N}e$

profile to Table 6.3, indicating that *S. semifasciatus* cannot be solely implicated in the anomaly between the simulated and empirical data.

Table 6.6 Effect of *S. commerson Ne* estimates when adding non-target species. Starting with *S. commerson* data with 174 outliers removed by nine CA iterations, *Ne* estimates at different P_{crit} thresholds were determined after progressive addition of grey mackerel (*S. semifasciatus*) genotypes.

Grey mackerel genotypes added	P_{crit}						
	0.05	0.02	0.01	0.001	0.0005	0.0001	0.0000
0	-32371[a]	-67105	550582	-420513	-48637	-14059	-14059
1	-32382	-67686	566612	-410564	-48310	1303	1303
2	-32315	-67583	719220	-356551	-47594	1031	1031
4	-35620	-70777	159027	-966684	-50839	1138	1138
8	-36871	-79371	95957	206370	3930	1179	1179
16	-37624	-94247	43218	2030	1088	1238	1238
32	-45964	-1040355	16140	1104	985	1233	1233
64	626218	5420	2896	700	776	974	1014
100	23439	5946	813	553	654	806	862
200	2189	418	233	376	455	547	620

[a] Negative $\hat{Ne}$ estimates indicate a large undefined Ne.

6.4.5 Simulation of genetically divergent populations

Ten populations simulated after divergence from a common founder population had average pairwise F_{ST} values of 0.004, 0.010, 0.027, 0.048 and 0.091 after 100, 200, 500, 1000 and 2000 generations respectively. With no mixing of the populations during genotype sampling, *Ne* estimates approximated the simulated population size (*N*=10000, Table 6.7).

Table 6.7 Harmonic mean of $\hat{Ne}$ before and after outlier genotypes removed. Harmonic mean of $\hat{Ne}$ at two P_{crit} thresholds in simulated populations with *N*=10000 and sample size *S*=5413 containing no immigrants or with 100 genotypes drawn from a single immigrant population. The immigrants are from populations diverging after a different number of generations from a common population. The harmonic mean in each column was based on *n* separate $\hat{Ne}$ estimates before and after outlier genotypes were removed using the CA algorithm.

	Before outlier genotypes removed		After outlier genotypes removed	
	No immigrants n=10	With immigrants n=90	No immigrants n=10	With immigrants n=90
Generations				
	P_{crit}=0.000			
100	9896	6236	13911	17100
200	10543	3037	11947	13973
500	10029	1282	11151	11558
1000	97734	571	10548	11049
2000	11834	176	12359	12295
	P_{crit}=0.010			
100	10732	11096	10841	11267
200	10557	10932	10670	11094
500	10211	9420	10217	10003
1000	9595	7629	9691	9736
2000	10407	4456	10508	10564

Ninety populations with 100 immigrants were created from pairs of the ten divergent populations. Across these 90 populations CA analysis found an average (standard deviation) of 7 (4), 18 (8), 44 (12), 74 (12) and 93 (6) immigrants after 100, 200, 500, 1000 and 2000 generations respectively. The average number of CA iterations required before no more immigrants could be detected were 3.4, 3.6, 3.6, 3.1, 3.0 after 100, 200, 500, 1000 and 2000 generations respectively. As a comparison, the program STRUCTURE was not able to distinguish the immigrants, even after 2000 generations of divergence. When two populations were specified in STRUCTURE ninety-seven of the 100 immigrants and 47.3% of the remaining 5313 samples were partitioned into the same population. This indicated that there was not sufficient genetic divergence between the populations to cluster the small proportion of immigrants into a separate population.

In the presence of 100 immigrants, there was a downward bias in $\hat{Ne}$ of the focal population for P_{crit} values of 0.00 and 0.01 (Table 6.7) as the number of generations of divergence increased. After outlier genotypes were removed *Ne* estimates were more consistent with an expected value of *N*=10000. After outlier (i.e. immigrant) genotypes were removed by CA, the smallest bias and highest accuracy of *Ne* occurred when P_{crit} =0.01.

6.5 Discussion

Palstra and Ruzzante (2011) urged further theoretical developments in order to avoid a downward bias in estimating linkage disequilibrium *Ne* in naturally-occurring metapopulations. The results have demonstrated that under certain circumstances even estimates for focal populations can be downwardly biased. It is believed that this bias amongst the samples taken for estimation could be due to the presence of 1) non-target-species and 2) immigrant genotypes from diverged populations. Importantly, only a few 'contaminant' genotypes can severely bias *Ne* estimates. The contaminant genotypes are not at equilibrium in the recipient population, so the results from this study are not in disagreement with a study showing that linkage disequilibrium estimates of effective population size are robust to equilibrium migration (Waples and England, 2011).

The correspondence analysis algorithm (CA) performed well in identifying and removing non-target genotypes that were added to simulated population samples. Standard methods of population clustering such as STRUCTURE (Pritchard et al., 2000), were incapable of identifying the simulated immigrants. The threshold value of 2 used in the CA algorithm was developed by trial and error as a reasonable threshold to exclude outlier genotypes without removing too many target population genotypes. A series of scatter plots on principal coordinates is shown after each iteration of removing outliers on the threshold (Figure 6.S6). The pattern in this series was typical for many of the simulation runs where a final cluster of points becomes clearly visible. As expected, as the F_{ST} between non-target and the target populations decreased, it was more difficult to detect the non-target genotypes. Further development is required to test and refine the CA algorithm under a broader range of allele frequencies and number of loci. The general message is that it is worthwhile to detect and remove putative non-target genotypes prior to LDNE analysis.

The simulated divergent populations were implemented using a simple Wright-Fisher model with mating modified such that gametes were chosen from populations having equal numbers in each sex. This model was used by Waples (2006) however many other models could have been used including those with mutation and selection (Der et al., 2011). These additional processes would cause a larger divergence at the same number of generations compared to the simple genetic drift model used in the study.

The investigation suggests that mackerel genotypes collected around Darwin contained a small proportion from genetically divergent *S. commerson* population(s) or from congeneric species. It is possible that tissue samples of closely related species were taken inadvertently thus mimicking an admixed *S. commerson* population. The 100 grey mackerel (*S. semifasciatus*) samples amplified at 5 of the 7 loci used in this study, while another closely related endemic species (*Scomberomorus queenslandicus*) amplifies at all the 7 loci (unpublished data). The fact that all grey mackerel genotypes were successfully removed by the correspondence analysis method does indicate that the method works well when non-target species are implicated. It would be expected to have intermediate results when populations are at varying levels of population divergence as indicated by simulations in this study.

No genotyping errors were assumed when estimating linkage disequilibrium, although pre-screening of the data resulted in one locus being removed due to a deviation from Hardy Weinberg equilibrium. While this deviation might indicate the presence of a null allele error there could be other errors such as allelic dropout errors. Random dropout errors are not expected to change the expectation of linkage disequilibrium estimates nor the outcome of the expected *Ne* estimate.

Assuming that all samples represented *S. commerson*, it is likely that the population adjacent to Darwin is an admixed population containing small numbers of individuals from genetically distinct populations. These individuals could also have been transient vagrants of genetically distinct populations of *S. commerson* (Sulaiman and Ovenden, 2010; Fauvelot and Borsa, 2011) that were sampled in the same geographical region. The hypothesis that a small (rather than large) number of immigrant genotypes were present in the empirical genotypes is supported by the observations that (a) most adults in a mark-recapture study were found to move less than 100km per year parallel to the shore (Buckworth et al., 2012) and (b) isotope signatures in the sagittal otolith carbonate of *S. commerson* indicated spatial separation across northern Australia (Newman et al., 2009).

In the *S. commerson* data it was very difficult to get a precise estimate of $\hat{N}e$. Before 'cleaning' the data with CA, *Ne* estimates varied at different P_{crit} levels including some negative estimates of *Ne.* Using a P_{crit} value of 0.01 the likely $\hat{N}e$ seems very large with an estimate of 110000 from empirical data. This estimate was believed to be unreliable as inferred from the lack of sufficiency of the data when estimating the mean *Ne* with *N*=600000 (Figure 6.S4).

Negative estimates of Ne are counter-intuitive and indicate that the true *Ne* is large and undefined. Waples and Do (2010) point out that even if the Ne estimate is negative, if adequate data is available the lower bound of the confidence interval may be finite and can provide useful information. This finding was also supported by the simulations with large *N* values where the lower 95% confidence interval for *S. commerson* appears to be much more stable than the estimate and upper limits. Using a P_{crit} value of 0.01, the lower 95% confidence interval gave a harmonic mean of $\hat{N}e$=24000 from empirical data. More loci could be used to achieve more precise estimates of *Ne.* However there was sufficiency in the data to detect *Ne* when *N*=30000 (Figure 6.S3, P_{crit}=0.01). There was also sufficiency in the data when estimating the lower 95% confidence of *Ne* with *N*=60000 giving $\hat{N}e_{lower}$=22188 (Figure 6.S5, P_{crit} =0.01). It was concluded from these simulations that the empirical Ne_{lower} estimate of 24000 is reasonably reliable. In ecological terms 24000 represents a large and stable genetic population size and it would be expected to reach a similar conclusion with the addition of more loci.

