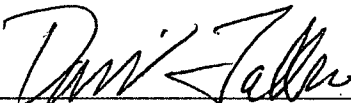


POPULATION GENETICS AND MIXED STOCK ANALYSIS OF CHUM
SALMON (ONCORHYNCHUS KETA) WITH MOLECULAR GENETICS

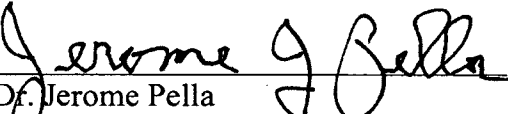
By

Michael R. Garvin

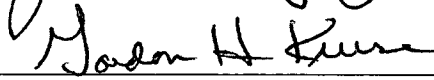
RECOMMENDED:



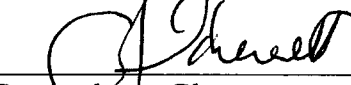
Dr. David Tallmon



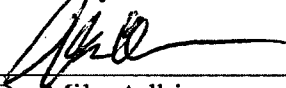
Dr. Jerome Pella



Dr. Gordon Kruse

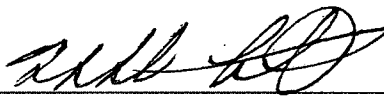


Dr. Anthony Gharrett
Advisory Committee Chair

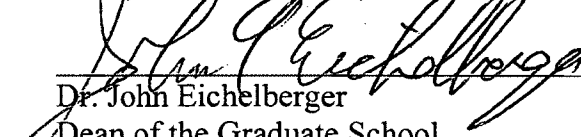


Dr. Milo Adkison
Chair, Graduate Program in Fisheries Division

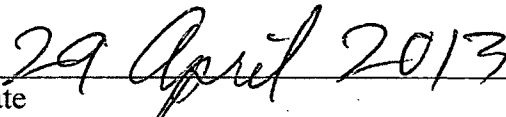
APPROVED:



Dr. Michael Castellini
Dean, School of Fisheries and Ocean Sciences



Dr. John Eichelberger
Dean of the Graduate School



Date

POPULATION GENETICS AND MIXED STOCK ANALYSIS OF CHUM SALMON
(ONCORHYNCHUS KETA) WITH MOLECULAR GENETICS

A
THESIS

Presented to the Faculty
of the University of Alaska Fairbanks

in Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

By

Michael R. Garvin, M.S.

Fairbanks, Alaska

December 2012

UMI Number: 3573014

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3573014

Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Chum salmon (*Oncorhynchus keta*) are important for subsistence and commercial harvest in Alaska. Variability of returns to western Alaskan drainages that caused economic hardship for stakeholders has led to speculation about reasons, which may include both anthropogenic and environmental causes in the marine environment.

Mixed stock analysis (MSA) compares genetic information from an individual caught at sea to a reference baseline of genotypes to assign it to its population of origin. Application of genetic baselines requires several complex steps that can introduce bias. The bias may reduce the accuracy of MSA and result in overly-optimistic evaluations of baselines. Moreover, some applications that minimize bias cannot use informative haploid mitochondrial variation. Costs of baseline development are species-specific and difficult to predict. Finally, because populations of western Alaskan chum salmon demonstrate weak genetic divergence, samples from mixtures cannot be accurately assigned to a population of origin.

The chapters of this thesis address three challenges. The first chapter describes technical aspects of genetic marker development. The second chapter describes a method to evaluate the precision and accuracy of a genetic baseline that accepts any type of data and reduces bias that may have been introduced during baseline development. This chapter also includes a method that places a cost on baseline development by predicting the number of markers needed to achieve a given accuracy. The final chapter explores the reasons for the weak genetic structure of western Alaskan chum salmon populations. The results of those analyses and both geological and archaeological data

suggest that recent environmental and geological processes may be involved. The methods and analyses in this thesis can be applied to any species and may be particularly useful for other western Alaskan species.

Table of Contents

| | Page |
|---|------|
| Signature page..... | i |
| Title page | ii |
| Abstract..... | iii |
| Table of Contents..... | v |
| List of Figures..... | ix |
| List of Tables | xi |
| Dedication and Acknowledgements | xii |
| General Introduction..... | 1 |
| Chapter 1..... | 8 |
| Abstract..... | 9 |
| Introduction | 10 |
| Single nucleotide polymorphism discovery | 13 |
| Sanger Sequencing..... | 15 |
| Next Generation Sequencing..... | 18 |
| Restriction-site associated DNA (RAD) markers | 23 |
| TILLING | 24 |
| Use of previously published data | 26 |
| Single nucleotide polymorphism genotyping assays..... | 27 |
| Criteria for choosing SNP genotyping assays..... | 28 |

| | |
|---|------------|
| Linked single nucleotide polymorphisms: The power of haplotypes | 33 |
| An analysis of SNP discovery in non-model species | 36 |
| Summary/Conclusion | 38 |
| Acknowledgements | 40 |
| Figures and Tables..... | 41 |
| References | 46 |
| Chapter 2..... | 74 |
| Abstract..... | 75 |
| Introduction | 76 |
| Materials and Methods | 85 |
| Generation of test baselines and stock mixtures | 85 |
| Measurements of diversity | 87 |
| Baseline summary statistics | 88 |
| BAYES LTO..... | 91 |
| SPAM LTO | 92 |
| SPAM simulation | 92 |
| Simulated SNP loci: how many SNPs would equal the discriminatory power of the combined SNP and microsatellite baseline | 93 |
| Results | 96 |
| Evaluation of combined microsatellite and SNP baseline | 96 |
| Evaluation of the bias and accuracy of estimates with BAYES and SPAM..... | 98 |
| Evaluation of potential bias in baseline..... | 100 |
| How many SNP loci are needed to exceed or equal the combined SNP and microsatellite baseline?..... | 101 |

| | |
|---|------------|
| Discussion..... | 102 |
| Acknowledgements | 109 |
| Figures and Tables..... | 110 |
| References | 128 |
| Chapter 3..... | 137 |
| Abstract..... | 138 |
| Introduction | 139 |
| Materials and Methods | 143 |
| Populations and genotype data..... | 143 |
| Hierarchical G-test | 143 |
| Measures of divergence..... | 144 |
| Principal components analysis (PCA)..... | 145 |
| Trees | 145 |
| Outlier analysis..... | 145 |
| Isolation by distance..... | 146 |
| Dispersal distance..... | 148 |
| Results | 150 |
| Measures of divergence..... | 150 |
| Principal components analysis | 151 |
| Neighbor-joining trees..... | 151 |
| Outlier analysis..... | 152 |
| Isolation by distance..... | 152 |
| Dispersal distance on the Kuskokwim River | 153 |
| Discussion..... | 154 |

| | |
|---------------------------|-----|
| Acknowledgements | 163 |
| Figures and Tables..... | 164 |
| References | 179 |
| General Conclusions | 190 |
| General References | 192 |

List of Figures

| | Page |
|---|------|
| Figure 1.1 The importance of the SNP discovery scheme..... | 41 |
| Figure 1.2 The improved T_m -shift assay | 42 |
| Figure 1.3 Linkage phase resolution..... | 43 |
| Figure 2.1 Area of study | 110 |
| Figure 2.2a Mean stock composition estimates for 25 reporting groups..... | 111 |
| Figure 2.2b BAYES LTO for 25 reporting groups..... | 112 |
| Figure 2.2c SPAM LTO for 25 reporting groups | 113 |
| Figure 2.2d SPAM simulations for 25 reporting groups..... | 114 |
| Figure 2.3a Mean stock mixture estimates for 14 reporting groups | 115 |
| Figure 2.3b BAYES LTO for 14 reporting groups..... | 116 |
| Figure 2.3c SPAM LTO for 14 reporting groups | 117 |
| Figure 2.3d SPAM simulations for 14 reporting groups..... | 118 |
| Figure 2.4 Bias in stock proportion estimates..... | 119 |
| Figure 2.5 Standard deviation of stock proportion estimates | 120 |
| Figure 2.6 Mean squared error ^{1/2} of the stock proportion estimates..... | 121 |
| Figure 2.7 G_{ST} values to assess possible introduction of bias..... | 122 |
| Figure 2.8 Mean proportion of mixture individuals correctly assigned..... | 123 |
| Figure 3.1a Area of study..... | 164 |
| Figure 3.1b Historical Yukon-Kuskokwim connections | 165 |
| Figure 3.1c Historical Kuskokwim-Nushagak connections..... | 166 |

| | |
|--|-----|
| Figure 3.2 PCA of western Alaskan chum salmon populations | 167 |
| Figure 3.3a The neighbor-joining tree drawn with genetic chord distances..... | 168 |
| Figure 3.3b Consensus neighbor-joining tree | 169 |
| Figure 3.4 Outlier analysis..... | 170 |
| Figure 3.5a IBD on Yukon and Kuskokwim with present geographical distances | 171 |
| Figure 3.5b IBD on Yukon and Kuskokwim with past geographical distances..... | 171 |
| Figure 3.6a IBD on Kuskokwim and Nushagak with present geographical distances | 172 |
| Figure 3.6b IBD on Kuskokwim and Nushagak with past geographical distances | 172 |

List of Tables

| | Page |
|--|------|
| Table 1.1 Summary of SNP discovery efforts in non-model organisms | 43 |
| Table 2.1 Stocks used in the study and their sources..... | 124 |
| Table 2.2 Measures of genetic diversity | 126 |
| Table 2.3 Sum of mean error for three methods | 127 |
| Table 3.1 Geographical and run timing information of the samples | 173 |
| Table 3.2 Summary statistics of the loci..... | 174 |
| Table 3.3 Log-likelihood ratio (G) tests for the [25P70L data set] | 176 |
| Table 3.4 Log-likelihood ratio (G) tests for the [21P50L data set] | 177 |
| Table 3.5 Dispersal distances of chum salmon..... | 178 |

Dedication and Acknowledgements

*Phlebas the Phoenician, a fortnight dead,
 Forgot the cry of gulls, and the deep sea swell
 And the profit and loss.
 A current under sea
 Picked his bones in whispers. As he rose and fell
 He passes the stages of his age and youth
 Entering the whirlpool.
 Gentile or Jew
 O you who turn the wheel and look windward,
 Consider Phlebas, who was once handsome and tall as you.*

-T.S Elliot

As I sit here at my family's house in Casco Bay on Lake Coeur d' Alene staring at the water, I recall hours spent fishing for west-slope cutthroat trout and running through ponderosa pine-scented woods in bare feet that were calloused from a summer of living with no shoes. I can't help but think that I was insane to spend the last eight years trying to understand the complexities of salmon evolution and molecular genetics. This isn't helped by the fact that I'm broke, 43 years old, pretty much every possession I own is falling apart, and I am walking into one of the worst job markets since the Great Depression. What the hell was I thinking?

I'm sure every graduate student asks themselves that question and I have no idea what their answers are, but I have come to realize that it's really about two things: belly grabbing laughter and intense frustration. The relationships that I established in the last eight years – friend and foe alike – have elicited both of those reactions. It's probably why I am still young at heart and smiling, but have lost a considerable amount of hair. What's left is gray. Because of that laughter and that frustration (the images in my mind

are vivid), I was able to fully explore the talents that I brought into this world, and was able to develop skills that I needed but definitely lacked when I set out. Both of these endeavors have been equally rewarding.

First and foremost I would like to thank my committee. Their deep knowledge spans a significant breadth of science. My advisor, Tony Gharrett's abilities are probably best described by the many students that he has produced during his tenure. They work for institutions throughout Alaska and the Pacific Northwest and continue to heavily influence conservation and management of various species. Tony constantly challenged me to improve whatever it was I was working on – frustrating for sure, but the finished product was always infinitely better and in the end I learned much. As a matter of fact, I wonder if I will be as productive during the next stage of my career without him there to challenge me. Dave Tallmon more than anything has been a role model for me. A successful and internationally known geneticist (who is my age) has encouraged me to look outside of Alaska and outside of my field of salmon genetics to see things from a different view and to question assumptions that I made but failed to see. Gordon Kruse is probably one of the hardest working people I have ever met. It's hard to find him among the stacks of papers in his office and I've no idea how he absorbs all of that information. I came to Alaska to obtain a PhD in Fisheries. I always complained that most of my classes were about statistics and not fish. Gordon provided focus there. His comprehensive exam questions and preparing for this defense made me explore the bigger picture – how is what I am doing relevant to fisheries? Sometimes I get lost amongst the pipets and the PCR tubes. Finally Jerry Pella, who essentially invented the

field of mixed stock analysis, probably taught me the most because his field was the most difficult for me to understand and because he is a fantastic teacher with a great sense of humor. If you don't understand something the first time, he explains it another way. And he will keep going until you get it.

I would love to list all of the other important people that have been part of this fabulous journey because they are the ones who got me here. Alone I would never have accomplished this much. Unfortunately it would equal the length of this thesis and I doubt I could explain the complexities of each of those relationships. But they have been invaluable. In the end, it is what I walk away with. Maybe I'll write a book some day.

General Introduction

In this doctoral dissertation I address several challenges that have emerged from efforts to explain recent declines of chum salmon (*Oncorhynchus keta*) that resulted in disaster declarations in western Alaskan by the Governor of Alaska (Zimmerman 2006). Chum salmon have the broadest distribution of any species of Pacific salmon (Augerot 2005). They are an important component of subsistence use by rural Alaskans and represent a substantial commercial fishery for some of those same communities (Wolfe & Spaeder 2009). Recent increased demands for wild Alaskan salmon products, which include the roe from chum salmon, have led to increased profits to the commercial sector outside of rural communities as well (Knapp 2007). Much like in financial markets, decreases in returns cause duress among the participants and commonly initiate calls to understand the basis for the losses; and if the variables that are responsible for both increases and decreases in abundances can be identified, it may allow stakeholders to buffer against hardship when returns are low and maximize the returns when they are high.

Studies that use commercial catch as an indicator of salmon biomass indicate that abundances of chum salmon have varied by more than 3-fold in the North Pacific Ocean on a decadal scale (Beamish & Bouillon 1993; Eggers 2009). Commercial catches from the Yukon River varied from 28 800 to well over 1 000 000 in less than two decades (Kruse 1998). Salmon returns are typically predicted from historical spawner and recruitment data (Ricker 1954), which can provide valuable information but lacks information from the ocean phase of anadromous salmon. Much of a salmon's success (return to spawn) or failure (ocean mortality) is tied to the productivity of the marine

environment. Mortality can also result when commercial fisheries catch one species when they are targeting another (bycatch). Recently, the bycatch of chum salmon by the Bering Sea groundfish fishery (Gisclair 2009) and sockeye salmon (*O. nerka*) fisheries (Seeb et al. 2004) has caused concern because chum salmon abundances for some western Alaskan drainages have decreased at the same time that bycatch in those fisheries has increased.

Mixed-stock analysis (MSA) is a tool that is gaining wide use for conservation and management of species worldwide. It uses morphological (phenotypic) or genetic (genotypic) data from individuals (usually) comprised of a mixture of populations, either to assign individuals from that mixture to a population of origin or to estimate the composition of the mixture from source populations. The method relies on a baseline of genotypes, phenotypes, or both that represents the geographic range of the target species and provides a reference to compare with the samples from the mixtures (unknowns). This method is ideal for addressing many conservation and management questions because samples can be taken non-lethally and individuals can oftentimes be assigned to a population or region of origin with high accuracy. For Pacific salmon, if the assignment with MSA is accurate, then the abundances of specific stocks of salmon returns can be correlated with environmental and anthropogenic variables in the ocean phase of their life-history.

Mixed-stock analysis has typically been performed with data from genetic markers called microsatellites (e.g. Beacham et al. 2009), which likely result from strand slippage during DNA replication (Hancock 1999). Recently, much effort has been

devoted to the use of genetic markers called single nucleotide polymorphisms (SNPs) because these markers resolve one of the major challenges to management that is based on genetic data: sharing information among users. The scoring of microsatellite data is often lab-specific, which is inconsequential if comparisons among data sets are within that laboratory; but if comparisons are made between laboratories, then standardizations are necessary. This becomes a problem as the number of users and the sizes of the databases increase. The current chum salmon baseline includes tens of thousands of samples taken from populations that span the entire Pacific Rim from Korea to Oregon. The data from SNP markers is typically binary and, therefore, simple to share among laboratories but necessitates many SNPs to equal the power of a single microsatellite. SNPs have the advantages of being distributed throughout the genome and are plentiful, which has led most labs to begin to establish SNP baselines for MSA.