This study was primarily focussed on the bias in the linkage disequilibrium estimation of *Ne* when a population may include genetically divergent conspecifics. There are many other approaches used to estimate *Ne* that have different underlying assumptions (Barker, 2011) and which should be evaluated as being suitable for the estimation of *Ne* in large, naturally-occurring populations. A natural progression in this area of research is to develop inferences of census population sizes based on effective size estimates (Palstra and Ruzzante, 2008) and how these could be used to assist management of natural resource species.

6.6 Conclusion

Realistic simulations have shown that it is possible to make effective population size estimates using the linkage disequilibrium method with finite confidence limits up to several thousand depending on the number of loci and genotypes assayed. Estimates of effective size made from samples taken from naturally occurring populations need to be treated with caution. It is recommend that screening of outliers of the sampled genotypes should be undertaken, particularly if the population being studied is sympatric with closely-related species, or is possibly receiving immigrants from adjacent populations.

6.7 Acknowledgements

The authors of an earlier FRDC project (No 2002/011) are thanked for making the *S. commerson* genotypes available and to Raewyn Street and Marcus McHale for their technical support. Helpful comments were provided by Professor Robin Waples, Dr Phillip England, Dr Friso Palstra and two anonymous referees.

6.8 References

Barker JSF (2011) Effective population size of natural populations of *Drosophila buzzatii*, with a comparative evaluation of nine methods of estimation. Molecular Ecology 20:4452-4471.

Blower DC, Pandolfi JM, Gomez-Cabrera MDC, Bruce BD, Ovenden JR (2012) Population genetics of Australian white sharks reveals fine-scale spatial structure, transoceanic dispersal events and low effective population sizes. Marine Ecology Progress Series 455:229-244.

Beacham TD, Candy JR, McIntosh B, MacConnachie C, Tabat A, Kaukinen K, Deng L, Miller KM, Withler RE (2005) Estimation of stock composition and individual identification of Sockeye salmon on a Pacific rim basis using microsatellite and major histocompatibility complex variation. Transactions of the American Fisheries Society 134:1124-1146.

Bekkevold D, Clausen LAW, Mariani S, Andre C, Christensen TB, Mosegaard H (2007) Divergent origins of sympatric herring population components determined using genetic mixture analysis. Marine Ecology Progress Series 337:187-196.

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (1996-2004) GENETIX 4.05, logiciel sous Windows TM pour la genetique des populations. Laboratorie Genome, Populations, Interactions, CNRS UMR 5171, Universite de Montpellier II, Montpellier (France).

Buckworth RC, Newman SJ, Ovenden JR, Lester RJG, McPherson GR (2007) The stock structure of northern and western Australian Spanish mackerel. Darwin, Australia: Dept. of Primary Industry, Fisheries and Mines. Fishery Report No. 88.

Buckworth RC, Ovenden JR, Broderick D, Macbeth GM, McPherson GR, Phelan MJ (2012). GENETAG: Genetic mark-recapture for real-time harvest rate monitoring: Pilot studies in northern Australia Spanish Mackerel fisheries. Northern Territory Government, Australia. Fishery.Report No. 107.

Collette BB, Nauen CE (1983) FAO species catalogue. Vol.2. Scombrids of the world. An annotated and illustrated catalogue of tunas, mackerels, bonitos and related species known to date. Food and Agriculture Organization Fish Synopsis (125); i-vii, 1-137, Rome.

Der R, Epstein CL, Plotkin JB (2011) Generalised population models and the nature of genetic drift. Theoretical Population Biology 80:80-99.

Fauvelot C, Borsa P (2011) Patterns of genetic isolation in a widely distributed pelagic fish, the narrow-barred Spanish mackerel (*Scomberomous commerson*). Biological Journal of the Linnean Society 104:886-902.

Hare MP, Nunney L, Schwartz MK*,* Ruzzante DE, Burford M, Waples RS, Ruegg K, Palstra F (2011) Understanding and estimating effective population size for practical application in marine species management. Conservation Biology 25:438-449.

Hedgecock D, Launey S, Pudovkin AI, Naciri Y, Lapegue S, Bonhomme F (2007) Small effective number of parents (N-b) inferred for a naturally spawned cohort of juvenile European flat oysters *Ostrea edulis*. Marine Biology 150:1173-1182.

Jorgensen SJ, Reeb CA, Chapple TK, Anderson S, Perle C, Van Sommeran SR, Fritz-Cope C, Brown AC, Klimley P, Block BA (2009) Philopatry and migration of Pacific white sharks. Proceedings of the Royal Society of Biological Sciences doi:10.1098/rspb.2009.1155

Luikart G, Ryman N, Tallmon DA, Schwartz MK, Allendorf FW (2010) Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. Conservation Genetics 11:255-373.

Macbeth GM, Broderick D, Ovenden JR, Buckworth RC (2011) Likelihood-based genetic mark-recapture estimates when genotype samples are incomplete and contain typing errors. Theoretical Population Biology 80:185-196.

McPherson GR (1988) A review of large coastal pelagic fishes in the South Pacific Region, with special reference to *Scomberomorus commerson* in north-east Australian waters. South Pacific Commission/Inshore Fisheries Resources/WP.15.

Newman SJ, Buckworth RC, Mackie M, Lewis P, Wright I, Williamson P, Bastow TP, Ovenden JR (2009) Spatial subdivision of adult assemblages of Spanish mackerel, *Scomberomorus commerson* (Pisces: Scombridae) across northern Australia: implications for fisheries management. Global Ecology and Biogeography 18:711-723.

Ovenden J, Peel D, Street R, Courtney AJ, Hoyle SD, Peel SL, Podlich H (2007) The genetic effective and adult census size of an Australian population of tiger prawns (*Penaeus esculentus*). Molecular Ecology 16:127-138.

Palstra FP, Ruzzante DE (2008) Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? Molecular Ecology 17:3428-3447.

Palstra FP, Ruzzante DE (2011) Demographic and genetic factors shaping contemporary metapopulation effective size and its empirical estimation in salmonid fish. Heredity 107:444-455.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Pudovkin AI, Zaykin DV, Hedgecock D (1996) On the potential for estimating the effective number of breeders from heterozygote excess in progeny. Genetics 144:383-387.

Schaid DJ (2004) Linkage Disequilibrium Testing When Linkage Phase Is Unknown. Genetics 166:505-512.

Smith BL, Alvarado-Bremer JR (2010) Inferring population admixture with multiple nuclear genetic markers and Bayesian genetic clustering in Atlantic swordfish (*Xiphlas gladius*). Collective Volume of Scientific Papers ICCAT 65:185-190.

Sulaiman ZH, Ovenden JR (2010) Population genetic evidence for the east–west division of the narrow-barred Spanish mackerel (*Scomberomorus commerson, Perciformes: Teleostei*) along Wallace’s Line. Biodiversity and Conservation 19:563-574.

Sved JA (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theoretical Population Biology 2:125–141.

Svedang H, Righton D, Jonsson P (2007) Migratory behaviour of Atlantic cod *Gadus morhua*: natal homing is the prime stock-separating mechanism. Marine Ecology Progress Series 345:1-12.

Tillett BJ, Meekan MG, Field IC, Thorburn DC, Ovenden JR (2012) Evidence for reproductive philopatry in the bull shark, *Carcharhinus leucas* in northern Australia. Journal of Fish Biology 80:2140-2158.

Waples RS (1989) A Generalized Approach for Estimating Effective Population Size From Temporal Changes in Allele Frequency. Genetics 121:379-391.

Waples RS (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conservation Genetics 8:167-184.

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

Waples RS, 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 RS, England PR (2011) Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. Genetics 189:633-44.

Weir B (1996) Genetic Data Analysis II. Sinauer Associates, Sunderland, MA. p 137.

Wright S (1930) Evolution in mendelian populations. Genetics 16:97-159.

Zhdanova O, Pudovkin AI (2008) Nb_HetEx: A Program to Estimate the Effective Number of Breeders. Journal of Heredity 99:694-695.

6.9 Appendix A – Correspondence analysis

Correspondence analysis R script: The genotype file is presented as a matrix *Z* having columns for each allele within every locus e.g. L1A1, L1A2, L1A3, L2A1, L2A2 indicates the data has 5 columns with locus one (L1) having three alleles (A1..A3) and locus two having two alleles (A1..A2). For each and every genotype a single row marking the number of alleles at each locus and at each allele with values ‘0’ for no alleles, ‘1’ for one heterozygous allele or ‘2’ for homozygote alleles. The total count within each locus across all alleles should sum to two.

For illustrative purposes only a small *Z* file with four genotypes will have a format like:

```
L1A1, L1A2, L1A3, L2A1, L2A2
1,0,1,0,2
1,1,0,1,1
0,2,0,1,1,
0,1,1,2,0
```

The R code modified from Nenadic and Greenacre (2006) converts the incidence matrix to a format that can be read and manipulated by R (R Development Core Team 2011) with the first two principal components *PC1* and *PC2* determined as:

```
Z   <-  data.matrix(Z)                      # convert to matrix
P   <- Z / sum(Z)                           # proportional contribution
rm  <- apply(P, 1, sum)                     # sum rows
cm  <- apply(P, 2, sum)                     # sum columns
eP  <- rm %*% t(cm)                         # multiply by transpose
dec <- svd((P - eP) / sqrt(eP))             # singular value decomposition
PC1 <- dec$u[,1] * dec$d[1] / sqrt(rm)      # Principal component 1
PC2 <- dec$u[,2] * dec$d[2] / sqrt(rm)      # Principal component 2
```

6.10 Supplementary information

6.10.1 Genotype methods

Tissue samples were taken from fish and stored in 90% ethanol or a saturated $NaCl_2$ solution containing 20% dimethyl sulphate. In total, 5413 genotypes from seven polymorphic microsatellite loci were collected between 2003 and 2006.

Samples were genotyped with seven di-nucleotide microsatellite loci; *90RTE* (Van Herwerden et al., 2000), *SCA8, SCA30, SCA47, SCA49* (Gold et al., 2002), *SM3* (GenBank AY700810.1) and *SM37* (GenBank AY700844.1). Genomic DNA was extracted using the salting-out method (Sambrook et al., 1989). Microsatellite amplifications for the seven loci were performed in four multiplexed reactions in 96-well plates using Perkin Elmer (Waltham, MA, U.S.A.) 9600 and 9700 series thermocyclers. The PCR volume per well was six microliters with QIAGEN® (Hilden, Germany) master mix (containing Taq polymerase and magnesium chloride) and QIAGEN® (Hilden, Germany) Q-solution was used to facilitate multiplexing. Mineral oil was used to control evaporation during cycling. Cycling conditions consisted of denaturation at 95°C for 15 min, followed by 37 cycles of 94°C for 30 sec at 56°C for 45 sec and 72°C for 1 min 30 sec. A final extension at 72°C for 45 min was used to ensure complete addition of adenine to the PCR product. Microsatellite gel separation and scoring was performed on a Life Technologies™ (Carlsbad, CA, U.S.A.) ABI™ 3130*xl* Genetic Analyser. Life Technologies™ Genemapper™ 3.7 software was used to score alleles, to assign them to bin classes and export genotype information for subsequent analyses.

Empirical data was tested for deviations from Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium using Genepop-on-the-web v4.0.10 (Rousset, 2008). For HWE tests, all locus *x* population combinations were tested. Tests for linkage disequilibrium considered all combinations of locus pairs for each population. Tests were made with successively larger batch sizes in Genepop until a stable result was obtained. Bonferroni corrections for simultaneous tests were applied commencing with an α level of 0.05. The software Microchecker (Van Oosterhout et al., 2004) was used to examine cases of deviation for Hardy-Weinberg equilibrium for microsatellite data. Microsatellite data was analysed in blocks of less than 500 samples to avoid the upper limit of the Microchecker software.

6.10.2 Genotype results

Average observed heterozygosity across seven microsatellite loci was 0.762 and the average expected heterozygosity was 0.802. Tests for Hardy-Weinberg equilibrium rejected the null hypothesis for all seven loci. Locus-by-locus analysis with Microchecker showed that for some alleles there was a difference in the observed and expected number of homozygotes, inferring null alleles may be present. Nulls were predicted by the software at loci *Sca49* and *Sca47* at frequencies ranging from 0.03 to 0.09, and nulls were detected at lower frequencies at some other loci. Graphical representation by Microchecker of the observed and expected frequency of heterozygotes, plotted against the number of base pairs separating the two alleles, revealed a deficit in heterozygotes when alleles were separated by two base pairs and a compensatory increase in the observed number of homozygotes (Figure 6.S7). This could be explained by a slight scoring error, which may have been responsible for the null allele predictions made by Microchecker and which may have been compounded by large sample sizes in the HWE tests. Wakefield (2010) confirms that rejection of the null hypothesis using conventional *p*-values is more likely when sample sizes are large and recommends a Bayesian framework in these cases. Thus, a small proportion of heterozygote genotypes were underrepresented in the microsatellite data. There was unlikely to be a systematic bias in the microsatellite data, as the controlling factor in their omission was similarity in allele size, which should occur evenly across alleles independent of their frequency or size, and across samples independent of biological factors.

6.10.3 Simulation Results

Figures 6.S1 to 6.S6 are refered to within the main manuscript. Briefly the frequency distribution of *Ne* estimates is a good indicator of the precision obtained from the seven polymorphic loci used in this study. Ideally a tight cluster of *Ne* estimates is desirable (Figure 6.S1, *Pcrit*=0.01). When there is insufficient genotype data negative and or very large estimates can occur (Figure 6.S4). The lower 95% confidence interval of *Ne* (Figure 6.S5) was less variable than the mean expectation (Figure 6.S5).

6.10.4 Supplementary References

Gold JR, Pak E, DeVries DA (2002) Population structure of king mackerel (*Scomberomorus cavalla*) around peninsular Florida, as revealed by microsatellite DNA. Fisheries Bulletin 100:491-509.

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

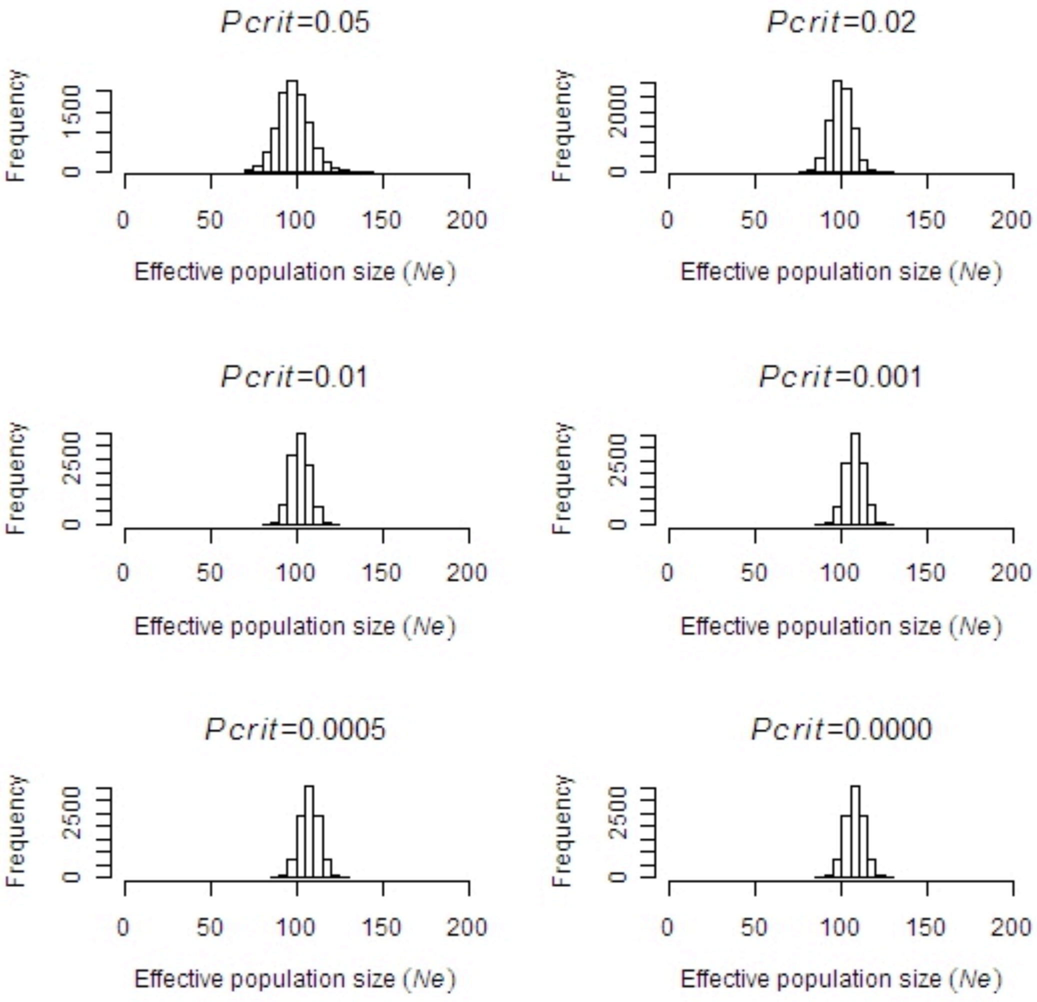
Sambrook J, Fritsch EF, Maniatis T (1989) Molecular Cloning: A Laboratory Manual, 2nd edn. Cold Spring Harbour Laboratory Press, Cold Spring, New York.