The development of a program to perform MSA requires three main steps: (1) discovery of genetic markers, (2) use of those markers to genotype individuals from samples that will be used for the reference baseline, and (3) evaluation of the baseline to determine its accuracy and precision for MSA. The first two chapters of this PhD thesis address these three criteria of baseline development. Chapter 1 is a review article that I was invited to write by the editors of *Molecular Ecology Resources* in 2010. A manuscript that I wrote for my Master's thesis stimulated their interest in the technical aspects of genetic marker development. The majority of the work for chapter 1 was an extensive search of the literature for the current state of genetic marker development that also drew on my background in technology development when I worked in the

biotechnology industry and also technology that I developed in A.J. Gharrett's laboratory during my Master's degree. The third co-author, K. Saitoh, collaborated on some of the technology assessment within the document.

The second chapter of this thesis is focused on baseline evaluation, but addresses SNP discovery and genotyping in a broader context because two publications by Eric Anderson from the Southwest Fisheries Science Center in Santa Cruz, CA, and colleagues described several sources of bias that can be introduced into the baseline development process (Anderson 2010; Anderson et al. 2008) and can severely affect MSA efforts. The main impetus behind chapter 2 stemmed from the fact that computer programs that are currently available to evaluate genetic baselines either give overly-optimistic assessments, or they do not accept haploid mitochondrial data.

I converted RFLP data that was obtained previously by D. Churikov in A. J. Gharrett's lab to SNP assays, which demonstrated that mitochondrial DNA variation in chum salmon is informative and should be used for MSA applications (Garvin et al. 2010). In addition, my advisor A.J. Gharrett, our co-author J. P. Bielawski of Dalhousie University in Nova Scotia and I found that the mitochondrial genomes of Pacific salmon may have experienced positive Darwinian selection during their evolution (Garvin et al. 2011). If the mitochondrial genomes experienced positive selection intra-specifically, then some variants may show divergence at the population level where other neutral markers do not. This could be invaluable for stocks that need to be separated for management or conservation but look similar with neutral markers.

As stated earlier, there is currently no software package available to evaluate a genetic baseline that contains haploid data. Therefore, I worked closely with J. Pella and M. Masuda from the National Marine Fisheries Auke Bay Laboratories to develop a new method to perform these analyses. Much of the statistical framework was developed from previously published articles from J. Pella and M. Masuda. I developed code in the R environment to integrate their methods with my baseline development project; and the final product resulted from several meetings between myself, A. J. Gharrett, J. Pella, and M. Masuda to troubleshoot the process as it evolved. The manuscript in this thesis is the result of edits that I obtained from co-authors and three reviewers from the Canadian Journal of Fisheries and Aquatic Sciences, who rejected the original manuscript but encouraged resubmission after suggested changes. S. A. Fuller, R. R. Riley, V. Brykov, and R. Wilmot contributed data for the genetic baseline used in the manuscript and provided editorial comments and suggestions.

The final chapter of this thesis addresses an interesting issue that has been challenging to Alaskan MSA efforts for chum salmon. Populations in western Alaska, specifically from southern Norton Sound, the Lower Yukon River, the Kuskokwim River, and northern Bristol Bay are very similar genetically. As a result, any individuals from a geographic area that is the combined size of Idaho, Oregon, and Washington are oftentimes merged into a single reporting group; stocks are assigned to “coastal western Alaska” rather than to specific drainages. This causes difficulties because chum salmon are used differently over this large geographic area. In some places they are used primarily to feed dogs, which are important for bear protection and access to fishing

grounds, whereas in other locations the salmon are used for subsistence and commercial harvest. Because of this weak genetic structure, it is not possible in marine samples of mixtures to identify specific stocks from this large geographic area. Although several theories have been proposed for the weak genetic structure of the chum salmon populations here, none have been explored in detail.

The variation in DNA has been used for decades to explore the demography and structure of populations. As molecular genetics continues to merge with molecular biology, interest has focused on identifying the functional changes that result from mutations that provide the molecular markers used for population genetics analyses. Some molecular markers that are discovered may alter the physiology of individuals that have them, and those changes may provide for adaptation to specific environments. My original idea for the third chapter was to identify genetic markers that may have experienced positive selection with outlier tests; the Kuskokwim River offered a suitable model system because it encompasses multiple habitats within a single drainage and samples were available for numerous populations.

This idea rapidly evolved as genetic data from additional markers and populations became available from several outside sources. In addition to outlier tests, I was able to test several interesting hypotheses that centered on the paleo-geography of western Alaska to explain the weak structure of chum salmon populations. The initial results of that exploration quickly led me to archaeological and oceanographic studies that also provided interesting additional evidence to explain the low genetic divergence in coastal southwestern chum salmon populations. The additional data resulted in the inclusion of

several co-authors. The bulk of the work was developed from ideas discussed between myself and A.J. Gharrett, and some additions developed from discussions with W. Templin at the Alaska Department of Fish and Game. The chapter presented in this thesis is the result of edits from all co-authors and will be submitted for publication in a peer-reviewed journal after my doctoral defense. The results of the third chapter raise some interesting conclusions that may be applied to other species that inhabit western Alaska.

Chapter 1

Application of Single Nucleotide Polymorphisms to Non-model species:

A Technical Review¹

¹ Michael R. Garvin, Kenji Saitoh, and Anthony J. Gharrett. *Molecular Ecology Resources* 10(6): 91-108 2010

Abstract

Single nucleotide polymorphisms (SNPs) have gained wide use in humans and model species and are becoming the marker of choice for applications in other species.

Technology that was developed for work in model species may provide useful tools for SNP discovery and genotyping in non-model organisms. However, SNP discovery can be expensive, labor intensive, and introduce ascertainment bias. In addition, the most efficient approaches to SNP discovery will depend on the research questions that the markers are to resolve as well as the focal species. We discuss advantages and disadvantages of several past and recent technologies for SNP discovery and genotyping and summarize a variety of SNP discovery and genotyping studies in ecology in evolution.

Introduction

Identification of DNA sequence variation with single nucleotide polymorphisms (SNPs) has become a routine application in fields as diverse as human forensics (Weir 2003), crop improvement (Till et al. 2007), marker assisted breeding (Schaeffer 2006), aquaculture (Liu & Cordes 2004), conservation (Seddon et al. 2005), and resource management (Smith et al. 2005a), in addition to the wide use of these markers in humans for diagnostic applications (McCarthy et al. 2008). Single nucleotide polymorphisms are becoming useful markers in ecological and evolutionary studies as well. In a study of white spruce (*Picea glauca*), several SNPs were identified in genes that appeared to be under positive selection, one of which controlled flowering time and reproductive success in another species (Namroud et al. 2008). Similarly, adaptive genetic divergence between seasonal runs of chinook salmon (*Oncorhynchus tshawytscha*) correlated to a SNP in the clock locus, which is involved in regulating the circadian rhythm (O'Malley et al. 2007). In another application, 54 693 SNPs from domestic cattle were used to resolve the phylogeny of 61 Pecoran species, which had proved difficult to determine with other markers (Decker et al. 2009). Willing et al. (2010) used over 1000 SNPs in wild guppies (*Poecilia reticulata*) to demonstrate previously unknown shared ancestry, gene flow, and admixture among populations. They also determined that two loci, which had previously mapped to a QTL that contributed to ornamental traits, were under directional selection. Clearly SNPs present an exciting development for ecological and evolutionary studies of non-model organisms.

Genetic markers that were applied in past studies fall into two general categories: markers that identify anonymous genetic variation and markers that identify genetic variation in specific segments of the genome. Markers in the first category include those that are derived from amplified fragment length polymorphisms (AFLPs) (Vos et al. 1995), random amplification of polymorphic DNA (RAPDs) (Williams et al. 1990), and restriction fragment length polymorphisms (RFLPs) of genomic DNA, which detects variation as anonymous restriction sites. Markers of the second type include allozymes, which indirectly measure variation of DNA in coding regions of proteins, restriction site analysis to identify variation at specific loci in the mitochondrial and chloroplast genomes or specific nuclear genes (Avisé 1994), and microsatellites, which carry variation in numbers of repeats at specific non-coding sequences and have well described mutation mechanisms (Schlotterer & Tautz 1992). Single nucleotide polymorphisms fall into the second category, but are broadly distributed and can represent variation in all of the different genomic regions (coding, non-coding, microsatellite, mitochondrial, and chloroplast DNA). Discussions of the advantages and disadvantages of the various markers have been presented elsewhere (e.g. Sunnucks 2000, Moran 2002, Vignal et al. 2002).

Specific types of molecular markers are appropriate for some, but not all, studies for a variety of reasons that include: their locations in the genome or their roles in gene expression; their mode of mutation or lack of co-dominant expression; or their practicality or economy of application. Single nucleotide polymorphisms can be applied to a wide variety of studies because they are distributed throughout the genome, are

simple to score, can be inexpensive to genotype, and represent co-dominant markers with a simple, well-defined mutation model. For the purpose of this review, we examined three general types of applications.

The first type of application (which will be referred to as 'population genetics' studies) typically seeks information about the demography and structure of populations and requires a random sample of variable loci that represent the entire genome. Applications may include measurement of variation to describe population structure and to estimate effective population size (N_e) (Tenesa et al. 2007), gene flow (Keller et al. 2008) or dispersal (Bensch et al. 2002), population growth or declines (Emerson et al. 2001; Hyten et al. 2006), and inbreeding (Zenger et al. 2007). For these applications, loci are generally expected to conform to neutral expectations and 'outlier' loci that violate neutrality are often removed to improve neutral parameter estimates (Luikart et al. 2003).

The second type of application will be referred to as 'classification' studies because they include efforts that attempt to delineate individuals or groups of individuals from each other, such as studies of cryptic species (Garvin et al. 2011), molecular systematics (Edwards 2009), mixed stock analysis (Smith et al. 2005a; Negrini et al. 2008; Ogden 2008), parentage analysis (Heaton et al. 2002), or identification of mutations that are involved in local adaptation (Namroud et al. 2008). Unlike population genetics applications, classification studies do not require that loci behave as selectively neutral. In fact, loci that experience different selective regimes in different populations may allow greater ability to distinguish individuals from those populations (e.g., Bensch et al. 2002, Smith et al. 2005a).

The third application will be referred to as ‘mapping’ studies because they typically include analyses that attempt to identify chromosomal segments or SNPs that correlate with a phenotypic trait or genetic marker. These studies include identification of quantitative trait loci (QTL) of known phenotypic traits in pedigrees for various studies that include conservation efforts (Boulding et al. 2008) or marker-assisted selection to improve agricultural and aquacultural production (Collard et al. 2005). Mapping studies are similar to classification studies because they seek to identify markers that show divergence among groups (usually phenotypes or QTL), but unlike classification studies, mapping studies seek to identify markers to create a physical genome map or compare markers to a physical genome map.

Given the distribution and abundance of SNPs in most genomes, only a limited subset is practical to genotype. The type of application to which the research question belongs determines which subsets of SNPs are appropriate, and the subset of SNPs that are identified depends in part on the methods used to discover them. Consequently, each study will most likely differ in (1) how the SNPs should be discovered, (2) what platform(s) are efficient for genotyping, and (3) whether linked SNPs (haplotypes) would be useful. This review discusses these three topics, and provides a summary of selected studies that discovered SNPs in non-model species for ecological and evolutionary work.

Single nucleotide polymorphism discovery

The primary goal of SNP discovery is to identify markers that provide genetic variation for resolution of the kinds of questions posed above. A variety of methods have

been used to discover SNPs, but the approach chosen to address a particular question will depend primarily on the economic and technical resources available to the investigator and the amount and type of sequence information that exists for the focal species. However, the strategy that is applied to the discovery process may differ among the three types of applications because the SNPs that are considered 'informative' may differ in each instance.

A poorly conceived SNP discovery scheme may identify SNPs that produce bias in estimates of the parameters for which they were developed (ascertainment bias) (Fisher 1934; Kuhner et al. 2000; Nielsen 2004; Clark et al. 2005; Rosenblum & Novembre 2007; Anderson 2010) or may result in the inclusion of uninformative markers in some analyses (Figure 1.1). Failure to minimize ascertainment bias or introduction of uninformative markers during the discovery process can confound neutral parameter estimates (Wilding et al. 2001; Luikart et al. 2003) and result in wasted time, effort, and resources. Wise choice of a discovery process is one of the most important decisions in initiating a SNP-based study. Also, statistical methods that can account for ascertainment strategies are available, but they do not take into account population structure of non-human organisms, which can be highly structured (Wakeley et al. 2001; Polanski & Kimmel 2003; Nielsen & Signorovitch 2003; Marth et al. 2004; Rosenblum & Novembre 2007). Consequently, efficient SNP discovery schemes must be designed for each specific application. For example, the discovery process for population genetics applications should survey many loci in both coding and non-coding DNA (with the caveat that outlier loci should be evaluated as discussed previously), and it should include

many individuals that represent the breadth of the geographic distribution. Exclusion of genomic regions or omission of individuals based on their geographic location may introduce bias. Alternatively, if SNPs are used for classification applications, the discovery panel should also include many individuals that represent the geographic range of the study, but should not be assembled randomly. Rather, the discovery panel should be assembled so that informative markers can be distinguished from uninformative markers before proceeding to the development of specific marker assays. Markers that are informative for mixed stock analysis and species identification should maximize divergence among groups or targeted geographic regions (Garvin & Gharrett 2007), whereas markers that are informative for parentage analysis should maximize heterozygosity and variability within populations (Chakraborty et al. 1999; Krawczak 1999; Morin et al. 2004). Because mapping studies involve both linkage mapping and linkage disequilibrium mapping, their ascertainment panel should be organized by phenotype or QTL if known. In addition to consideration of the introduction of ascertainment bias, the SNP discovery method should also be chosen based on cost, throughput, effort, and available resources.

Sanger Sequencing

Sanger sequencing has been the workhorse for *de novo* sequencing of genomes and identification of molecular markers. Slab-gel systems such as the LI-COR (Lincoln, NE) DNA analysis system routinely provide data for approximately 900 base pairs, whereas capillary systems such as the Genetic Analyzer from Applied Biosystems (Foster

City, CA) provide slightly shorter reads (with standard capillaries and reagents) but have higher throughput and much of the process can be automated. These technologies have been used successfully to provide SNPs for many non-model organisms, (e.g. Elfstrom et al. 2005; Smith et al. 2005b; Morin et al. 2007; Rosenblum et al. 2007; Ferber et al. 2008; Paduan & Ribolla 2008). However, with mass sequencing of multiple loci, it is not practical to include a sufficiently large number of individuals for an efficient discovery panel (Figure 1.1). Although costs have declined recently, sequencing many individuals during the discovery process is still expensive and time consuming, which limits the use of Sanger sequencing as the primary source of SNP discovery in all three types of applications.

A second difficulty is that sequencing from genomic DNA generally requires primers that are specific to the target sequence. In some instances, primers have been designed from closely related species, (e.g. Primmer et al. 2002; Smith et al. 2005c) or from highly conserved candidate genes for which *a priori* assumptions based on information that exists in the reference species are made about the nature of the variability at that locus, (e.g. Aitken et al. 2004; Canino et al. 2005). Another source of sequence data in a species for which little sequence information exists is information from AFLP bands (Roden et al. 2009). In this approach, the AFLP sequence itself was used to isolate random genomic DNA fragments, which were sequenced to identify potential SNPs. For many non-model organisms, little sequence information exists, although that is changing as sequencing costs decline; but typically, more expressed sequence tags (ESTs), which are sequences of cDNA from processed mRNA, are known

than are available for the unedited genomic DNA sequence. The use of ESTs for SNP discovery can often be useful, but may confound parameter estimates in studies that require variation at neutral, randomly distributed loci (population genetics and mapping work) because such surveys only include variation in coding regions, which may be subject to different selection regimes (e.g. convergent selection) than non-coding regions and may not represent the entire unprocessed gene. Scanning coding sequences may provide useful markers for studies that attempt to identify loci that are under selection, but several studies have demonstrated that non-coding regions of DNA may also provide important sequences for selection and evolution, which would not be included in an EST-based discovery process (King & Wilson 1975; Stone & Wray 2001; Begun et al. 2007; Chouard 2010). Also, if the ultimate goal is to convert sequence information to a genotyping assay, an EST-based discovery scheme can result in low conversion rate because of intervening introns.

A third problem with mass (Sanger) sequencing for SNP discovery is that insertions and deletions (INDELs) in sequences from heterozygous individuals often make interpretation of sequences downstream from the INDEL difficult and eliminate potentially informative SNPs. Recent work suggests that the density of SNPs in genomes increases nearer INDELs (Dacheng et al. 2008), and loss of these regions in the discovery process can limit the kinds of SNPs that are chosen. For moderate to low frequency alleles, an individual that is sequenced to identify the SNP will quite likely be a heterozygote. Strategies that pool individuals to increase the sample sizes of the discovery panel also increase the chance of sequencing through an INDEL, which

reduces the usefulness of mass sequencing. Although traditional Sanger sequencing may not be used extensively for SNP discovery purposes in the long term, it will continue to be a valuable tool for *de novo* sequencing of genomes, for validating polymorphisms, and for gathering long-read sequence information.

Next Generation Sequencing

Next-generation sequencing technologies have been touted as a breakthrough for many reasons (Margulies et al. 2005; Khaitovich et al. 2006; Hauser & Seeb 2008; Rokas & Abbot 2009). For reviews on the various technologies see Hudson (2008) and Shendure & Ji (2009). Each run on any of the currently available next-generation platforms can produce mega- or giga-bases of sequence information in a few days. No *a priori* assumptions need to be made about which regions of the genome to sequence, but specific genomic regions can be targeted if that is desired. This flexibility provides a platform that theoretically would be useful for discovering SNPs for all applications. In addition, the cost per base may be several orders of magnitude less expensive than standard Sanger sequencing methods (Hudson 2008). However, several potential problems should be considered before applying this technology to SNP discovery.

The first challenge is the ability to handle and analyze large amounts of data. Data files from many next-generation sequencing platforms can be in the terabyte range. Long term data storage can be unwieldy and costly, although this technology sector is advancing rapidly, which will likely resolve this issue in the near future. The main effort of most projects will shift from data acquisition to bioinformatics, which presents many

challenges with next-generation sequence data (for a review see Pop & Salzberg 2008). The introduction of ascertainment bias or identification of uninformative SNPs, which is a concern with Sanger sequencing, can present an even greater challenge with next-generation sequencing. The mega- or giga-bases of sequence information from a single run are typically obtained from a single individual or at most several individuals. Variation discovered in a few individuals cannot be assumed to be representative of the variation across the range of a species. Variation should be derived from DNA sequences of many individuals from wide geographic or temporal ranges involved in most studies. Because the current cost per run on these platforms is thousands of dollars, increasing the number of individuals in the discovery panel is not yet practical; however, as this technology continues to mature, many of these problems may be resolved.

In theory, tens, hundreds, or even thousands of individuals can be included in the discovery panel for a single run on some next-generation sequencing instruments (Meyer et al. 2007; Meyer et al. 2008; Erlich et al. 2009; Patterson & Gabriel 2009; Prabhu & Pe'er 2009); however, less sequence information is obtained from each individual and the tradeoff creates other difficulties (Holt & Jones 2008). With this strategy, the same portion of the genome (genome equivalency) must be sampled from each individual in order to get repeat coverage (a large number of overlapping contiguous DNA sequences or contigs), which is possible with organisms that have small, simple genomes such as *Caenorhabditis elegans* and *Arabidopsis thaliana* because a large portion of the genome can be sequenced in a single run. Single nucleotide polymorphism discovery in organisms that have larger, more complex genomes is more difficult because less of the

genome can be sequenced in a single run, and repetitive sequences and duplicated loci can interfere with sequence alignment. Equivalency is more easily achieved if a relatively small portion of the genome is sampled from which repetitive sequences and duplicated loci have been eliminated. One approach is to sequence only a handful of targets (Rigola et al. 2009) or ‘capture’ a subset prior to sequencing (Albert et al. 2007; Okou et al. 2007; Porreca et al. 2007), but this approach requires that the target sequences are known *a priori*. Other methods digest the DNA with restriction endonucleases and generate a ‘reduced representation’ sample (van Tassell et al. 2008), similar to the approach for AFLP analyses (Zabeau & Vos 1995) and other applications (Meissner et al. 2005; Roden et al. 2009). A recent improvement of the method fractionates the endonuclease-treated sample by size on a polyacrylamide gel that allows the removal of repetitive DNA sequences and provides the ability to select target fragments by gel excision (van Tassell et al. 2008). Restriction site recognition sequences also provide known anchor sequences, which make sequence alignment more efficient and accurate (Ng et al. 2006).

Another strategy for reducing genome complexity is to sequence the transcriptome of a pool of individuals (Barbazuk et al. 2007; Toth et al. 2007; Vera et al. 2007; Collins et al. 2008; Novaes et al. 2008; Renaut et al. 2010). Although this approach might seem to be a reasonable strategy to produce genome equivalency, it succeeds only if the same tissue is used from each individual and if the ‘expression states’ of individuals are identical. Expression of mRNA can, however, vary substantially, even among individuals in subtly different environments (Novak et al. 2002). Unequal genome contributions among individuals can confound accurate SNP detection because it can

reduce the number of contigs that are used to identify a probable SNP. However, library normalization can remove rare mRNAs from a sample (Patanjali et al. 1988). This strategy can be useful for some applications; however, it suffers the same problems that were discussed for Sanger sequencing of ESTs because it skews the discovery process toward the coding regions of the genome, although small portions of 5' and 3' untranslated regions can be included.

Accuracy reported for sequence information produced by next-generation sequencing is high. For example, a recent estimate of less than a 0.5% error rate was reported (Huse et al. 2007). However, these numbers can be misleading because most next-generation sequencing studies that have demonstrated low error rates and high numbers of valid SNPs used either a reference sequence for alignment, which was already available for the organism (termed resequencing), or highly conserved EST sequences (e.g. Barbazuk et al. 2007; Collins et al. 2008; Craig et al. 2008; Novaes et al. 2008; Sarin et al. 2008). Absence of an accurate reference sequence or highly conserved sequences from closely related species makes alignment of the short contigs from these platforms difficult. However, recent work that used paired end reads and newer alignment algorithms has resulted in accurate sequence assembly without a reference sequence (Li et al. 2010). Some platforms, such as the “two-base” encoding system used in the SOLiD platform, provide greater accuracy, but short read lengths and downstream data analysis introduce other difficulties (Holt & Jones 2008).

Sanger sequencing technology has provided the majority of the *de novo* reference sequences available for next-generation sequencing alignments. Next-generation

sequence technology and bioinformatics will need to improve in order to provide reference sequences with accuracy comparable to those generated with Sanger sequencing. A recent study that applied next-generation sequencing to SNP discovery in humans reported false positive error rates between 11% and 70% and false negative error rates between 10% and 90% (Craig et al. 2008). False positive errors can result from sequencing errors, alignment errors, or paralogous sequence variants (variants from duplicated regions of the genome). False negative errors can result from too few overlapping contigs that include a SNP or in regions that are difficult to align or sequence. Distinguishing true SNPs from false SNPs can be improved if a candidate SNP is present in multiple overlapping sequence alignments, which can derive from a single individual or multiple individuals. Accuracy of SNP identification can also increase if the SNP site is correlated among individuals from a known pedigree or is identified in numerous individuals from the same population (Xue et al. 2009); however, the strategy that is used to assemble a SNP discovery panel may not provide population structure or pedigree information. Choosing only those SNPs that are identified in numerous individuals from the same population may bias the SNP discovery process because it favors the most variable sites and eliminates rare SNPs. Exclusion of rare polymorphisms can be problematic. For instance, inference of gene trees from a collection of loci from which rare variants were omitted would remove many of the tree tips (Brumfield et al. 2003); and in humans, it has been shown that rare variants are responsible for important phenotypes (Cohen et al. 2004; McClellan et al. 2007). Clearly, next-generation sequencing is a powerful and useful technology, but caution should be exercised in its

application to SNP discovery in non-model organisms because it is a rapidly changing field. For example, single molecule real-time instruments such as Pacific Biosciences' PacBioS promise read lengths of 10 kilobases, which is a considerable increase from the current 35 to 500 basepair (bp) reads (Metzker 2009). It is likely that less expensive and more accurate systems will be available in the near future that will resolve most of the difficulties with the current technology (Eid et al. 2009; Li & Wang 2009; Rusk 2009).

Restriction-site associated DNA (RAD) markers

Restriction-site associated DNA (RAD) markers, like RFLP- and AFLP-based methods, identify SNPs that alter a restriction site (Miller et al. 2007). Genome coverage is increased because many restriction fragments can be analyzed simultaneously with microarray technology. Genomic DNA is digested with a specific endonuclease, the cleaved recognition sites are biotin labeled, and the DNA is sheared to a few hundred basepairs by sonication. Next, the biotin-labeled fragments are separated from unlabeled fragments with streptavidin beads, recovered from the beads by restriction digestion, and fluorescently labeled. The resultant fragments represent only the unmutated restriction sites, which correspond to SNP or single feature polymorphism, and can be identified with a microarray, which is either already available or constructed from clones of RAD fragments.

Restriction-site associated DNA markers have been used both to discover and to map thousands of SNPs in fungi (*Neurospora crassa*) (Lewis et al. 2007), zebrafish (*Danio rerio*) (Miller et al. 2007), and threespine stickleback (*Gasterosteus aculeatus*)

(Miller et al. 2007). They have also been applied in conjunction with next-generation sequencing, to reduce genome complexity prior to sequencing (Baird et al. 2008). Most of the applications have involved mapping studies in model organisms for which it offers several advantages such as subtractive hybridization to remove similar sequences between samples and increase the resolution between samples. The use of specific restriction endonucleases (i.e. common vs. rare recognition sites) allows the user to determine the marker density or vary the coverage of the genome. In principle, this method could also be used for classification studies because arrays would be enriched for divergent (informative) markers. However, discovery ascertainment bias may still be an issue because the platform is currently designed to develop arrays from DNA of pairs (or two pools) of individuals and genotyping for discovery would require an array for each individual. Theoretically, the RAD method could be used for SNP discovery in any organism; however, all of the studies to date have been in model organisms for which significant genomic resources are available. In addition, generation of libraries and printing arrays can be labor intensive or beyond the capabilities of many laboratories that study non-model organisms.

TILLING

Targeting Induced Lesions IN Genomes (TILLING) is a method that was developed for reverse genetics and used to identify mutations associated with a desired (often mutant) phenotype of a species (Oleykowski et al. 1998; Yang et al. 2000; Colbert et al. 2001; Henikoff et al. 2004; Till et al. 2004). Eco-TILLING (Sokurenko et al. 2001;

Comai et al. 2003; Gilchrist et al. 2006; Till et al. 2006; Till et al. 2007) and Deco-TILLING (Garvin & Gharrett 2007) are modifications of TILLING and have been used to discover and survey rare mutations in humans (Till et al. 2006), and natural mutations in wild populations of various organisms (Comai et al. 2003; Gilchrist et al. 2006; Till et al. 2007). Application of TILLING-based methods to SNP discovery has several advantages. First, ascertainment bias is reduced and more informative markers may be discovered because many individuals can be pooled in discovery panels (Garvin & Gharrett 2007); and the panels can represent a broad geographic range or appropriate cross-section of the target species. Second, pooling many individuals creates flexibility, which allows assembly of a discovery panel to identify SNPs at random, for population genetics and mapping applications; or assembling one that reflect the groups that need to be separated so that informative SNPs can be identified early in the process (classification and mapping applications). Third, costs are reduced substantially when samples are pooled as compared to sequencing the same numbers of individuals. Finally, pooling individuals for Sanger sequencing can introduce unreadable sequences from INDELS or be difficult to interpret from rare variants in multiple individuals. The problems introduced by pooling in other methods are not experienced with TILLING because homozygous individuals can often be identified during the discovery process, and provide readable sequencing data; pooling more individuals increases chances of identifying a heteroduplex. Lastly, the cleavage information from the individuals used in the TILLING reactions can be used to validate the SNP genotyping assays.

TILLING-based methods do have several drawbacks. First, PCR primers must be

designed for each target, which requires sequence information. However, a recent study identified SNP sites by heteroduplex cleavage of restriction digested, tailed, and PCR amplified fragments (Xu et al. 2009). Heteroduplexes were digested to nick the SNP site, and the 3' end was elongated with biotin-dUTP. The biotinylated DNA fragments were separated on streptavidin beads, cloned and sequenced to determine the SNP.

Unfortunately, this approach fails to address ascertainment bias concerns. A second drawback of TILLING is that the majority of studies used slab gel systems, which can be labor intensive and are not amenable to automation, although capillary instruments can and are being used for TILLING. Third, some nucleotide pair heteroduplex mismatches are recognized by the endonucleases better than others; consequently, some types of SNP can be missed (Oleykowski et al. 1998). And lastly, DEco-TILLING reduces costs associated with marker discovery but reduces the sensitivity and accuracy of the TILLING method.

Use of previously published data

Data from previous studies and published information can be valuable sources for SNP discovery. *In silico* data mining of GenBank sequences is probably the most common approach (Picoult-Newberg et al. 1999; Kota et al. 2003; Labate & Baldo 2004). Data mining can be cost-effective but it is, of course, only applicable to organisms for which substantial sequence information is available. In addition, the criteria that were used to assemble the individual sequences and the number of individuals used for generating the sequence are rarely known, and most GenBank sequences for non-model

organisms are EST sequences. Any of these factors can introduce ascertainment bias or identify uninformative markers.

Historically, allozymes provided useful genetic markers and, in some instances, the underlying mutation can be discovered and used as a SNP (Brunelli et al. 2008), although the conversion may not be straightforward in non-model and polyploid species. Other types of genetic markers may also be used as a source for discovery. Data from AFLP work has been used to identify markers in brown trout (Nicod & Largiader 2003). Restriction site analyses, which have been conducted in mitochondrial and chloroplast DNA of numerous species (Avisé 2000; Avisé 2004) indirectly identify SNPs in restriction sites and provide information about their genetic and geographic structure. This information can be used to identify informative SNPs (Vysotskaia et al. 2001). For example, we used a mitochondrial haplotype tree that was constructed from restriction sites in chum salmon to identify several variable restriction sites that were responsible for major, distinct clusters in the tree (Garvin et al. 2010a). With the mitochondrial tree, we identified specific individuals that possessed both forms of variants for informative restriction sites and sequenced them to identify and develop useful SNP assays.

Single nucleotide polymorphism genotyping assays

A multitude of SNP genotyping assays are available; for detailed reviews see Syvanen 2001, Tsuchihashi et al. 2002, Vignal et al. 2002, Sobrino et al. 2005, Giancola et al. 2006, Kim & Misra 2007, Bagge & Lubberstedt 2008, Gupta et al. 2008, and Ragoussis 2009. The focus of this review is the application of SNPs to evolutionary and

ecological studies; consequently, we focus on assays that are routinely used for these types of studies.

Criteria for choosing SNP genotyping assays

Assays for markers such as allozymes, microsatellites, AFLPs, and RAPDs can be easily transported and validated between species. In contrast, nearly every SNP assay, regardless of the platform, must be designed and validated empirically for each species. The choice of assay does not generally depend on the type of study that is being undertaken but is more closely related to the size of the project, which includes both the number of SNPs and the number of individuals that are to be genotyped (Giancola et al. 2006; Bagge & Lubberstedt 2008). The choice of assay should also consider whether the project might expand to include additional SNPs or more individuals. Finally, the choice of technology usually requires a cost analysis.

We consider a small- or medium-sized project as one that genotypes tens to about one hundred SNPs on a few hundred to a few thousand individuals. Projects of this size can be accomplished with single tube/single SNP PCR assays such as the TaqmanTM assay (Holland et al. 1991; Higuchi et al. 1993; Lee et al. 1993), the InvaderTM assay (Olivier 2005), the T_m-shift assay (Wang et al. 2005), Molecular BeaconsTM (Giesendorf et al. 1998), Amplifluor[®] (Nazarenko et al. 1997), and simple primer extension assays that require relatively little expertise. Primers and probes can be designed in house, but most fluorescent probes, primers, and enzymes for the SNP detection reactions must be ordered from a commercial source, which can increase costs. Costs can be offset by

reducing PCR volumes (5 μ l is used routinely) and by using equipment that may already be in the laboratory, such as a real-time quantitative PCR (QPCR) instrument, capillary sequencer, or plate reader. For instance, the TaqmanTM assay has gained wide application for gene expression analyses, which interrogates the fluorescence signal after each cycle of a PCR, and requires a thermocycler with the ability to detect fluorescence (Lee et al. 1993). However, the assay for allelic discrimination (Higuchi et al. 1993) can also be done by detecting fluorescent levels at the end of the PCR. In particular, the TaqmanTM assay for allele discrimination can be conducted by using a standard thermocycler to carry out the amplification and a plate reader to quantify the fluorescence, which can be less expensive to purchase and maintain than a QPCR instrument, and can be used for other analyses. Some added benefits from the use of plate readers is that many of them can detect a wider variety of fluorescent dyes than some of the standard QPCR instruments, and like the QPCR instruments, many of them come with genotyping software.