Van Herwerden L, Benzie J, Peplow L, Davies C (2000) Microsatellite markers for coral trout (*Plectropomus laevis*) and red throat emperor (*Lethrinus miniatus*) and their utility in other species of reef fish. Molecular Ecology 9:1919-1952.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICROCHECKER: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes 4:535-538.

Wakefield J (2010) Bayesian methods for examining Hardy-Weinberg equilibrium. Biometrics 66:257-265.

Figure 6.S1 Frequency of 10000 *Ne* estimates when simulating a population size of *N*=100 at different P_{crit} values.



Pcrit=0.05

Pcrit=0.02

Pcrit=0.01

Pcrit=0.001

Pcrit=0.0005

Pcrit=0.0000

Figure 6.S2 Frequency of 10000 *Ne* estimates when simulating a population size of *N*=1000 at different P_{crit} values.

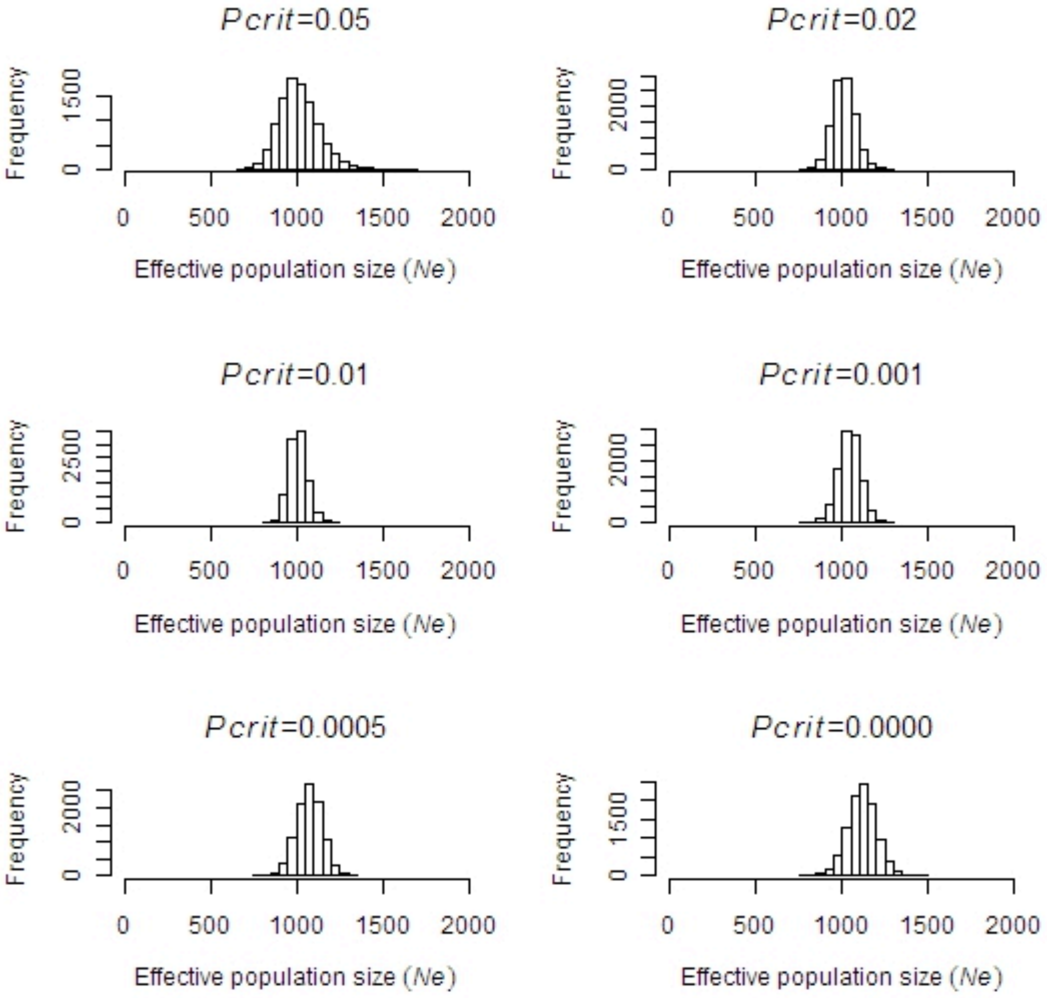


Figure 6.S3 Frequency of 10000 *Ne* estimates when simulating a population size of *N*=30000 at different P_{crit} values. The frequency of all *Ne* estimates less than 100000 and greater than 100000 were pooled and are indicated on the *x*-axis limits of each graph.

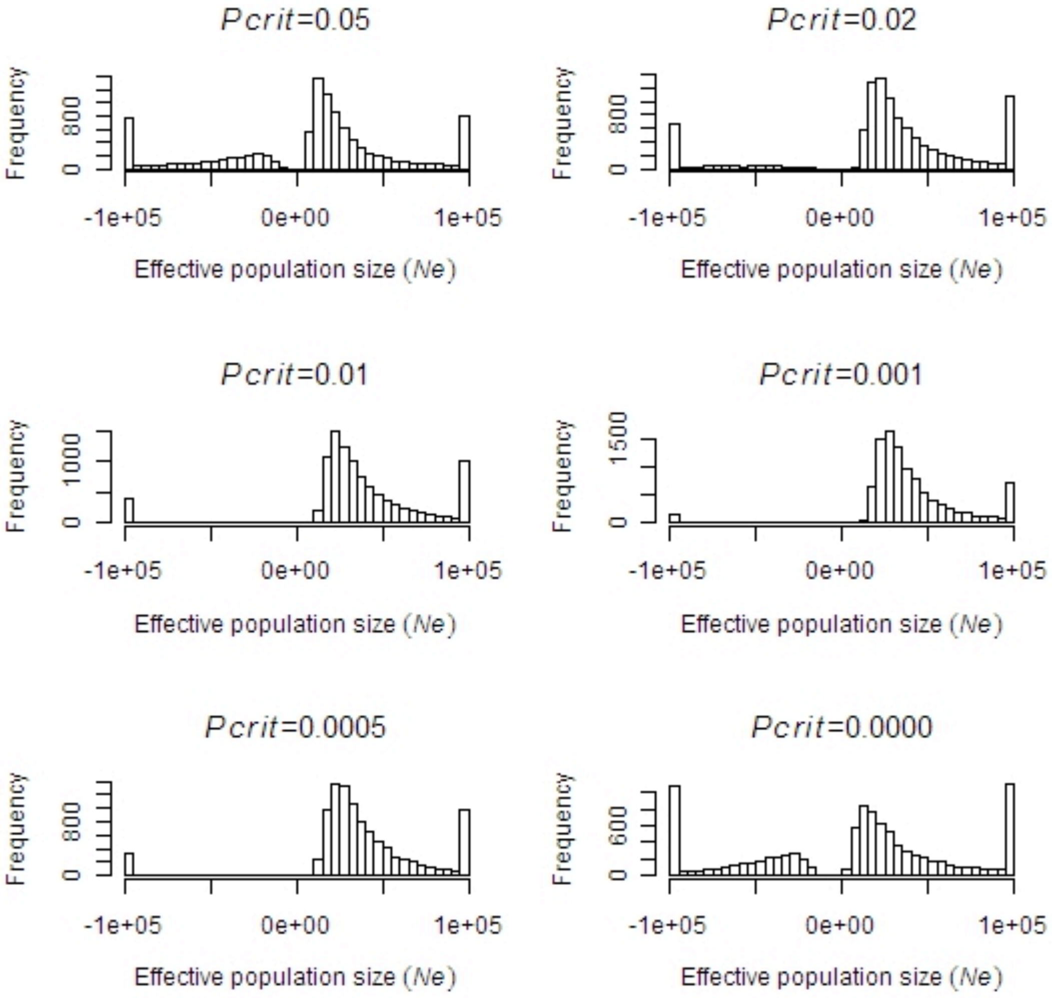


Figure 6.S4 Frequency of 10000 *Ne* estimates when simulating a population size of *N*=60000 at different P_{crit} values. The frequency of all *Ne* estimates less than 100000 and greater than 100000 were pooled and are indicated on the *x*-axis limits of each graph.