In our lab, we use a modified version of the T_m -shift assay (Wang et al. 2005) (Figure 1.2). We added locked nucleic acids (LNAs) to the 3' end of the allele specific primers, which has improved other PCR-based assays (Latorra et al. 2003; Mouritzen et al. 2003; Takatsu et al. 2004; You et al. 2006) and increased the specificity of our assays (Figure 1.2). The modified T_m -shift assay requires an instrument such as a Light Scanner from Idaho Technologies (Salt Lake City, UT) or a real-time QPCR instrument to perform a melting curve analysis but can include multiple 384-well thermocyclers to increase throughput. In a small lab, the cost for the T_m -shift assay is about an order of

magnitude lower than the Taqman[™] assay because the only major reagent investments are a generic *Taq* DNA polymerase, SYBR[™] green dye, dNTPs, and LNA-modified primers, all of which can be purchased from a wide variety of vendors at competitive prices. (The LNA primers add about \$US60 to the cost of a set of primers). Also, the genotyping is based on a profile of data points (a curve) rather than a point estimate (single point fluorescent reads), the latter of which can be misleading when samples that contain varying levels of DNA are analyzed on the same PCR plate.

Large projects may involve genotyping hundreds or even thousands of SNPs on a few individuals, genotyping a large number of individuals with a few SNPs, or genotyping many SNPs on many individuals. Most of the technology development for SNP genotyping has been driven by research on humans, which has focused on surveying very large numbers of SNPs in a few individuals. Consequently, many methods can efficiently and accurately genotype tens of thousands or even millions of SNPs simultaneously (multiplexing). Multiplexing can be achieved in single tube PCR-based assays by using different combinations of dyes on the probes or primers that are specific to each SNP, but most instruments are limited in the number of dyes that can be detected. Also, increasing the number of separate amplifications in a single PCR can result in an increased likelihood of primer-primer and primer-probe interactions that produce false signals or failed PCRs, although new microfluidics technology may eventually resolve this (Blow 2009). In addition, multiple target amplification can exhaust PCR reagents such as the dNTPs. Finally, SNPs from different genomes (e.g. nuclear versus mitochondrial) may be incompatible because of differences in copy number.

Several very accurate and robust methods can be used to perform multiplex genotyping. The iPLEX (Sequenome, San Diego, CA), SNPstream (Beckman Fullerton, CA), and SNaPshot (Applied Biosystems, Foster City, CA) assays can multiplex from 2 to about 50 SNP assays per reaction and typically use a 96 or 384 well format. Higher density SNP assays are usually performed on arrays or beads. Solid support arrays (microarrays) usually have physically attached oligonucleotides that include the target SNP site internally. Genomic DNA from an individual is queried on each array. The SNP is usually identified with a scanner that detects fluorescence at the location of a successful reaction (e.g. annealing or primer extension) on the array. Standard microarrays can include tens of thousands or even millions of SNP assays. However, they need to be manufactured for each experiment, usually cannot be reused, and tend to have high failure rates (Syvanen 2001; Tsuchihashi & Dracopoli 2002). Off-the-shelf arrays are available for model species and humans; but custom arrays must be designed for non-model species, which can make these methods expensive. Bead-based assays such as Illumina's GoldenGate assay offer a more flexible assay environment (Shen et al. 2005). Each type of bead carries an assay for a SNP. The different SNP assays are identified by a tag on the bead, which can be a DNA sequence (Shen et al. 2005) or an etched barcode (Lin et al. 2009). Other bead-based assays identify the specific assay with the fluorescence property of the beads themselves (Xu et al. 2003). They can be queried with flow-cytometers, or they can be attached to solid support universal microarrays after the assay reaction and read with an array reader. The bead format reduces the failure rate observed with standard microarrays, in part because the assays are performed in solution

prior to attachment to the solid support. Bead-based assays also easily allow the incorporation of additional SNP assays, whereas the standard microarray platforms do not. The throughput for most of the highly parallel genotyping methods are standard 96- or 384-well format (one individual, but multiple SNPs per well).

Systems that can genotype thousands of SNPs in thousands of individuals rapidly and inexpensively are not yet available; but for projects that have large numbers of individuals, increased throughput can be achieved by increasing the number of thermocyclers in the lab, by converting assays to a higher plate density (from 96 to 384 or 1536 wells per PCR plate), or by using liquid handling systems such as are used in the drug discovery and other fields. For example, the JanusTM (Perkin Elmer, Waltham MA) and the BiomekTM (Beckman Coulter, Fullerton, CA) can pipet sub-microliter volumes accurately into hundreds or thousands of PCR plates per day. Alternatively, assays can be transferred to a completely different genotyping platform. For instance, TaqmanTM, Molecular BeaconTM, and Scorpion[®] assays can be run on Fluidigm (South San Francisco, CA) instruments that use automation and microfluidics to perform 96 single assays on 96 individual samples in a single microfluidic plate. The reduction in assay volumes substantially decreases reagent costs, although there are costs for consumables, which are required for the instruments.

Expanding from assays of small or medium numbers of SNPs to large numbers of SNPs usually requires a substantial investment. The equipment costs alone range from tens of thousands to hundreds of thousands of dollars. Instruments require servicing and maintenance. Consumable reagents for the assays are often sold exclusively by the

instrument manufacturer and can be expensive. Finally, assay design and optimization are of paramount importance for multiplexing because different SNP assays must perform equally well under the same reaction conditions, and cross reactivity with other target sequences as well as non-target genomic DNA sequences must be minimized. The design of multiplex assays is usually performed by the commercial manufacturer for a fee or as part of the reagent costs; but multiplex assays that are developed for non-model species are usually not guaranteed because reference genomes are unavailable for a thorough bioinformatic analysis, which is conducted during assay development. Alternatively, the number of private institutions that provide genotyping services is increasing rapidly; and outsourcing the genotyping phase of a project to a contractor or a core lab may be more cost effective because investments in instruments, consumables, maintenance, and personnel are reduced or eliminated.

Linked single nucleotide polymorphisms: The power of haplotypes

Linkage disequilibrium is “the non-random assortment of alleles in a population at two or more loci into gametes” (Hedrick 2005). Linkage can exist if SNPs are positioned closely on a chromosome, and this can provide useful information (see below), but the SNPs cannot be treated as independent alleles because they violate Mendel’s Law of independent assortment. Linked SNPs often occur in an inherited haploid section of DNA (a haplotype). For most population genetics applications, the haplotype must be determined either empirically or probabilistically. Coalescence theory, which provides accurate estimates in population genetics applications of demography of populations, is

based on haplotypic data (Fu & Li 1999; Emerson et al. 2001; Leblois & Slatkin 2007). For classification applications, haplotypes can be the basis for species identification (Hajibabael et al. 2007), and haplotypes may be more useful than singly inherited SNPs for associating loci with disease (Davidson 2000; Drysdale et al. 2000; Hoehe et al. 2000), as well as for parentage analysis (Jones et al. 2009) and standard population genetics studies (Morin et al. 2009). Mapping applications depend on linked haplotypes to increase the resolution of linkage maps and to more accurately identify QTL. Obviously, for many applications, haplotype sequence data can provide important information.

Gathering haplotype information from linked SNPs may not be straightforward because exchanges between homologous chromosomes during meiosis can result in shuffling of haplotypes from the previous generation. If linkage is complete, conserved haplotypic sequences (haplotype blocks) will result (Wall & Pritchard 2003; Guryev et al. 2006). Because these blocks of DNA sequence are inherited, some studies can be done more quickly and inexpensively by monitoring only a few SNPs that represent the haplotype block (tagSNPs), instead of the entire collection of SNPs (Carlson et al. 2004). Although these blocks are conserved, some recombination may occasionally occur within them and create different linkage phases (Schaschl et al. 2006), which must be resolved in order to perform proper analyses for some kinds of studies. Statistical programs such as PHASE and FastPHASE (Stephens et al. 2001; Stephens & Donnelly 2003; Scheet & Stephens 2006), can estimate the haplotypic frequency in a population with maximum likelihood (ML) algorithms under the assumption of Hardy-Weinberg equilibrium

(HWE). This is useful for some population genetics estimates and for developing baseline populations for classification applications that are used for stock admixture analysis, as long as the unit of interest is a population that conforms to HWE. In contrast, statistical software packages such as SPAM (Debevec et al. 2000) BAYES (Pella & Masuda 2001), and ONCOR (Anderson et al. 2008) determine the probability that each individual (multi-locus genotype or haplotype) from a mixed stock sample (a mixture of populations) originated from the various baseline populations. Population mixtures are usually not in Hardy-Weinberg nor linkage equilibrium (the Wahlund Effect), and the linkage phase cannot be determined by ML estimation. Software programs such as Phase and other probabilistic methods, which randomly assign linkage phase to individuals, do not determine haplotypes of specific individuals in a mixture. Consequently, for stock identification applications, linkage phase must be determined empirically for each multiple heterozygote in a sample mixture.

Most haplotype information is generated from sequences or restriction digests, which are accurate for haploid mitochondrial DNA. Methods that sequence diploid genomic DNA are unable to directly determine complete haplotypes because both alleles at a locus are sequenced simultaneously. Next-generation sequencing platforms do sequence single molecules, and therefore generate haploid data, but the shorter fragment lengths generated by these methods (30 to 400 base pairs) can obscure linkage relationships. However, improvements in the technology continue to provide longer read lengths for some platforms such as Roche's 454, which can generate fragment lengths of around 500 basepairs, and so-called Third Generation Sequencing claims 10kb fragments

are routine (Metzker 2009). One way to resolve the haplotype phase is to clone individual DNA molecules, but this is not practical for large studies. Standard genotyping methods suffer from the same problem and, therefore, create what is known as unphased diploid data (sequence data in which the phase of the double heterozygotes was undetermined) (Wall & Pritchard 2003). We extensively modified a gel-based assay (Eitan & Kashi 2002) to resolve the linkage phase of double heterozygous individuals with a high-throughput SYBR-green-based assay that also incorporates LNAs at the 3' end of the primers (Garvin & Gharrett 2010b, and (Figure 1.3). This assay allows us to resolve the linkage phase of some loci for more accurate stock admixture analysis. Other assays such as Single Strand Conformation Polymorphism, or SSCP can also score complete haplotypes (Sunnucks et al. 2000).

An analysis of SNP discovery in non-model species

Single nucleotide polymorphisms have been discovered and reported in a variety of non-model organisms, which included all three types of studies that we described (Table 1.1). Although our list is not meant to be an exhaustive survey, it is representative. The majority of SNP discovery efforts in this survey used Sanger sequencing, which is not surprising given that it has been available longer than other methods, and it is accessible to most labs. Most of the studies attempted to address ascertainment bias by including representative samples in the discovery panel, and some were able to use large sample sizes, but most included only a handful of samples. The number of potential SNPs discovered and reported ranged from 2 to 1,700, which

demonstrates that large numbers of SNPs can be discovered with Sanger sequencing, although costs associated with the larger studies were not available.

Six reports used next-generation sequencing for SNP discovery in non-model organisms. Two of the studies had reference genomes available for sequence alignment, and four did not. Ascertainment bias was not addressed in these studies; either a single individual was used, or multiple individuals were pooled to increase genetic variability. Although the number of next-generation sequencing studies is small, it is clear that thousands or millions of potential SNPs can be discovered with this method. Unfortunately, the proportion of SNPs that were validated from potential SNPs detected (Table 1.1) reveals that a major challenge for data generated from next-generation SNP discovery efforts is to select the SNPs that can be validated before conversion to a genotyping assay. The study by Renaut et al. (2010) demonstrates the promise and pitfalls of next-generation sequencing for SNP discovery. The authors estimated the d_n/d_s ratio, the transition/transversion ratio, and genes that were under positive selection in two species of whitefish. Parameter estimates were derived from 6042 “potential” SNPs from the sequencing survey, but attempts were made to validate only 31 of the 6042 “potential” SNPs, 6 of which were not validated. Such parameter estimates should be viewed with caution, and the authors suggest that many of the SNPs are likely paralogous sequence variants. All SNP discovery methods for next-generation sequencing used the quality of the raw data as part of the criteria for SNP conversion; however, unlike the next-generation sequencing studies reported here, studies that used other methods for discovery were able to take into account whether SNPs were linked and how informative

they were in addition to data quality when deciding which SNPs to develop, which can be a critical issue given the cost and time needed to develop genotyping assays.

In silico data mining and TILLING were the other two methods that were heavily represented in the data set. It should be noted that RAD technology was not included in Table 1.1 because all of the studies to date involved model organisms, which is not the focus of this review. *In silico* data mining was used to discover large numbers of SNPs inexpensively; however, the drawback to this method is the inability to deal with ascertainment bias because the discovery process was rarely known. Methods that applied TILLING-based technologies were used in various species to discover varying numbers of SNPs. Most of the studies involved crop species, and identified SNPs in specific genes that were responsible for advantageous phenotypes. Finally, most of the studies reported the discovery of the SNPs themselves, but provided little or no information as to whether or not they were informative for the type of study for which they developed. From this survey, it is difficult to draw conclusions as to which methods are better at discovering informative markers, but it is anticipated that many follow-up studies with these markers will soon appear in the literature.

Summary/Conclusion

The application of SNPs to genetic studies will continue to expand because SNPs are abundant, co-dominant markers that are broadly distributed in genomes, simple to score, and amenable to high-throughput screening. The main drawbacks to applying SNPs are that the development of numerous, informative markers can be labor intensive,

can incorporate ascertainment bias, and methods to genotype them can be expensive. The importance of the SNP discovery strategy cannot be overemphasized. We have also shown that in the same application, linked SNPs may have advantages over single, independently inherited SNPs; but linkage phases must be resolved, often empirically.

As projects incorporate larger numbers of SNPs and individuals, new technologies will be developed. The large investment by private corporations in the human diagnostic and drug discovery fields has resulted in the introduction of many powerful and useful technologies for SNP discovery and genotyping. However, caution should be exercised when applying technology that was developed for the study of humans to non-model organisms because certain assumptions may not apply (e.g. a fully annotated sequence for humans is available, humans are diploid organisms, many technologies were developed for classification applications and not population genetics applications, etc.).

Finally, new technology may be costly; and, given the rapid pace of development, technologies often become obsolete in a short time. Nevertheless, as more genetic information is accumulated for non-model organisms, and as development of technology evolves toward efficient and inexpensive alternatives, SNPs will more easily and economically become incorporated into laboratories.

Acknowledgements

This represents a portion of M. Garvin's master's and doctoral work at the University of Alaska Fairbanks. We thank Sharon Hall and Rachel Riley for technical assistance. This project was supported by a grant from the University of Alaska Fairbanks Pollock Cooperative Conservation Research Center to AJG. MRG received support from the University of Alaska Experimental Program to Stimulate Competitive Research (EPSCoR). W. Smoker, D. Tallmon, J. Pella, G. Kruse, M. Canino, J. Guyon, M. McPhee, and P. Bentzen provided constructive comments on the manuscript.

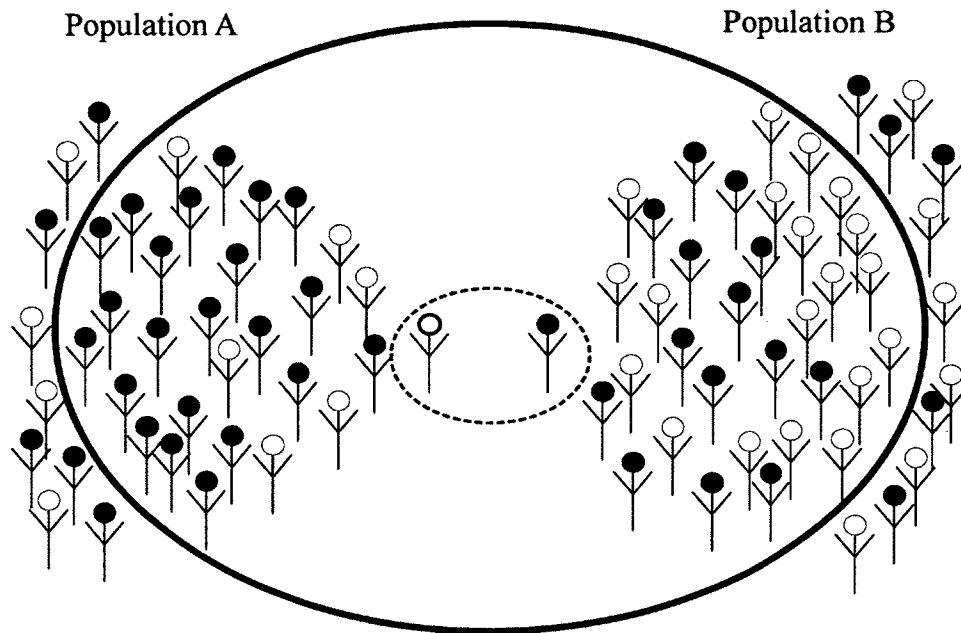


Figure 1.1. The importance of the SNP discovery scheme. A poorly designed SNP discovery scheme (represented by the dashed-line ellipse) can identify SNPs that are not representative (the unfilled circle), produce bias in parameter estimates (ascertainment bias), or include markers in an analysis that cannot distinguish between the two populations (wasted resources). A poorly designed discovery scheme may also fail to identify representative or informative markers (the gray circles), which would be more efficient for distinguishing between the two populations. For typical Sanger and next-generation sequencing studies, few individuals are selected for the discovery process. Methods such as Eco-TILLING and Deco-TILLING reduce ascertainment bias of parameter estimates and identify informative markers by including more individuals and populations in the discovery panel (solid-line ellipse).

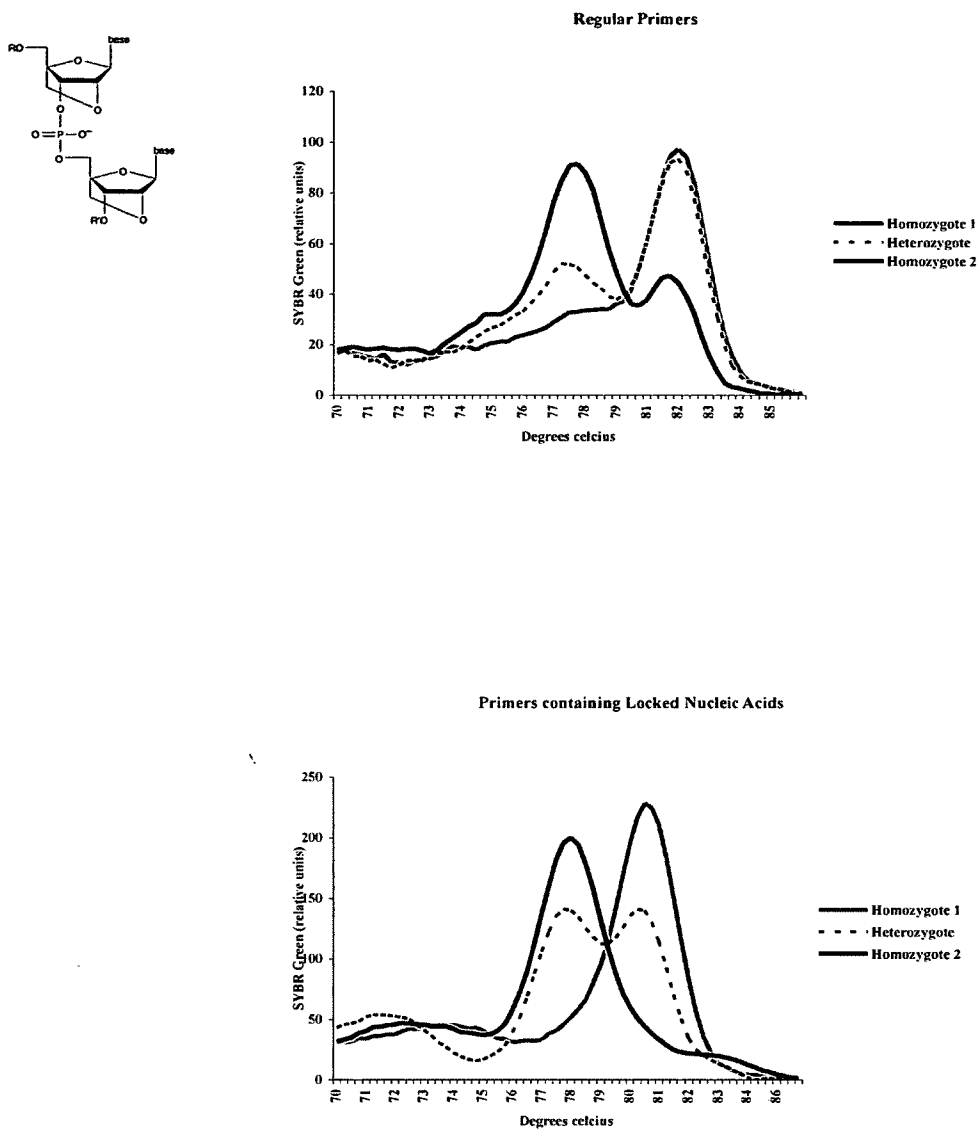


Figure 1.2. The improved T_m -shift assay. The assay uses locked nucleic acids (LNAs) placed at the 3' ends that complement the SNP site. LNAs lock the ribose group of the nucleic acid in the chair form (upper left drawing), which increases its affinity for its complementary nucleotide (You et al. 2006). The top graph shows the assay with unmodified primers. The bottom graph shows the assay on the same samples that incorporates LNAs into the SNP specific-3' position of the primers.

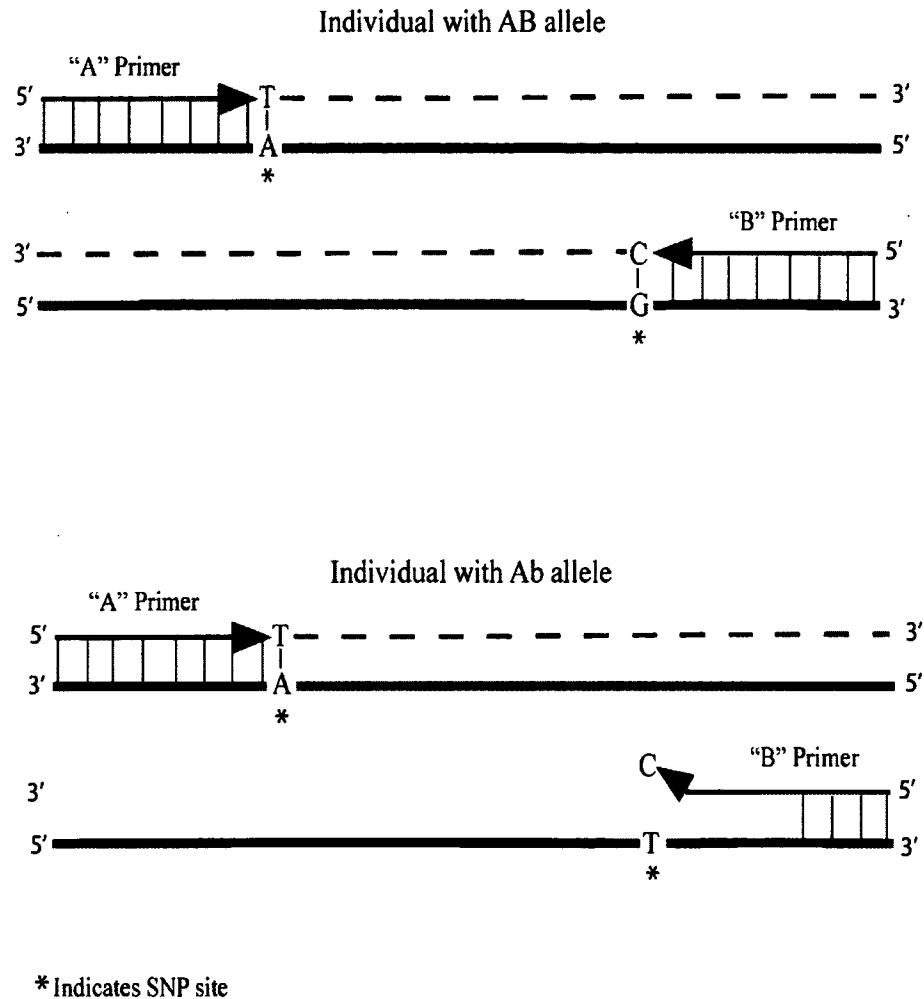


Figure 1.3. Linkage phase resolution. We developed a rapid assay that resolves linkage phases in linked SNPs, which is an improvement of an older method (Eitan & Kashi 2002). Primer pairs were designed so that the 3' end of the forward and the reverse primers terminated at the first and second SNP sites in the haplotype, respectively. An LNA modified nucleotide was placed at the 3' end of the primer. Here, we refer to the two alleles at SNP site 1 as A/a and the alleles at SNP site 2 as B/b; there are two forward (A, a) and two reverse (B, b) primers. A detectable PCR product can only be generated if nucleotides at both SNP sites complement the primers. In order to develop each assay, we used four primer pairs (A-B, a-B, A-b, and a-b). Only one combination of primers can amplify a specific haplotype. In practice, the results of two reactions will determine the phase of double heterozygotes, one that uses the primer pair that amplifies one of the repulsion phase (A-b or a-B) and one that amplifies one of the coupling phase (a-b or A-B) haplotypes.

Table 1.1. Summary of SNP discovery efforts in non-model organisms.

| Species | Study Type | SNP Discovery Method | Ascertainment bias addressed ? | Max # individuals in ascertainment panel | # potential SNPs* |
|--|--|--|--------------------------------|--|-------------------|
| Arctic char (<i>Salvelinus alpinus</i>) | Undefined | Sanger Sequencing | Yes | 5 | 2 |
| Atlantic cod (<i>Gadus morhua</i>) | Classification | Sanger Sequencing | No | 5* | 724 |
| Atlantic salmon (<i>Salmo salar</i>) | Undefined | Sanger Sequencing | Yes | 15 | 19 |
| Brown trout (<i>Salmo trutta</i>) | Undefined | Sanger Sequencing | Yes | 5 | 15 |
| Brown trout (<i>Salmo trutta</i>) | Population Genetics | Sanger Sequencing/AFLP | No | 16 | 24 |
| Chimpanzee (<i>Pan troglodytes verus</i>) | Undefined | Sanger Sequencing | Yes | 19 | 26 |
| Chinook salmon (<i>Oncorhynchus tshawytscha</i>) | Classification | Sanger Sequencing | Yes | 347 | 40 |
| Chinook salmon (<i>Oncorhynchus tshawytscha</i>) | Classification | Sanger Sequencing | Yes | 10 | 114 |
| Chinook salmon (<i>Oncorhynchus tshawytscha</i>) | Classification | Sanger Sequencing | Yes | 32 | 54 |
| Chum salmon (<i>Oncorhynchus keta</i>) | Classification | Sanger Sequencing | Yes | 32 | 13 |
| Chum salmon (<i>Oncorhynchus keta</i>) | Classification | Sanger Sequencing | Yes | 50 | 107 |
| Chum salmon (<i>Oncorhynchus keta</i>) | Classification | Sanger Sequencing | Yes | 10 | 155 |
| Chum salmon (<i>Oncorhynchus keta</i>) | Classification | TILLING | Yes | 5'x 96 | 19 |
| Coho salmon (<i>Oncorhynchus kisutch</i>) | Classification | Sanger Sequencing | Yes | 37 | 46 |
| Collared Flycatcher (<i>Ficedula albicollis</i>) | Population Genetics | Sanger Sequencing | No | 8 | 61 |
| Coral (<i>Acropora millepora</i>) | Population Genetics | NGS - 454** | No | Unknown | 48,046 |
| Eastern fence lizard (<i>Sceloporus undulatus</i>) | Population Genetics | Sanger Sequencing | Yes | 91 | 158 |
| Eelgrass (<i>Zostera marina</i>) | Population Genetics | <i>In silico</i> data mining/Sanger Sequencing | Yes | 16 | 152 |
| Giant Panda (<i>Ailuropoda melanoleura</i>) | Population Genetics | NGS - Illumina | No | 1 | 2,700,000 |
| Grapevine (<i>Vitis vinifera</i>) | Population Genetics/Mapping/Classification | Sanger Sequencing | Yes | 10 | 1625 |

| # Validated SNPs ^a | # characterized SNPs | How were validated SNPs characterized? | Method used to validate SNP | Reference |
|-------------------------------|----------------------|--|-------------------------------------|-------------------------------|
| 2 | 2 | Nucleotide diversity | Sanger Sequencing | Ryynanen & Primmer 2006 |
| 318/594 ^A | 318 | Population assignment, F_{ST} | MassARRAY Sequenome | Moen <i>et al.</i> 2008 |
| 19 | 19 | Nucleotide diversity | Sanger Sequencing | Ryynanen & Primmer 2006 |
| 15 | 15 | Nucleotide diversity | Sanger Sequencing | Ryynanen & Primmer 2006 |
| 12 | 12 | Resequencing | Sanger Sequencing | Nicod <i>et al.</i> 2003 |
| 19/26 ^A | 14 | Genotyping assay success | SNaPshot | Aiken <i>et al.</i> 2004 |
| 40 | 10 | MAF, FIS, UPGMA, AMOVA | Taqman TM | Smith <i>et al.</i> 2005 |
| 41 | 41 | MAF, not paralogous | Taqman TM | Smith <i>et al.</i> 2005 |
| 13 | 12 | MAF | Taqman TM | Campbell <i>et al.</i> 2008 |
| 13 | 13 | MAF, H_O and F_{ST} | Taqman TM | Smith <i>et al.</i> 2005 |
| 107 | 36 | MAF, H_O and F_{ST} and unlinked | Taqman TM | Elfstrom <i>et al.</i> 2007 |
| 55 | 55 | MAF, not paralogous | Taqman TM | Smith <i>et al.</i> 2005 |
| 15 | 6 | MAF | T_m -shift & Taqman TM | Garvin & Gharrett 2007 |
| 21 | 19 | MAF, H_O and F_{ST} and unlinked | Taqman TM | Smith <i>et al.</i> 2005 |
| 61 | 61 | Nucleotide diversity, H_O | Sanger Sequencing | Primmer <i>et al.</i> 2002 |
| 14/20 ^A | 14 | Resequencing | Resequencing | Meyer <i>et al.</i> 2009 |
| 158 | 19 | MAF, Nucleotide diversity, H_O | Sanger Sequencing | Rosenblum <i>et al.</i> 2007 |
| 47/152 ^A | 37 | MAF, H_O and F_{ST} and unlinked | SNaPshot | Ferber <i>et al.</i> 2008 |
| 0 | 0 | N/A | N/A | Li <i>et al.</i> 2010 |
| 80/96 ^A | 80 | MAF | SNPplex | Lijavetzky <i>et al.</i> 2007 |

Table 1.1 continued.

| Species | Study Type | SNP Discovery Method | Ascertainment bias addressed ? |
|---|------------------------------------|--|--------------------------------|
| Grayling (<i>Thymallus thymallus</i>) | Undefined | Sanger Sequencing | Yes |
| Green algae (<i>Chlamydomonas reinhardtii</i>) | Mapping | <i>In silico</i> data mining, RFLP mapping | No |
| Green sea turtle (<i>Chelonia mydas</i>) | Population Genetics | Sanger Sequencing/AFLP | Yes |
| Half-smoothed tongue sole (<i>Cynoglossus semilaevis</i>) | Mapping | TILLING | No |
| Lake whitefish (<i>Coregonus</i> spp.) | Classification | NGS - 454** | No |
| Melon (<i>Cucumis melo</i> L.) | Classification | TILLING | Yes |
| Mosquito (<i>Aedes aegypti</i>) | Classification | Sanger Sequencing | Yes |
| Mung bean (<i>Vigna radiata</i>) | Classification/Population Genetics | TILLING | Yes |
| Pacific Oyster (<i>Crassostrea gigas</i>) | Mapping | <i>In silico</i> data mining | No |
| Paddy weed (<i>Monochoria vaginalis</i>) | Classification | TILLING | Yes |
| Pied Flycatcher (<i>Ficedula hypoleuca</i>) | Population Genetics | Sanger Sequencing | No |
| Rainbow trout (<i>Oncorhynchus mykiss</i>) | Mapping/Classification | NGS - 454 | No |
| Rice (<i>Oryza sativa</i> L.) | Classification | TILLING | Yes |
| Rose Gum (<i>Eucalyptus grandis</i>) | Classification | NGS - 454 | No |
| Sockeye salmon (<i>Oncorhynchus nerka</i>) | Classification | Sanger Sequencing | Yes |
| Sperm Whale (<i>Physeter macrocephalus</i>) | Population Genetics | Sanger Sequencing | Yes |
| Sugarcane (<i>Saccharum officinarum</i>) | Undefined | <i>In silico</i> data mining | No |
| Trinidadian guppy (<i>Poecilia reticulata</i>) | Mapping | Sanger Sequencing | Yes |
| Turkey (<i>Meleagris gallopavo</i>) | Undefined | NGS - Illumina** | No |
| Weathervane scallop (<i>Patinopecten caurimus</i>) | Population Genetics/Classification | Sanger Sequencing | No |
| Western black cottonwood (<i>Populus trichocarpa</i>) | Population Genetics | TILLING | Yes |
| White spruce (<i>Picea glauca</i>) | Undefined | <i>In silico</i> data mining | Yes |
| Windflower (<i>Anemone coronaria</i>) | Population Genetics | Sanger Sequencing | No |