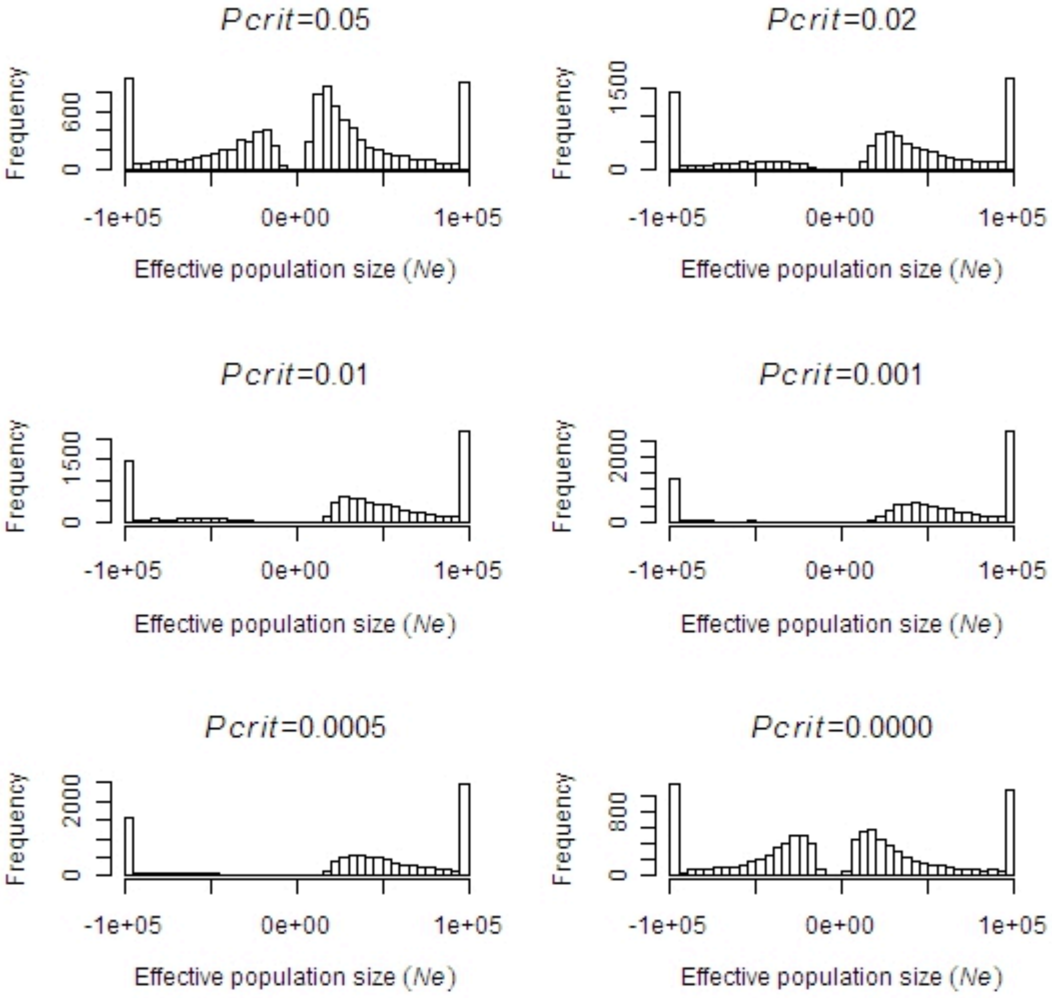


Figure 6.S5 Frequency of lower 95% confidence interval of $\hat{Ne}$ from 10000 estimates when simulating a population size of *N*=60000 at different P_{crit} values. The frequency of all *Ne* estimates less than 100000 and greater than 100000 were pooled and are indicated on the *x*-axis limits of each graph.

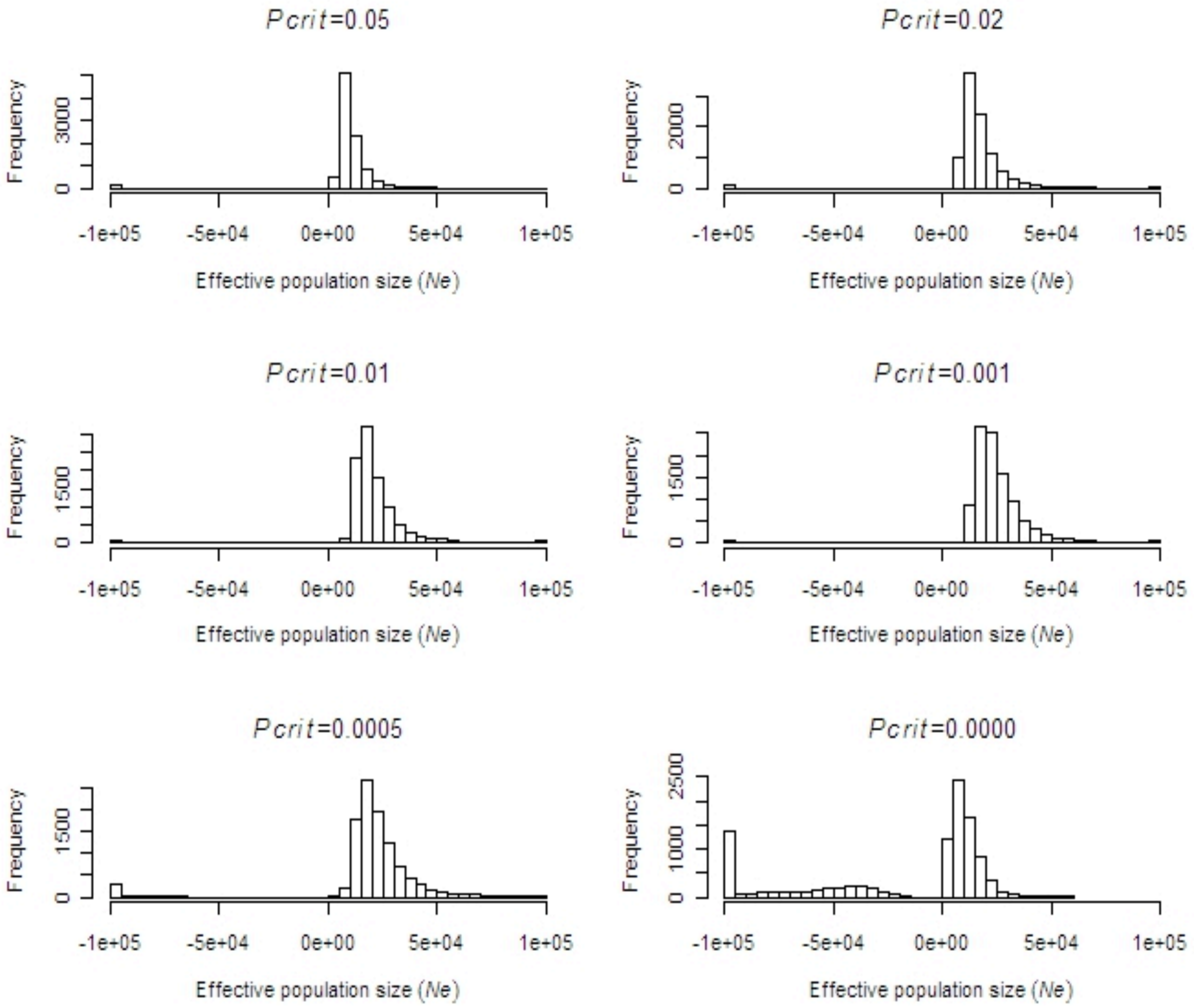


Figure 6.S6 Correspondence analysis plots after nine iterations of removing outliers in the empirical mackerel data that satisfied the threshold $\sqrt{(PC1^2 + PC2^2)} > 2$ where *PC1* and *PC2* are the first and second principal components. Iterative steps are from top left to right moving down rows. The last plot shows a cluster ball of genotypes after removing 116 genotypes from 5413 genotypes. One more iteration (not shown) removed 4 additional points.