MAF = Minor Allele Frequency; HO = Observed Heterozygosity; FST = Wright's FST, UPGMA = Unweighted Pair Group Method
 *SNPs also includes INDELS

^aA subset of SNPs was chosen from potential SNPs

^bValidated SNP means it was validated with a method other than the discovery method

^cIndividuals were pooled

**Reference sequence or conserved reference sequence available for alignment

| Max # individuals in ascertainment panel | # potential SNPs [†] | # Validated SNPs [‡] | # characterized SNPs | How were validated SNPs characterized? | Method used to validate SNP | Reference |
|--|-------------------------------|-------------------------------|----------------------|--|-----------------------------|-------------------------------|
| 5 | 14 | 14 | 14 | Nucleotide diversity | Sanger Sequencing | Ryynanen & Primmer 2006 |
| 2 | 204 | 186/204 ^A | 186 | Resequencing | Sanger Sequencing | Vysotskaia <i>et al.</i> 2001 |
| 40 | 68 | 37 | 29 | MAF, Nucleotide polymorphism, H ₀ | Amplifluor | Roden <i>et al.</i> 2009 |
| 10* | 41 | 23/41 ^A | 9 | Resequencing | Sanger Sequencing | Xu <i>et al.</i> 2009 |
| 24* | 6042 | 25/31 ^A | 6042 | dn/ds, tv/ts ratio | SNPplex | Renaut <i>et al.</i> 2010 |
| 112 | 6 | 6 | 6 | Nonsynonymous mutation | Sanger Sequencing | Nieto <i>et al.</i> 2007 |
| 15 | 87 | 8 | 8 | MAF, H ₀ , and unlinked | SNaPshot | Paduan & Ribolla 2009 |
| 24 | 157 | 157 | 157 | % polymorphic sites | Sanger Sequencing | Barkely <i>et al.</i> 2008 |
| Unknown | 51 | 51 | 20 | Polymorphic/MAF | Sanger Sequencing | Bai <i>et al.</i> 2009 |
| 4 | 2 | 2 | 2 | Nonsynonymous mutation | Sanger Sequencing | Wang <i>et al.</i> 2007 |
| 8 | 52 | 52 | 52 | Nucleotide diversity, H ₀ | Sanger Sequencing | Primmer <i>et al.</i> 2002 |
| 96* | 13,140-24,627 | 183/384 ^A | 183 | MAF, H ₀ | Golden Gate Assay | Sanchez <i>et al.</i> 2009 |
| 57 | 6 | 6 | 6 | Genotype/Phenotype association | Sanger Sequencing | Kadaru <i>et al.</i> 2006 |
| 21* | 30,108 | 279/337 ^A | 279 | dN/dS ratio | Resequencing | Novaes <i>et al.</i> 2008 |
| 10 | 114 | 39 | 39 | MAF, not paralogous | Taqman™ | Smith <i>et al.</i> 2005 |
| 6 | 39 | 39 | 18 | MAF and unlinked | Luminex | Morin <i>et al.</i> 2007 |
| Unknown | 1588 | 58/180 ^A | 48 | Polymorphic, dosage of polyploid species | Pyrosequencing | Cordeiro <i>et al.</i> 2006 |
| 5 | 1700 | 400 | 235 | Resequencing/Sequence alignment to other populations | Sanger Sequencing | Dreyer <i>et al.</i> 2007 |
| 6* | 7,952 | 355 | 384 | MAF | Golden Gate Assay | Kerstens <i>et al.</i> 2009 |
| 95 | 27 | 12 | 12 | MAF, H ₀ | Taqman™ | Elfstrom <i>et al.</i> 2005 |
| 41 | 63 | 63 | | Resequencing, H ₀ , UPGMA | Sanger Sequencing | Gilchrist <i>et al.</i> 2006 |
| 12 | 12,264 | 245/325 ^A | 245 | Resequencing | Sanger Sequencing | Pavy <i>et al.</i> 2006 |
| 1 | 155 | 30 | 9 | MAF, H ₀ , Genetic Distance | MALDI-TOF MS | Shanay <i>et al.</i> 2006 |

[†]with Arithmetic mean, AMOVA = Analysis of Molecular Variance

References

- Aitken, N., Smith, S., Schwartz, C., and Morin, P.A. 2004. Single-nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach. *Molecular Ecology* **13**: 1423-1431.
- Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., Weinstock, G.M., and Gibbs, R.A. 2007. Direct selection of human genomic loci by microarray hybridization. *Nature Methods* **4**: 903-905.
- Anderson, E. 2010. Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology*. *Molecular Ecology* **10**: 701-710.
- Anderson, E., Waples, R.S., and Kalinowski, S.T. 2008. An improved method for estimating the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences* **65**: 1475-1486.
- Avise, J.C. 1994. *Molecular Markers, Natural History and Evolution*. Chapman Hall, Sunderland, MA.
- Avise, J.C. 2000. *Phylogeography: The History and Formation of Species*. Harvard University Press, Cambridge.
- Avise, J.C. 2004. *Molecular markers, natural history and evolution*. 2nd ed. Sunderland and Associates, Inc., Sunderland, MA.
- Bagge, M., and Lubberstedt, T. 2008. Functional markers in wheat: technical and economic aspects. *Molecular Breeding* **22**(3): 319-328.

- Bai, J., Li, Q., Kong, F., and Li, R. 2009. Characterization of 20 single nucleotide polymorphism markers in the Pacific oyster (*Crassostrea gigas*). *Animal Genetics* **40**: 1004.
- Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., and Cresko, W.A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**(10): e3376.
- Barbazuk, W.B., Emrich, S.J., Chen, H.D., Li, L., and Schnable, P.S. 2007. SNP discovery via 454 transcriptome sequencing. *The Plant Journal* **51**: 910-918.
- Barkley, N.A., Wang, M.L., Gillaspie, A.G., Dean, R.E., Pederson, G.A., and Jenkins, T.M. 2008. Discovering and verifying DNA polymorphisms in a mung bean [*V. radiata* (L.) R. Wilczek] collection by EcoTILLING and sequencing. *BMC Research Notes* **1**: 28-34.
- Begun, D.J., Holloway, A.K., Stevens, K., Hiller, L.W., Poh, Y.-P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., Pachter, L., Myers, E., and Langley, C.H. 2007. Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology* **5**(11): 2534-2559.
- Bensch, S., Åkesson, S., and Irwin, D. 2002. The use of AFLP to find an informative SNP: genetic differences across a migratory divide in willow warblers. *Molecular Ecology* **11**(11): 2359-2366.
- Blow, N. 2009. Genomics: catch me if you can. *Nature Methods* **6**(7): 539-544.
- Boulding, E.G., Culling, M., Glebe, B., Berg, P.R., Lien, S., and Moen, T. 2008. Conservation genomics of Atlantic salmon: SNPs associated with QTLs for

- adaptive traits in parr from four trans-Atlantic backcrosses. *Heredity* **101**(4): 381-391.
- Brumfield, R.T., Beerli, P., Nickerson, D.A., and Edwards, S.V. 2003. The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution* **18**(5): 249-256.
- Brunelli, J.P., Thorgaard, G.H., Leary, R.F., and Dunnigan, J.L. 2008. Single-nucleotide polymorphisms associated with allozyme differences between inland and coastal rainbow trout. *Transactions of the American Fisheries Society* **37**(5): 1292-1298.
- Campbell, N.R., and Narum, S.R. 2008. Identification of novel single-nucleotide polymorphisms in Chinook salmon and variation among life history types. *Transactions of the American Fisheries Society* **137**: 96-106.
- Canino, M.G., O'Reilly, P.T., Hauser, L., and Bentzen, P. 2005. Genetic differentiation in walleye pollock (*Theragra chalcogramma*) in response to selection at the pantophysin (PanI) locus. *Canadian Journal of Fisheries and Aquatic Sciences* **62**: 2519-2529.
- Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., and Nickerson, D.A. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics* **74**: 106-120.
- Chakraborty, R., Stivers, D.N., Su, B., Zhong, Y., and Budowle, B. 1999. The utility of short tandem repeat loci beyond human identification: Implications for development of new DNA typing systems. *Electrophoresis* **20**: 1682-1696.

- Chouard, T. 2010. Revenge of the hopeful monster. *Nature* **463**(7283): 864-867.
- Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H., and Nielsen, R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* **15**: 1496-1502.
- Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**: 869-872.
- Colbert, T., Till, B.J., Tompa, R., Reynolds, S., Steine, M., Yeung, A.T., McCallum, C.M., Comai, L., and Henikoff, S. 2001. High-throughput screening for induced point mutations *Plant Physiology* **126**: 480-484.
- Collard, B.C.Y., Jahufer, M.Z.Z., Brouwer, R.B., and Pang, E.C.K. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* **142**: 169-196.
- Collins, L.J., Biggs, P.J., Voelckel, C., and Joly, S. 2008. An approach to transcriptome analysis of non-model organisms using short-read sequences. *Genome Informatics* **21**: 3-14.
- Comai, L., Young, K., Till, B., Reynolds, S., Green, E.A., Codomo, C.A., Enns, L.C., Johnson, J.E., Burtner, C., Odden, A.R., and Henikoff, S. 2003. Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. *The Plant Journal* **37**(5): 778-786.

- Cordeiro, G.M., Elliott, F., McIntyre, C.L., Casu, R.E., and Henry, R.J. 2006. Characterisation of single nucleotide polymorphisms in sugarcane ESTs. *Theoretical Applied Genetics* **113**: 331-343.
- Craig, D., Pearson, J., Szelinger, S., Sekar, A., and Redman, M. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods* **5**(10): 887-893.
- Dacheng, T., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., Nagylaki, T., Hudson, R., Bergelson, J., and Chen, J.-Q. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455**: 105-108.
- Davidson, S. 2000. Research suggests importance of haplotypes over SNPs. *Nature Biotechnology* **18**: 1134-1135.
- Debevec, E., Gates, R., Masuda, M., Pella, J., Reynolds, J., and Seeb, L. 2000. SPAM (version 3.2): Statistics Program for Analyzing Mixtures. *Journal of Heredity* **91**: 509-510.
- Decker, J., Pires, J., Conant, G., McKay, S., Heaton, M., Chen, K., Cooper, A., Vilkki, J., Seabury, C., Caetano, A., Johnson, G., Brenneman, R., Hanotte, O., Eggert, L., Wiener, P., Kim, J., Kim, K., Sonstegard, T., Tassell, C.V., Neibergs, H., McEwan, J., Brauninn, R., Coutinho, L., Babar, M., Wilson, G., McClure, M., Rolf, M., Kim, J., Schnabel, R., and Taylor, J. 2009. Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proceedings of the National Academy of Sciences* **106**(44): 18644-18649.

- Dreyer, C., Hoffmann, M., Lanz, C., Willing, E.-M., Riester, M., Warthmann, N., Sprecher, A., Tripathi, N., Henz, S.R., and Weigel, D. 2007. ESTs and EST-linked polymorphisms for genetic mapping and phylogenetic reconstruction in the guppy, *Poecilia reticulata*. *BMC Genomics* **8**: 269-277.
- Drysdale, C.M., McGraw, D.W., Stack, C.B., Stephens, J.C., Judson, R.S., Nandabalan, K., Arnold, K., Ruano, G., and Liggett, S.B. 2000. Complex promoter and coding region β 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proceedings of the National Academy of Sciences* **97**(19): 10483-10488.
- Edwards, S.V. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* **63**(1): 1-19.
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**(5910): 133-138.

- Eitan, Y., and Kashi, Y. 2002. Direct micro-haplotyping by multiple double PCR amplifications of specific alleles (MD-PASA). *Nucleic Acids Research* **30**(12): 1-8.
- Elfstrom, C., Gaffney, P., Smith, C., and Seeb, J. 2005. Characterization of 12 single nucleotide polymorphisms in weathervane scallop. *Molecular Ecology Notes* **5**: 406-409.
- Elfstrom, C., Smith, C., and Seeb, L. 2007. Thirty-eight single nucleotide polymorphism markers for high-throughput genotyping of chum salmon. *Molecular Ecology Notes* **7**: 1211-1215.
- Emerson, B.C., Pardis, E., and Thebaud, C. 2001. Revealing the demographic histories of species using DNA sequences. *Trends in Ecology and Evolution* **16**(12): 707-716.
- Erlich, Y., Chang, K., Gordon, A., Ronen, R., Navon, O., Rooks, M., and Hannon, G.J. 2009. DNA sudoku - harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Research* **19**: 1243-1253.
- Ferber, S., Reusch, T.B.H., Stam, W.T., and Olsen, J.L. 2008. Characterization of single nucleotide polymorphism markers for eelgrass (*Zostera marina*). *Molecular Ecology Resources* **8**: 1429-1435.
- Fisher, R. 1934. The effects of methods of ascertainment upon the estimation of frequencies. *Annals of Human Eugenics* **6**: 13-25.
- Fu, Y.-X., and Li, W.-H. 1999. Coalescing into the 21st century: An overview and prospects of coalescent theory. *Theoretical Population Biology* **56**: 1-10.

- Garvin, M.R., and Gharrett, A.J. 2007. DEco-TILLING: An inexpensive method for SNP discovery that reduces ascertainment bias. *Molecular Ecology Notes* **7**: 735-746.
- Garvin MR, Saitoh K, Brykov V, Churikov D, Gharrett AJ (2010a) Single nucleotide polymorphisms in chum salmon (*Oncorhynchus keta*) mitochondrial DNA derived from restriction site haplotype information. *Genome* **53**, 501-507.
- Garvin MR, Gharrett AJ (2010b) Application of SNP markers to chum salmon (*Oncorhynchus keta*): Discovery, genotyping, and linkage phase resolution. *Journal of Fish Biology* **77**, 2137-2162.
- Garvin MR, Marcotte RW, Palof KJ, et al. (2011) Diagnostic single-nucleotide polymorphisms identify Pacific ocean perch and delineate blackspotted and roughey rockfish. *Transactions of the American Fisheries Society* **140**, 984-988
- Giancola, S., McKhann, H.I., Berard, A., Camilleri, C., Durand, S., Libeau, P., Roux, F., Reboud, X., Gut, I.G., and Brunel, D. 2006. Utilization of the three high-throughput SNP genotyping methods, the GOOD assay, Amplifluor and Taqman, in diploid and polyploid plants. *Theoretical and Applied Genetics* **112**: 1115-1124.
- Giesendorf, B.A.J., Vet, J.A.M., Tyagi, S., Mensink, E.J.M.G., Trijbels, F.J.M., and Blom, H.J. 1998. Molecular beacons: a new approach for semi-automated mutation analysis. *Clinical Chemistry* **44**: 482-486.
- Gilchrist, E., Haughn, G.W., Ying, C.C., Otto, S.P., Zhuang, J., Hamberger, B., Aboutorabi, F., Kalynkay, T., Johnson, L., Bohlmann, J., Ellis, B., Douglas, C.J., and Cronk, Q.C.B. 2006. Use of Ecotilling as an efficient SNP discovery tool to

- survey genetic variation in wild populations of *Populus trichocarpa*. *Molecular Ecology* **15**: 1367-1378.
- Gupta, P., Rustgi, S., and Mir, R. 2008. Array-based high-throughput DNA markers for crop improvement. *Heredity* **101**: 5-18.
- Guryev, V., Smits, B., van de Belt, J., Verheul, M., Hubner, N., and Cuppen, E. 2006. Haplotype block structure is conserved across mammals. *PloS Genetics* **2**(7): 1111-1118.
- Hajibabael, M., Singer, G.A.C., Hebert, P.D.N., and Hickey, D.A. 2007. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics* **23**(4): 167-172.
- Hauser, L., and Seeb, J.E. 2008. Advances in molecular technology and their impact on fisheries genetics. *Fish and Fisheries* **9**: 473-486.
- Heaton, M.P., Harhay, G.P., Bennett, G.L., Stone, R.T., Grosse, W.M., Casas, E., Keele, J.W., Smith, T.P.L., Chitko-McKown, C.G., and Laegreid, W.W. 2002. Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. *Mammalian Genome* **13**(5): 272-281.
- Hedrick, P. 2005. *The Genetics of Populations*. Third ed. Jones and Bartlett, Sudbury, MA.
- Henikoff, S., Till, B.J., and Comai, L. 2004. TILLING: Traditional mutagenesis meets functional genomics *Plant Physiology* **135**: 1-7.

- Higuchi, R., Fockler, C., Dollinger, G., and Watson, R. 1993. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Biotechnology* **11**(9): 1026-1030.
- Hoehe, M.R., Kopke, K., Wendel, B., Rohde, K., Flachmeier, C., Kidd, K.K., Berrettini, W.H., and Church, G.M. 2000. Sequence variability and candidate gene analysis in complex disease: association of μ opioid receptor gene variation with substance dependence. *Human Molecular Genetics* **9**(19): 2895-2908.
- Holland, P.M., Abramson, R.D., Watson, R., and Gelfand, D. 1991. Detection of specific polymerase chain reaction product by utilizing the 5' \rightarrow 3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proceedings of the National Academy of Sciences* **88**: 7276-7280.
- Holt, R.A., and Jones, S.J.M. 2008. The new paradigm of flow cell sequencing. *Genome Research* **18**: 839-846.
- Hudson, M.E. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources* **8**: 3-17.
- Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and Welch, D.M. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**: R143.
- Hyten, D.L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R.L., Costa, J.M., Specht, J.E., Shoemaker, R.C., and Cregan, P.B. 2006. Impacts of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences* **102**(45): 6666-16671.

- Jones, B., Walsh, D., Werner, L., and Fiumera, A. 2009. Using blocks of linked single nucleotide polymorphisms as highly polymorphic genetic markers for parentage analysis. *Molecular Ecology Resources* **9**(2): 487-497.
- Kadaru S.B., Ydav A.S., Fjellstrom R.G., Oard J.H. 2006. Alternative EcoTILLING protocol for rapid, cost-effective single nucleotide polymorphism discovery and genotyping in rice (*Oryza sativa* L.). *Plant Molecular Biology Reporter* **24**: 3-22.
- Keller, I., Veltsos, P., and Nichols, R.A. 2008. The frequency of rDNA variants within individuals provides evidence of population history and gene flow across a grasshopper hybrid zone. *Evolution* **62**(4): 833-844.
- Kerstens, H.H., Crooijmans, R.P., Veenendaal, A., Dibbits, B.W., Chin-A-Woeng, T.F., Dunnen, J.T.d., and Groenen, M.A. 2009. Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics* **10**: 479-489.
- Khaitovich, P., Enard, W., Lachmann, M., and Paabo, S. 2006. Evolution of primate gene expression. *Nature Reviews* **7**: 693-702.
- Kim, S., and Misra, A. 2007. SNP genotyping: Technology and biomedical applications. *Annual Review of Biomedical Engineering* **9**: 289-320.
- King, M., and Wilson, A. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107-116.

- Kota, R., Rudd, S., Facius, A., Kolesov, G., Thiel, T., Zhang, H., Stein, N., Mayer, K., and Graner, A. 2003. Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare L.*). *Molecular Genetics and Genomics* **270**: 24-33.
- Krawczak, M. 1999. Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis* **20**: 1676-1681.
- Kuhner, M.K., Beerli, P., Yamato, J., and Felsenstein, J. 2000. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* **156**: 439-447.
- Labate, J.A., and Baldo, A.M. 2004. Tomato SNP discovery by EST mining and resequencing. *Molecular Breeding* **16**(4): 343-349.
- Latorra, D., Campbell, K., Wolter, A., and Hurley, J.M. 2003. Enhanced allele-specific PCR discrimination in SNP genotyping using 3' locked nucleic acid (LNA) primers. *Human Mutation* **22**(1): 79-85.
- Leblois, R., and Slatkin, M. 2007. Estimating the number of founder lineages from haplotypes of closely linked SNPs. *Molecular Ecology* **16**: 2237-2245.
- Lee, L.G., Connell, C.R., and Bloch, W. 1993. Allelic discrimination by nick-translation PCR with fluorogenic probes. *Nucleic Acids Research* **21**: 3761-3766.
- Lewis, Z.A., Shiver, A.L., Stiffler, N., Miller, M.R., Johnson, E.A., and Selker, E.U. 2007. High-density detection of restriction-site-associated DNA markers for rapid mapping of mutated loci in *Neurospora*. *Genetics* **177**: 1163-1171.
- Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z.,

Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O.A., Leung, F.C.-C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shang, G., Zhang, S., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C.C., Lam, T.T.-Y., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M.W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.-W., Yiu, S.-M., Liu, S., Zhang, H., Li, D., Huang, Y., Wan, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G.K.-S., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J., and Wang, J. 2010. The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**: 311-317.

Li, Y., and Wang, J. 2009. Faster human genome sequencing. *Nature Biotechnology* **27**(9): 820-821.

Lijavetzky, D., Cabezas, J.A., Ibanez, A., Rodriguez, V., and Martinez-Zapater, J.M. 2007. High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a resequencing approach and SNPlex technology. *BMC Genomics* **8**: 424-435.

- Lin, C.H., Yeakley, J.M., McDaniel, T.K., and Shen, R. 2009. Medium- to high-throughput SNP genotyping using VeraCode microbeads. *Methods in Molecular Biology* **496**: 129-142.
- Liu, Z., and Cordes, J. 2004. DNA marker technologies and their applications in aquaculture genetics. *Aquaculture* **242**: 735-736.
- Luikart, G., England, P.R., Tallmon, D., Jordon, S., and Taberlet, P. 2003. The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics* **4**: 981-994.
- Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-B., Knight, J.R., Lanza, J.R., Leamo, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., and Rothberg, J.M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.
- Marth, G., Czabarka, E., Murvai, J., and Sherry, S. 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**: 351-372.

- McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., and Hirschhorn, J.N. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**: 356-369.
- McClellan, J., Susser, E., and King, M. 2007. Schizophrenia: a common disease caused by multiple rare alleles. *British Journal of Psychology* **190**: 194-199.
- Meissner, A., Gnirke, A., Bell, G., Ramsahoye, B., Lander, E., and Jaenisch, R. 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research* **33**: 5868-5877.
- Metzker, M. 2009. Sequencing in real time. *Nature Biotechnology* **27**: 150-151.
- Meyer, E., Aglyamova, G.V., Wang, S., Muchanan-Carter, J., Abrego, D., Colbourne, J.K., Willis, B.L., and Matz, M.V. 2009. Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* **10**: 219-236.
- Meyer, M., Stenzel, U., and Hofreiter, M. 2008. Parallel tagged sequencing on the 454 platform. *Nature Protocols* **3**(2): 267-278.
- Meyer, M., Stenzel, U., Myles, S., Pruffer, K., and Hofreiter, M. 2007. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research* **35**(15): e97.
- Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., and Johnson, E.A. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated (RAD) markers. *Genome Research* **17**: 240-248.

- Moen, T., Hayes, B., Nilsen, F., Delghandi, M., Fjalestad, K.T., Fevolden, S.-E., Bert, P.R., and Lien, S. 2008. Identification and characterisation of novel SNP markers in Atlantic cod: Evidence for directional selection. *BMC Genetics* **9**: 18-26.
- Moran, P. 2002. Current conservation genetics: building an ecological approach to the synthesis of molecular and quantitative genetic methods. *Ecology of Freshwater Fish* **11**: 30-55.
- Morin, P., Aitken, N., Rubio-Cisneros, N., Dizon, N., and Mesnick, S. 2007. Characterization of 18 SNP markers for sperm whale (*Physeter macrocephalus*). *Molecular Ecology Notes* **7**: 626-630.
- Morin, P., Luikart, G., Wayne, R., and group, SNP Workshop Group. 2004. SNPs in ecology, evolution and conservation. *Trends in Ecology and Evolution* **19**(4): 208-216.
- Morin, P., Martien, K., and Taylor, B. 2009. Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources* **9**: 66-73.
- Mouritzen, P., Nielsen, A.T., Pfundheller, H.M., Choleva, Y., Kongsbak, L., and Moller, S. 2003. Single nucleotide polymorphism genotyping using locked nucleic acid. *Experimental Review and Molecular Diagnostics* **3**: 27-38.
- Namroud, M.-C., Beaulieu, J., Juge, N., Laroche, J., and Bousquet, J. 2008. Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology* **17**: 3599-3613.

- Nazarenko, I., Bhatnagar, S., and Hohman, R. 1997. A closed tube format for amplification and detection of DNA based on energy transfer. *Nucleic Acids Research* **25**: 2516.
- Negrini, R., Nicoloso, L., Crepaldi, P., Milanesi, E., Colli, L., Chegiani, F., Pariset, L., Dunner, S., Leveziel, H., Williams, J., and Marsan, P.A. 2008. Assessing SNP markers for assigning individuals to cattle populations. *Animal Genetics* **40**: 18-26.
- Ng, P., Tan, J., Ooi, H., Hong, S., Lee, Y.L., Chiu, K.P., Fullwood, M.J., Srinivasan, G.K., Perbost, C., Du, L., Sung, W.-K., Wei, C.-L., and Ruan, Y. 2006. Multiplex sequencing of paired-end tags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Research* **34**(12): E84.
- Nicod, J.-C., and Largiadere, C.R. 2003. SNPs by AFLP (SBA): a rapid SNP isolation strategy for non-model organisms. *Nucleic Acids Research* **31**(5): e19.
- Nielsen, R. 2004. Population genetic analysis of ascertained SNP data. *Human Genomics* **1**(3): 218-224.
- Nielsen, R., and Signorovitch, J. 2003. Correcting for ascertainment biases when analyzing SNP data: Applications to the estimation of linkage disequilibrium. *Theoretical Population Biology* **63**: 245-255.
- Nieto, C., Piron, F., Dalmais, M., Marco, C.F., Moriones, E., Gomez-Guillamon, M.L., Truniger, V., Gomez, P., Garcia-Mas, J., Aranda, M.A., and Bendahmane, A. 2007. EcoTILLING for the identification of allelic variants of melon eIF4E, a factor that controls virus susceptibility. *BMC Plant Biology* **7**: 34-42.

- Novaes, E., Drost, D.R., Farmerie, W.G., Jr, G.J.P., Grattapaglia, D., Sederoff, R.R., and Kirst, M. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**(312): 1-14.
- Novak, J.P., Sladek, R., and Hudson, T.J. 2002. Characterization of variability in large-scale gene expression data: Implications for study design. *Genomics* **79**(1): 104-113.
- O'Malley, K.G., Camara, M.D., and Banks, M.A. 2007. Candidate loci reveal genetic variation between temporally divergent migratory runs of Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology* **16**(23): 4930-4941.
- Ogden, R. 2008. Fisheries forensics: the use of DNA tools for improving compliance, traceability and enforcement in the fishing industry. *Fish and Fisheries* **9**: 462-472.
- Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J., and Zwick, M.E. 2007. Microarray-based genomic selection for high-throughput resequencing. *Nature Methods* **4**: 907-909.
- Oleykowski, C.A., Mullins, C.R.B., Godwin, A., and Yeung, A.T. 1998. Mutation detection using a novel plant endonuclease. *Nucleic Acids Research* **26**(20): 4597-4602.
- Olivier, M. 2005. The Invader assay for SNP genotyping. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **573**(1-2): 103-110.

- Paduan, K.S., and Ribolla, P. 2008. Characterization of eight single nucleotide polymorphism markers in *Aedes aegypti*. *Molecular Ecology Resources* **9**(1): 114-116.
- Patanjali, S.R., Parimoo, S., and Weissman, S.M. 1988. Construction of a uniform-abundance (normalized) cDNA library. *Proceedings of the National Academy of Sciences* **88**: 1943-1947.
- Patterson, N., and Gabriel, S. 2009. Combinatorics and next-generation sequencing. *Nature Biotechnology* **27**(9): 826-827.
- Pavy, N., Parsons, L.S., Paule, C., MacKay, J., and Bousquet, J. 2006. Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs. *BMC Genomics* **7**: 174.
- Pella, J., and Masuda, M. 2001. Bayesian method for analysis of stock mixtures from genetic characters. *Fishery Bulletin* **99**: 369-376.
- Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., and Boyce-Jacino, M. 1999. Mining SNPs from EST databases. *Genome Research* **9**(2): 167-174.
- Polanski, A., and Kimmel, M. 2003. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* **165**: 427-436.
- Pop, M., and Salzberg, S.L. 2008. Bioinformatics challenges of new sequencing technology. *Trends in Ecology and Evolution* **24**(3): 142-149.

- Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F., Gao, Y., Church, G.M., and Shendure, J. 2007. Multiplex amplification of large sets of human exons. *Nature Methods* **4**: 931-936.
- Prabhu, S., and Pe'er, I. 2009. Overlapping pools for high-throughput targeted resequencing. *Genome Research* **19**: 1254-1261.
- Primmer, C.R., Borge, T., Lindell, J., and Saetre, G.P. 2002. Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. *Molecular Ecology* **11**: 603-612.
- Ragoussis, J. 2009. Genotyping technologies for genetic research. *Annual Review of Human Genetics* **10**: 117-133.
- Renaut, S., Nolte, A.W., and Bernatchez, L. 2010. Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). *Molecular Ecology* **19**(Suppl. 1): 115-131.
- Rigola, D., Oeveren, J.v., Janssen, A., Bonne, A., Schneiders, H., Poel, H.J.A.v.d., Orsouw, N.J.v., Hogers, R.C.J., Both, M.T.J.d., and Eijk, M.J.T.v. 2009. High-throughput detection of induced mutations and natural variation using KeyPoint™ technology. *PloS ONE* **4**(3): e4761.
- Roden, S.E., Dutton, P.H., and Morin, P.A. 2009. AFLP fragment isolation technique as a method to produce random sequences for single nucleotide polymorphism

- discovery in the green sea turtle, *Chelonia mydas*. *Journal of Heredity* **100**(3): 390-393.
- Rokas, A., and Abbot, P. 2009. Harnessing genomics for evolutionary insights. *Trends in Ecology and Evolution* **24**(4): 192-200.
- Rosenblum, E.B., Belfiore, N.M., and Moritz, C. 2007. Anonymous nuclear markers for the eastern fence lizard, *Sceloporus undulatus*. *Molecular Ecology Notes* **7**(1): 113-116.
- Rosenblum, E.B., and Novembre, J. 2007. Ascertainment bias in spatially structured populations: A case study in the eastern fence lizard. *Journal of Heredity* **98**(4): 331-336.
- Rusk, N. 2009. Cheap third-generation sequencing. *Nature Methods* **6**(4): 244-245.
- Ryynanen, H., and Primmer, C. 2006. Single nucleotide polymorphism (SNP) discovery in duplicated genomes: intron-primed exon-crossing (IPEC) as a strategy for avoiding amplification of duplicated loci in Atlantic salmon (*Salmo salar*) and other salmonid fishes. *BMC Genomics* **27**: 192.
- Sanchez, C.C., Smith, T.P., Wiedmann, R.T., Vallejo, R.L., Salem, M., Yao, J., and III, C.E.R. 2009. Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics* **10**: 559-556.
- Sarin, S., Parbhu, S., O'Meara, M.M., Pe'er, I., and Hobert, O. 2008. *Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. *Nature Methods* **5**(10): 865-867.

- Schaeffer, L. 2006. Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics* **123**(4): 218-223.
- Schaschl, H., Wandeler, P., Suchentrunk, F., Obexer-Ruff, G., and Goodman, S. 2006. Selection and recombination drive the evolution of MHC class II DRB diversity in ungulates. *Heredity* **97**: 427-437.
- Scheet, P., and Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* **78**: 629-644.
- Schlotterer, C., and Tautz, D. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Research* **20**: 211-215.
- Seddon, J., Parker, H., Ostrander, E., and Ellegren, H. 2005. SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. *Molecular Ecology* **14**: 503-511.
- Shamay, A., Fang, J., Pollak, N., Cohen, A., Yonash, N., and Lavi, U. 2006. Discovery of c-SNPs in *Anemone coronaria* L and assessment of genetic variation. *Genetic Resources and Crop Evolution* **53**: 821-829.
- Shen, R., Fan, J.-B., Campbell, D., Chang, W., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Garcia, E.W., McBride, C., Steemers, F., Garcia, F., Kermani, B.G., Gunderson, K., and Oliphant, A. 2005. High-throughput SNP genotyping on universal bead arrays. *Mutational Research* **573**: 70-82.
- Shendure, J., and Ji, H. 2009. Next-generation DNA sequencing. *Nature Biotechnology* **26**(10): 1135-1145.