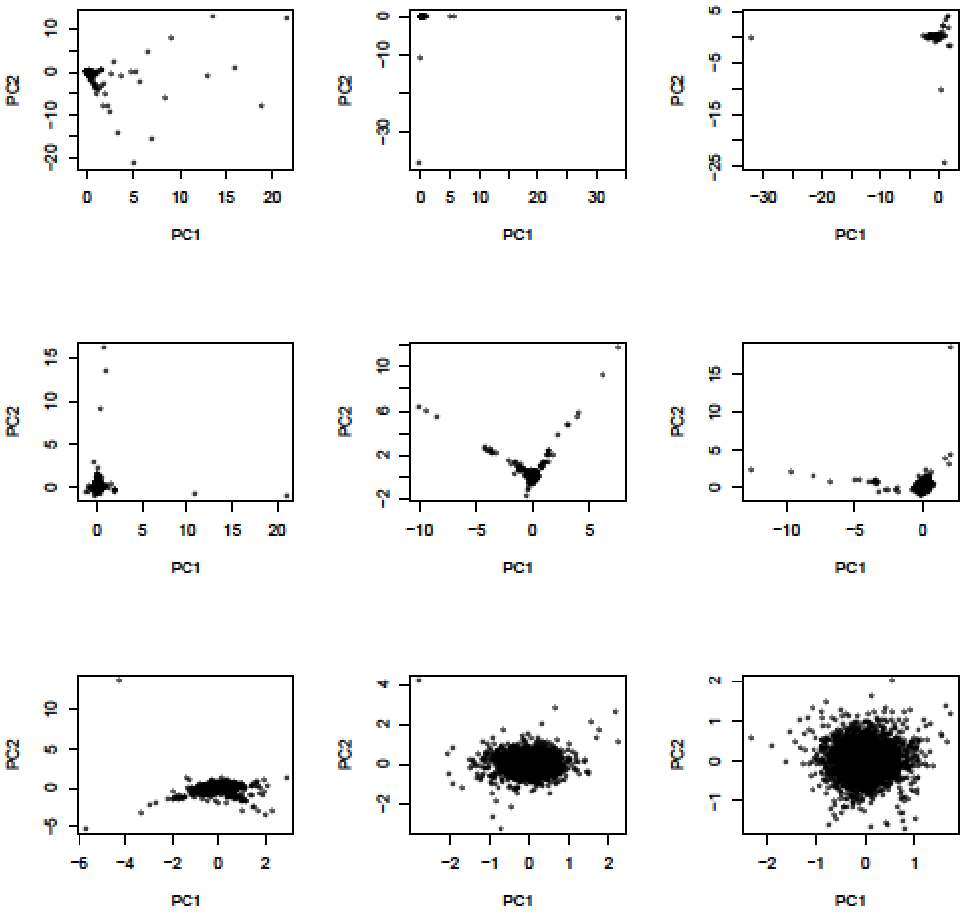
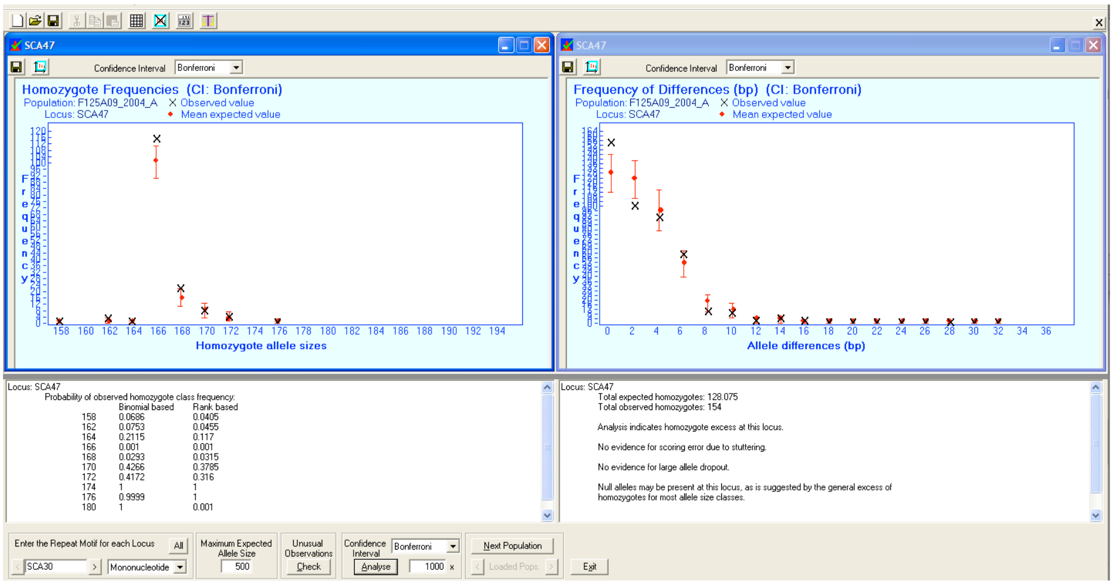


Figure 6.S7. Graphical output from Microchecker software showing observed (X) and expected (red vertical bars) frequency of homozygotes (left panel) and heterozygotes (right panel) for 500 genotypes from 2004 collected adjacent to Darwin.



Confidence Interval: Bonferroni

Population: F125A09_2004_A

Locus: SCA47

```
Locus: SCA47
        Probability of observed homozygote class frequency:
                        Binomial based      Rank based
                158     0.0686              0.0405
                162     0.0753              0.0455
                164     0.2115              0.117
                166     0.001               0.001
                168     0.0293              0.0315
                170     0.4266              0.3785
                172     0.4172              0.316
                174     1                   1
                176     0.9999              1
                180     1                   0.001
```

```
Locus: SCA47
        Total expected homozygotes: 128.075
        Total observed homozygotes: 154

        Analysis indicates homozygote excess at this locus.

        No evidence for scoring error due to stuttering.

        No evidence for large allele dropout.

        Null alleles may be present at this locus, as is suggested by the general excess of
        homozygotes for most allele size classes.
```

Enter the Repeat Motif for each Locus: SCA30 Mononucleotide

Maximum Expected Allele Size: 500

Unusual Observations

Confidence Interval: Bonferroni 1000 x

Chapter 7

Discussion - genetic markers applied to fish populations

Macbeth GM

The research results have been discussed in detail and compared with relevant literature at the end of Chapters 2 to 6. More broadly genetic markers have been usefully applied and their utility extended by this research. During the course of this PhD I have been able to explore new and novel ways of applying genetic markers to solve practical problems. I believe the future role of genetic markers applied to monitoring wild fishery stocks and domesticating stocks for aquaculture will expand even further to support global food security.

7.1 Major findings and implications

The major findings and implications discussed in this thesis on improving the utility of genetic markers in fish populations include:

- Genetic markers can be used at the onset of a fish breeding program to improve genetic gains for growth by an impressive 40% during a two year implementation phase. This improvement in growth is commercially significant and would take an estimated nine to 22 years to achieve in barramundi (*L. calcarifer*) using traditional selection methodology.
- Genetic markers can be used to determine rapid estimates of genotype by environmental (GxE) interactions in aquaculture populations. Rapid estimates of GxE are essential for the economic evaluation of national fish breeding programs where fish are grown under a different set of growing conditions. Rapid estimates of GxE are also useful in assessing new feed formulations.
- Genetic markers can be used to determine individual matches which are facilitated by the simultaneous estimation of Type I and Type II errors. The number of pairwise comparisons can increase by 80% compared to discarding genotypes with missing data. Wildlife forensic studies can now utilise more data in their studies as a result of this significant development.
- Genetic markers can be used to determine finite estimates of the number of wild fish in a fishing zone and the proportion of those fish caught (catchability). Catchability estimates have been used in stock assessment but the estimation of this parameter has been difficult to achieve using traditional catch data statistics. As many wild fish populations are increasingly being over-exploited this improved methodology offers a new effective monitoring tool for fisheries management.
- Effective population size is a measure of the health of breeding populations and is related to abundance. Genetic markers can be used to estimate effective population

size in wild fish populations. The simulations conclusively demonstrated that the LD method was highly sensitive to low numbers of transient individuals from spatially and genetically distinct populations being sampled within the local population being studied. Methods were devised to remove transient genotypes from Spanish mackerel data which then enabled finite estimates of the lower confidence limit of effective population size to be determined.

7.2 Caveats and future research

7.2.1 Caveats and future research - Chapter 2

A novel breeding program for improved growth in barramundi *Lates calcarifer* (Bloch) using foundation stock from progeny-tested parents.

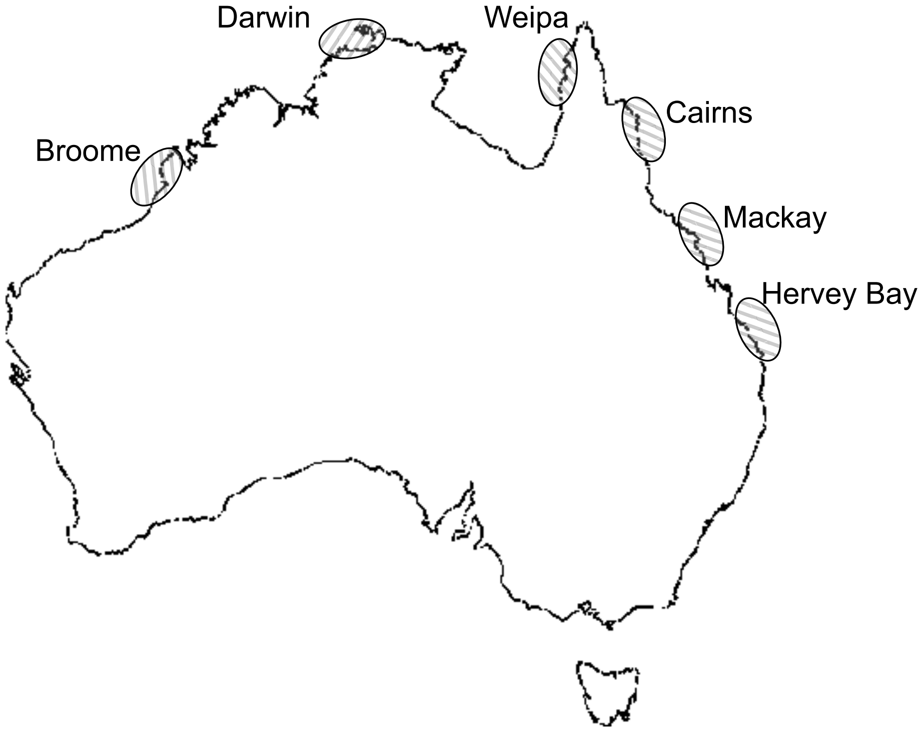
Genetic markers were used in this study as a means of identifying the sires of progeny genotyped. While the method of genotype analysis was not specified it is likely that microsatellite markers or SNP markers would be used for sire identification. Microsatellite markers have been extensively used for parentage assignment (Webster et al., 2000; Visscher et al., 2002). More recently SNP markers are gaining interest due to their abundance, low genotype error rates and potential for automation (Heaton et al., 2002; Fernandez et al., 2013; Anderson and Garza 2006). Other emerging genetic marker technologies may be used to identify the sires as this will have no bearing on the outcome of this study. As the dam and all sires can be genotyped it is a relatively straight forward task to identify the sires of each fish genotyped.

The quantitative genetic theory used in this study is well understood, proven over time and robust. Using a binary threshold to estimate breeding values for growth rate may have some critics. For example, not getting sufficient family representation in the top fish genotyped is one such criticism. This is not seen as a problem as only the best sires need to be identified for the proposed rapid genetic gains to be realised.

One assumption in the model is that all sires are unrelated. In reality when collecting and storing semen from wild males no guarantee can be made that the males are all unrelated. To reduce the risk of relatedness between samples the males could be collected over multiple spawning seasons during which milt from males can be harvested and at different locations (Figure 7.1). The locations were based on different estuary systems with evidence of genetic divergence (Jerry et al., 2013) with the locations also having vehicle access for transport of live fish. As reviewed in Chapter 1 the use genetic markers for relationship estimation may improve the accuracy of variance component estimation.

Likewise in trials that include semen from existing breeding programs the pedigree structure should be included in the analysis. This would mean that a program like ASREML (Gilmour 2001) would be required which utilises the numerator relationship matrix for variance component estimation.

Figure 7.1 Suggested locations for wild males and milt to be collected within Australia.



In practice, one difficulty in setting up a breeding program in barramundi is controlling sex reversal from males to females. In caged barramundi 45% had changed sex within 3 years of age (Guiguen et al., 1994) while sex reversal may occur as early as two years of age (Davis, 1982; Moore, 1979). With adequate redundancy in breeding stock rapid gains through selective breeding may be a challenge but achievable. The potential risks inherent to the managment of a distinct breeding strategy, such as male broodstock changing sex to females prior to contributing to the next generation, must be carefully considered, and are ideally accompanied by a risk management strategy which might for example foresee to keep cryopreserved milt of stock from current generations.

A current focus of research is controlling sex reversal in barramundi. As more data becomes available on sex reversal in barramundi under different environments it may, in future, be possible to calculate the necessary redundancy that may be required to ensure the required crosses between given family lines can be achieved. These mating designs are important to sustain low levels of inbreeding to ensure continued genetic progress can be made in future generations.

7.2.2 Caveats and future research - Chapter 3

> Rapid assessment of genotype by environmental interactions and heritability for growth rate in aquaculture species using *in vitro* fertilisation and DNA tagging.

Similar to Chapter 2 genetic markers to identify sires of fish from an artificial mating design using cryopreserved sperm to fertilise eggs from a single female is simulated. In this design however fish are randomly allocated in two different environments to assess genotype by environmental interactions (GxE). These GxE interactions are important. For example if fish are selected for breeding (e.g. for fast growth rate) in one environment and the fish from the breeding program are commercially grown in a second environment then the GxE will determine what proportion of genetic gains expected from the breeding program are expressed in the second environment.

The design which assumed equal heritability between environments allowed a simple univariate model which improved computer performance compared to bivariate analysis requiring the numerator relationship matrix (Faux and Gengler, 2013). This significant improvement in speed allowed simulations to be conducted in a timely manner. When analysing real experimental data a bivariate analysis would be recommended so that any potential difference in heritability between environments could be assessed.

The univariate analysis created an unforseen problem when analysing phenotypic data. For example in a design with only one dam and multiple sires it was important to include a scaling effect term in the statistical model. If the fixed term for environment is not in the statistical model the genetic sire variance between environments is inflated by the differences in the maternal genetic effect in each environment. As a test when using a full-factorial design with 50 dams no scaling effect was necessary as the average expectation of the maternal genetic effects within each environment were approximately equal. In practice the scaling effect of growth performance between environments can also be caused by non-genetic environmental effects. The caveat is that the univariate analysis of phenotypic data should always include a fixed effect term for environment in the model even if multiple dams are used. The fixed effect term for environment is not required in binomial data as scaling effects of performance are not observed.

Rapid estimates of GxE interactions have practical applications in assessing the effect of selection response not only in different physical environments but also to new formulated feeds. The replacement of fishmeal with land-based feeds is an important area of research (Rossi et al., 2013) not only for economic reasons but to reduce the environmental impact of fishmeal replacement. This study has the potential to assist in future research to improve the environmental sustainability of aquaculture.

7.2.3 Caveats and future research - Chapter 4

Likelihood-based genetic mark–recapture estimates when genotype samples are incomplete and contain typing errors.

As applied in Chapters 2 and 3 the use of genetic markers to identify sires is relatively straight forward. A difficulty arises when genotype samples are collected in the field under harsh tropical conditions with degraded DNA. Microsatellite markers are prone to loosing data in the form of null alleles artificially increasing the number of homozygotes where one allele drops out or in the case of two allele dropouts at a given locus forming a missing data point at that locus. The motivation for this study was to determine recaptures in Spanish mackerel (*S. commerson*) from microsatellite markers (Chapter 5) but before that could be achieved the mathematical theory had to be developed.

Testing of the theory led to the development of the program called SHAZA (Macbeth et al., 2011) where the writer's skills in computer programming were employed. In wildlife forensics budgets are typically constrained so all data should be utilised whenever possible. Unless there are valid statistical reasons, missing data should not be disregarded. The new methodology implemented in SHAZA allows all genotypes available to be implemented in genetic identification studies.

There are two potential caveats in SHAZA. The first is the assumption of an outbred population. As mentioned earlier this could be alleviated by fixing the number of Type I errors to a conservatively low level (*e.g.* V=0.01). It was also mentioned that if a kinship structure of a population is known then the ratio of Type I and Type II errors could be adjusted to provide unbiased estimates of recaptures. Knowing that false positives increase with an increased relationship structure in populations it is probably worthy of further research to investigate a new method that can estimate the proportion of full-sibs and possibly the proportion of half-sibs in genotype data (4.11 Appendix II – Percentage sibship estimation). This area of research seems rewarding as sibship estimates can be determined from genotype data which does not have sufficient genotype information to assign individual relationships which standard sibship analysis require (Wang, 2004).

The second caveat is that SHAZA corrects for missing data assuming an underlying binomial distribution using equation 4.5. This has shown to work extremely well when the chance of finding a recapture is low. For example the total number of pairwise comparisons with given sample size in *D* is: *D*(*D*-1)/2. When the number of recaptures is greater than *D* the binomial assumption starts to break down as clusters of recaptures form. This aspect needs more research and in practice has occurred in a study involving scat samples where genetic matches ('recaptures') were common.

7.2.4 Caveats and future research - Chapter 5

How many fish under the boat? Estimating abundance of narrow-barred Spanish mackerel (*Scomberomorus commerson*) using a genetic mark-recapture approach.

The theory developed in Chapter 4 is applied with real data using narrow-barred Spanish mackerel (*S. commerson*) genotypes. The microsatellite genotypes were collected in harsh field conditions where null alleles were common. The program SHAZA retrieved 80% more pairwise markers than what would have been achieved by disregarding genotypes will null alleles.

As a caveat of this study an assumption was made that the population was outbred. Apart from the large effective population size of Spanish mackerel (Macbeth et al., 2012) there is some additional evidence that suggests that the Spanish mackerel population is indeed an outbred population with few full-sibs. All of the 113 putative paired matches detected by SHAZA were within the same bag with the first known false positive (between bag match) occurring at the 128th highest likelihood ranked match. When estimating recaptures ($\hat{R}$) between fins from different collection bags it would be expected that $\hat{R}$ would approximate the defined cumulative false positive sum (*V*). When matching genotypes between sample bags and setting the false positive threshold to *V*=10 or *V*= 20 the corrected number of recaptures ($\hat{R}$) were 10.3 and 21.7 matches respectively. This provides support that the model was working as expected with $V \approx \hat{R}$ when no true recaptures were present. If there were many full-sib relationships in the data the expectation is that $\hat{R}$ would have been much larger than *V* and therefore it seems reasonable to suggest that Spanish mackerel genotypes were sampled from an outbred population.

To be more precise the estimate of abundance in this study reflects the abundance of wild feeding fish as there is no way of estimating the percentage of non-feeding fish. In future it may be possible to estimate fish density however more research is required to estimate the width of the swept area during sampling which is currently unknown.

It remains to be seen if the cumulative cost of estimating abundance using genetic mark-recapture methods could be better spent on genome sequencing the commercial species of interest so that accurate estimates of linkage disequilibrium effective population size can be determined.

7.2.5 Caveats and future research - Chapter 6

Linkage disequilibrium estimation of effective population size in Spanish mackerel (*Scomberomorus commerson*) with immigrants from divergent populations.

The limitations of estimating effective population size was tested in real data with seven polymorphic microsatellite loci. With the large population size of Spanish mackerel it was difficult to find finite confidence limits with the seven polymorphic loci used in this study. Although finite estimates of the lower 95% confidence interval were obtained the number of polymorphic loci were too few to determine finite estimates of the mean expectation and upper 95% confidence interval.

In future, single base pair mutations at a specific locus, called SNP markers, will be more likely to produce finite estimates of effective population size with much larger population sizes. This is because SNP markers are more plentiful, despite them not being as polymorphic as microsatellite data. The limitations of this study need to be studied further with future research investigating the limits of precision when using more loci. It is clear that future LD estimates of effective population size will use more markers.

While too expensive for current wildlife studies the future direction can be seen in species that have the full genome sequenced with contemporary and historical estimates of effective population estimated. Historical estimates are possible as the crossovers along the chromosome during meiosis are less likely to occur close to each other than further along the chromosome (Badke et al., 2012; Kim and Kirkpatrick, 2009) with closely linked markers likely to be more correlated in large populations (Waples, 2006). Also Hill (1981) suggested that linkage disequilibrium estimates of effective population size from closely linked markers would reflect ancient population history. Using SNP markers that are different distances from each other along the genome it is possible to estimate historical effective population over different time periods. This has been demonstrated in dairy cattle (Kim and Kirkpatrick, 2009; Shin et al., 2013), Swiss cattle (Flury et al., 2010), equine data (Corbin, 2012) and humans (Li and Durbin, 2011). Historical estimates would reflect carrying capacity estimates of the fishery ecosystem with future research potentially able to estimate maximum sustainable yields under simulated fish stock recovery programs.

Entire genome sequences are still not feasible in most fisheries applications and microsatellite markers are still expected to dominate linkage disequilibrium estimates of effective population size in the immediate future. The prospect of using genome wide estimates for fisheries management in high valued fisheries seems inevitable. The caveat is that effective population size is not a direct estimate of abundance and the relationship between them is still an area of active research (Waples et al., 2014, Dudgeon and Ovenden, 2015).

7.3 Conclusions

Genetic markers are a versatile tool in aquaculture and fisheries management. In the relatively short period of this PhD the utility of genetic markers have been developed and improved in a number of practical applications applied to fish populations. These applications covering selective breeding, variance component estimation (genetic correlations and heritability), improvements in genotype assignment of individuals, new methods of wild fisheries abundance estimates and improving the accuracy of effective population size estimates. The improvements made in the utility of genetic markers during this PhD is evidence that genetic marker research has not matured and that there is much scope for future research. This future appears strong as the growing applications of genetic markers in aquaculture and fisheries will continue to develop as the cost of genotyping falls.

Some areas where future research can be directed include:

(i) following up on the theory of achieving a predicted 40% improvement in growth rate in barramundi (*L. calcarifer*) and putting the design into practice,

(ii) the estimation of the percentage of full sibs, and possibly half sibs, in populations using false positive counts from likelihood pairwise matches (Figure 4.6),

(iii) the development of the SHAZA program to include simultaneous estimates of pedigree relationships and recapture estimates by implementing theory developed in area (ii) above,

(iv) investigate the feasibility of estimating fish density using the genetic mark-recapture abundance estimates,

(v) power testing of the linkage disequilibrium estimation of effective population size with large population sizes and more genotype data to determine finite upper limits of effective population size and

(vi) using historical effective population size estimates as an estimate of maximum carrying capacity with inferences on maximum sustainable yield.

7.4 References

Anderson EC, Garza JC (2006) The power of single-nucleotide polymorphisms for large-scale parentage inference. Genetics 172:2567-2582.

Badke YM, Bates RO, Ernst CW, Schwab C, Steibel JP (2012) Estimation of linkage disequilibrium in four US pig breeds. BMC Genomics 13:24 doi:10.1186/1471-2164-13-24

Corbin LJ, Liu AYH, Biship SC, Woolliams JA (2012) Estimation of historical effective population size using linkage disequilibria with marker data. Journal of Animal Breeding and Genetics 129:257-270.

Davis T (1982) Maturity and sexuality in barramundi, *Lates calcarifer* (Bloch), in the Northern Territory and south-eastern Gulf of Carpentaria. Marine and Freshwater Research 33:529-545.

Dudgeon CL, Ovenden JR (2015) The relationship between abundance and genetic effective population size in elasmobranchs: an example from the globally threatened zebra shark *Stegostoma fasciatum* within its protected range. Conservation Genetics doi:10.1007/s10592-015-0752-y

Faux P, Gengler N (2013) Inversion of a part of the numerator relationship matrix using pedigree information. Genetics Selection and Evolution 45:45. doi:10.1186/1297-9686-45-45

Fernández ME, Goszczynski DE, Lirón JP, Villegas-Castagnasso EE, Carino MH, Ripoli MV, Rogberg-Muñoz A, Posik DM, Peral-Garcia P, Giovambattista G. (2013) Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, tracability and assessment of parentage in an inbred Angus herd. Genetics and Molecular Biology 36:185-191.

Flury C, Tapio M, Sonstegard T, Drögemüller C, Leeb T, Simianer H, Hanotte O, Rieder S (2010) Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium. Journal of Animal Breeding and Genetics 127:339–347.

Gilmour AR, Cullis BR, Welham SJ, Thompson R (2001) ASREML User’s Manual. New South Wales Agriculture, Orange Agricultural Institute, Orange, NSW, Australia.

Guiguen Y, Cauty C, Fostier A, Fuchs A, Jalabert B (1994) Reproductive cycle and sex inversion of the seabas, Lates calcarifer, reared in sea cages in French Polynesia: histological and morphometric description. Environmental Biology of Fishes 39: 231-247.

Heaton MP, Harhay GP, Bennett GL, Stone RT, Grosse WM, Casas E, Keele JW, Smith TPL, Chitko-McKown CG, Laegreid WW (2002) Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. Mammaliam Genome 13:272-281.

Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium. Genetical Research 38:209-216.

Jerry DR, Smith-Keune C, Hodgson L, Pirozzi I, Carton AG, Hutson KS, Brazenor AK, Gonzalez AT, Gamble S, Collins G, VanDerWal J (2013) Vulnerability of an iconic Australian finfish (Barramundi – *Lates calcarifer*) and aligned industries to climate change across tropical Australia. Final Report Proect No. 2010/521 Fisheries Research Development Corporation.

Kim E-S, Kirkpatrick BW (2009) Linkage disequilibrium in the North American Holstein population. Animal Genetics 40:279-288.

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493-497.

Macbeth GM, Broderick D, Ovenden JR, Buckworth RC (2011) Likelihood-based genetic mark–recapture estimates when genotype samples are incomplete and contain typing errors. Theoretical Population Biology 80:185-196.

Macbeth GM, Broderick D, Buckworth RC, Ovenden JR (2012) Linkage disequilibrium estimation of effective population size in Spanish mackerel (*Scomberomorus commerson*) with immigrants from divergent populations. Genes Genomes and Genetics 3:709-717.

Moore R (1979) Natural sex inversion in the giant perch (*Lates calcarifer*). Marine and Freshwater Research 30:803-813.

Rossi Jr W, Moxely D, Buentello A, Pohlenz C. Gatlin III DM (2013) Replacement of fishmeal with novel plant feedstuffs in the diet of red drum *Sciaenops ocellatus*: an assessment of nutritional value. Aquaculture Nutrition 19:72-81.

Shin D-H, Cho K-H, Park K-D, Lee –J, Kim H (2013) Accurate estimation of effective population size in the Korean dairy cattle based on linkage disequilibrium corrected by genomic relationship matrix. Asian-Australasian Journal of Animal Science 26:1672-1679.

Visscher PM, Woolliams JA, Smith D, Williams JL (2002) Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection, Journal of Dairy Science 85:2368-2375.

Wang J (2004) Sibship reconstruction from data with typing errors. Genetics 166:1963-1979.

Waples RS (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conservation Genetics 8:167-184.

Waples RS, Tiago A, Luikart G (2014) Effects of overlapping generations on linkage disequilibrium estimates of effective population size. Genetics 197:769-780.

Webster MS, Chuang-Dobbs HC, Holmes RT (2000) Microsatellite identification of extrapair sires in a socially monogamous warbler. Behavioral Ecology 12:439-446.