Smith, C., Templin, W., Seeb, J., and Seeb, L. 2005a. Single nucleotide polymorphisms (SNPs) provide rapid and accurate estimates of the proportions of U.S. and Canadian Chinook salmon caught in Yukon River fisheries. *North American Journal of Fisheries Management* **25**: 944-953.

Smith, C.T., Baker, J., Par, L., Seeb, L.W., Elfstrom, C., Abe, S., and Seeb, J.E. 2005b. Characterization of 13 single nucleotide polymorphism markers for chum salmon. *Molecular Ecology Notes* **5**(2): 259-262.

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

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

Sokurenko, E.V., Tchesnokova, V., Yeung, A.T., Oleykowski, C.A., Trintchina, E., Hughes, K.T., Rashid, R.A., Brint, J.M., Moseley, S.L., and Lory, S. 2001. Detection of simple mutations and polymorphisms in large genomic regions. *Nucleic Acids Research* **29**(22): 1-8.

Stephens, M., and Donnelly, P. 2003. A comparison of Bayesian methods for the haplotype reconstruction from population genetic data. *American Journal of Human Genetics* **73**: 1162-1169.

Stephens, M., Smith, N., and Donnelly, P. 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**: 978-989.

- Stone, J.R., and Wray, G.A. 2001. Rapid evolution of *cis*-regulatory sequences via local point mutations. *Molecular Biology and Evolution* **18**(9): 1764-1770.
- Sunnucks, P. 2000. Efficient genetic markers for population biology. *Trends in Ecology and Evolution* **15**: 199-203.
- Sunnucks, P., Wilson, A., Beheregaray, L., Zenger, K., French, J., and Taylor, A. 2000. SSCP is not so difficult: the application and utility of single-stranded conformation polymorphism in evolutionary biology and molecular ecology. *Molecular Ecology* **9**: 1699-1710.
- Syvanen, A.-C. 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics* **2**: 930-942.
- Takatsu, K., Yokomaku, T., Kurata, S., and Kanagawa, T. 2004. A new approach to SNP genotyping with fluorescently labeled mononucleotides. *Nucleic Acids Research* **32**: e60.
- Tenesa, A., Navarro, P., Hayes, B.J., Duffy, D.L., Clarke, G.M., Goddard, M.E., and Visscher, P.M. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17**: 520-526.
- Till, B., Comai, L., and Henikoff, S. 2007. Tilling and Ecotilling for crop improvement. In: *Genomics-Assisted Crop Improvement*. Springer, Netherlands.
- Till, B., Zerr, T., Bowers, E., Greene, E., Comai, L., and Henikoff, S. 2006. High-throughput discovery of rare human nucleotide polymorphisms by Ecotilling. *Nucleic Acids Research* **34**(13): 1-12.

- Till, B.J., Reynolds, S.H., Weil, C., Springer, N., Burtner, C., Young, K., Bowers, E., Codomo, C.A., Enns, L.C., Odden, A.R., Greene, E.A., Comai, L., and Henikoff, S. 2004. Discovery of induced point mutations in maize genes by TILLING. *BMC Plant Biology* **4**: 1-8.
- Toth, A.L., Varala, K., Newman, T.C., Miguez, F.E., Hutchison, S.K., Willoughby, D.A., Simons, J.F., Egholm, M., Hunt, J.H., Hudson, M.E., and Robinson, G.E. 2007. Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science* **318**(5849): 441-444.
- Tsuchihashi, Z., and Dracopoli, N.C. 2002. Progress in high throughput SNP genotyping. *The Pharmacogenomics Journal* **2**: 103-110.
- van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., and Sonstegard, T.S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods* **5**: 247-252.
- Vera, J.C., Wheat, C.W., Fescemyer, H.W., Frilander, M.J., Crawford, D.L., Hanski, I., and Marden, J.H. 2007. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology* **17**(7): 1636-1647.
- Vignal, A., Milan, D., Cristobal, M.S., and Eggen, A. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution* **34**: 275-305.

- Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T.v.d., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kulper, M., and Zabeau, M. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* **23**(21): 4407-4414.
- Vysotskaia, V.S., Curtis, D.E., Voinov, A.V., Kathir, P., Silflow, C.D., and Lefebvre, P.A. 2001. Development and characterization of genome-wide single nucleotide polymorphism markers in green alga *Chlamydomonas reinhardtii*. *Plant Physiology* **127**: 386-389.
- Wakeley, J., Nielsen, R., Neen, S., Liu-Cordero, and Ardlie, K. 2001. The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. *American Journal of Human Genetics* **69**: 1332-1347.
- Wall, J.D., and Pritchard, J.K. 2003. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics* **4**: 587-597.
- Wang, G.-X., Tan, M.-K., Rakshit, S., Saitoh, H., Terauchi, R., Imaizumi, T., Ohsako, T., and Tominaga, T. 2007. Discovery of single-nucleotide mutations in acetolactate synthase genes by EcoTILLING. *Pesticide Biochemistry and Physiology* **88**: 143-148.
- Wang, J., Chuang, K., Ahluwalia, M., Patel, S., Umblas, N., Mirel, D., Higuchi, R., and Germer, S. 2005. High-throughput SNP genotyping by single-tube PCR with T_m -shift primers. *Biotechniques* **39**(6): 885-892.
- Weir, B. 2003. Uses of DNA and genetic markers for forensics and population studies. *Theoretical Population Biology* **63**: 171-172.

- Wilding, C.S., Butlin, R.K., and Grahame, J. 2001. Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *Journal of Evolutionary Biology* **14**: 611-619.
- Williams, J.G., Kubelik, A.R., Livak, K.J., Rafalski, J.A., and Tingey, S.V. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research* **18**(22): 6531-6535.
- Willing, E.-M., Bentzen, P., van Oosterhout, C., Hoffmann, M., Cable, J., Breden, F., Weigel, D., and Dreyer, C. 2010. Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. *Molecular Ecology* **19**(5): 968 - 984.
- Xu, H., Sha, M.Y., Wong, E.Y., Uphoff, J., Xu, Y., Treadway, J.A., Truong, A., O'Brien, E., Asquith, S., Stubbins, M., Spurr, N.K., Lai, E.H., and Mahoney, W. 2003. Multiplexed SNP genotyping using the Qbead system: a quantum dot-encoded microsphere-based assay. *Nucleic Acids Research* **31**(8): e43.
- Xu, J.-Y., Xu, G.-B., and Chen, S.-L. 2009. A new method for SNP discovery. *Biotechniques* **46**(3): 201-208.
- Xue, Y., Wang, Q., Long, Q., Ng, B.L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdallah, Z., Zhao, Y., Asan, MacArthur, D.G., Quail, M.A., Carter, N.P., Yang, H., and Tyler-Smith, C. 2009. Human Y chromosome based-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology* **19**: 1453-1457.

- Yang, B., Wen, X., Kosali, N.S., Oleykowski, C.A., Miller, C.G., Kulinski, J., Besack, D., Yeung, J.A., Kowalski, D., and Yeung, A.T. 2000. Purification, cloning and characterization of the CEL1 nuclease. *Biochemistry* **39**(13): 3533-3541.
- You, Y., Moreira, B., Behlke, M., and Owczarzy, R. 2006. Design of LNA probes that improve mismatch discrimination. *Nucleic Acids Research* **34**: e60.
- Zabeau, M., and Vos, P. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* **23**(21): 4407-4414.
- Zenger, K.R., Khatkar, M.S., Cavanagh, J.A.L., Hawken, R.J., and Raadsma, H.W. 2007. Genome-wide genetic diversity of Holstein Friesian cattle reveals new insights into Australian and global population variability, including impact of selection. *Animal Genetics* **38**(1): 7-14

Chapter 2

A Cross-validation Approach to Evaluate Genetic Baselines and Approximate the Number of Informative SNP Loci Needed for Mixed-Stock Analysis²

² Michael R. Garvin, Michele M. Masuda, Jerome J. Pella, Rachael R. Riley, Richard L. Wilmot, Vladimir Brykov, and Anthony J. Gharrett. Canadian Journal of Fisheries and Aquatic Sciences (submitted)

Abstract

Samples taken from nature for study are often mixtures from multiple populations, which frequently require mixed stock analysis (MSA) to estimate their composition.

Development of genetic baselines for MSA requires evaluations to predict their performance. Previous evaluation methods to evaluate baselines used simulated mixtures and were overly-optimistic of baseline power. More objective methods are available, but they do not accommodate potentially informative haploid data, and are based solely on maximum likelihood methods. We evaluated a combined single nucleotide polymorphism (SNP) and microsatellite baseline for chum salmon (*Oncorhynchus keta*) with a method called 'leave ten percent out cross validation' (LTO), which avoids optimism, uses observed genotypes, accepts haploid and diploid data, and applies either Bayesian or maximum likelihood-based methods. From simulated SNP data, we used logistic regression to estimate the number of SNP loci necessary to achieve a specified rate of correct assignment and provide a guide to develop genetic baselines.

Introduction

Mixed-stock analysis (MSA) uses multi-locus genotype data to estimate the composition of a mixture or to assign individuals to a stock of origin (Manel et al. 2005, Pella and Milner 1987). These types of analyses have been used for the study, management, and conservation of many species including turtles (Bowen et al. 2007), cattle (Negrini et al. 2008), elephants (Wasser et al. 2004), and salmon (Griffiths et al. 2010). The application of MSA generally requires a reference genetic baseline that includes samples from all the distinct stocks that may contribute to the mixture. For MSA, the term “stock” typically refers to a group of individuals defined for conservation or resource management purposes and is not necessarily a breeding population.

Applications that require samples from large numbers of stocks may be prohibitively costly, but a tendency for genetic similarity among neighboring stocks can compensate for incomplete baselines. In this case, regional representatives are chosen, and individuals in the mixture from non-baseline stocks will likely be correctly assigned to their geographic regions of origin (e.g. Beacham et al. 2009). The individuals sampled from each baseline stock are genotyped at numerous genetic loci, as is each individual in a sample from a mixture. Analyses are then conducted either to estimate the proportionate contribution of each baseline stock to the mixture or to assign each individual in the mixture sample to a baseline stock of origin.

The accuracy and precision of the genetic baselines are typically evaluated after their assembly and reevaluated when (1) new loci are added to the baseline; (2) new stocks are added to the baseline; or (3) only a subset of the baseline is used for MSA,

which for some applications may be desirable to conserve time and resources. Accuracy and precision of the genetic baselines refer to the expected performance of a statistical method trained with samples from the baseline either to estimate the stock composition of a mixture or to classify individuals from the mixture to their source stocks. During standard application of MSA, the individuals from the mixtures are obtained independently after baseline sampling. Therefore, the objective of baseline evaluation is to predict the performance of composition estimation and individual assignment of future mixtures from only the current baseline samples.

A widely used method to evaluate a genetic baseline is to repeatedly simulate from the baseline information samples of multi-locus genotypes to create hypothetical mixtures of a specified stock composition, and then to estimate the compositions of those mixtures from either the actual baseline samples or simulated baseline samples with the conditional maximum likelihood method. The simulations and estimation are accomplished with software programs such as SPAM (Debevec et al. 2000) or GMA (Kalinowski 2003). Recently, the methods used to create and estimate the composition of simulated mixtures with these programs were shown to overstate the accuracy and precision of baselines (Anderson et al. 2008). The optimism arises because sampling errors in the baseline result in apparent greater genetic divergence among stocks than actually exists (error-enhanced divergence).

The use of '100% mixtures' in which the entire mixture is composed of individuals from a single reporting group is commonly applied to baseline evaluation (e.g. Beacham et al. 2009; Seeb et al. 2011). However, the bias in estimates of stock

proportions is maximized for 100% mixtures because the repeat estimates concentrate inside and at the boundaries of the feasible space (i.e. you cannot have more than 100% nor less than 0% contribution), and their variances are also reduced for the same reason. With at least modest success in estimating the unknown stock proportions, i.e. the prior source probabilities of individuals in the mixture, the individuals in a 100% mixture are more easily identified to source than for a mixture nearer to equal proportions because the correct and unknown 100% prior is fully informative of individual sources (only one stock is possible) while the equal-proportions prior is uninformative (all stocks are equally probable). Ultimately, successful stock composition estimation requires identification of the sources of individuals in the mixture; and in this regard, equal-proportions mixtures are more challenging than 100% mixtures.

Several recent studies used ‘proof tests’ to evaluate the performance of genetic baselines (Habicht et al. 2010, Seeb et al. 2011, Templin et al. 2011), in which a few hundred individuals are removed from the baseline as a test mixture to be evaluated with the now reduced baseline. This use of baseline holdouts to create mixtures completely avoids the optimism caused by the simulation of mixture genotypes from baselines as discussed by Anderson et al. (2008), but it requires that baseline samples sizes are sufficiently large to accommodate the removal of a few hundred samples for the mixtures. The proof tests may result in understating baseline power because the reduced baseline sample sizes may be less effective at distinguishing sources. Although proof tests have been used only with 100% mixtures, the method could easily be adapted to

accommodate more challenging mixtures that have nearer equal stock proportions as we describe later.

The School of Fisheries and Ocean Sciences at the University of Alaska Fairbanks (UAFSOS), the Alaska Department of Fish and Game (ADF&G) and the National Marine Fisheries Service Auke Bay Laboratory (NMFSABL) are co-developing a genetic baseline for chum salmon (*Oncorhynchus keta*), whose incidental bycatch in the Bering Sea pollock fishery is of concern for fisheries management in Alaska (Gisclair 2009). The baseline consists of markers that are coded in both nuclear and mitochondrial DNA. Among nuclear markers, microsatellite loci commonly have large numbers of alleles, which provide more opportunities for genetic drift to reveal differences among populations than bi-allelic single nucleotide polymorphisms (SNPs). On the other hand, mitochondrial markers can be useful for MSA because they can show stronger levels of divergence among populations than nuclear markers. This is because mitochondrial DNA represents a smaller effective population size than nuclear DNA because it is haploid, undergoes no recombination, and is maternally inherited (Billington 2003). Indeed, strong divergence has been demonstrated in chum salmon (Garvin et al. 2010; $\phi_{ST} > 0.3$; Moriya et al. 2006). In addition, some species (including salmon) demonstrate positive directional selection in the mitochondrial genome (Brandt et al. 2012; Foote et al. 2010; Garvin et al. 2011; Grossman et al. 2004; Scott et al. 2010), which could further increase divergence among populations.

The use of SNPs in both nuclear and mitochondrial DNA is relatively new to MSA; but it is becoming more common because laboratory genotyping methods are

simple and their data are easily shared among users, although the number of SNPs required to address mixture problems usually far exceeds the number of multi-allelic microsatellite loci that would provide the same level of discrimination. The development of baselines from SNP data involves four steps: (1) SNP discovery, (2) SNP development, (3) SNP selection, and (4) SNP baseline evaluation. Most baseline development projects use Sanger or next-generation sequencing to discover tens or even thousands of SNPs from which a subset is developed into a laboratory assay and used to genotype baseline samples. In an attempt to reduce the number of SNPs to be used for MSA, a subset of the most promising are high-graded and then evaluated for precision and accuracy, usually with simulations of mixture samples from the baseline itself.

Anderson (2010) cautioned that a systematic upward bias in predicted accuracy can be introduced into baseline evaluation when SNPs are high-graded if the same samples are used to choose loci and to evaluate the new baseline. This bias is distinct from error-enhanced divergence described earlier by Anderson et al. (2008) and results from the fact that divergence estimates from a sample are larger or smaller than the true value, which can be corrected by a regression to the mean. SNP loci that are chosen based on high divergence estimates will likely not perform as well with different baseline and mixture samples, and loci that were excluded may have performed better than expected. High-grading bias is also different than so-called ‘ascertainment bias’, which occurs when too few individuals are used in the ascertainment panel for SNP discovery (although this is best described as sampling error, we will use the term most often reported in the literature).

Anderson (2010) discussed two double cross-validation methods to avoid high-grading bias in the context of MSA. In ‘Simple Training and Holdout’ (STH) the baseline samples from each stock are randomly divided into a training set and a holdout set. The training sets are used to identify the most informative loci and to constitute a baseline for MSA. The genotype data in the holdout sets are used only to create mixtures that are evaluated by MSA with the baseline of training sets. Because the STH method may substantially limit the data available for estimation of the frequencies of the genotypes from the mixture in the separate baseline stocks as required in MSA, Anderson (2010) suggests a modification called the ‘Training Holdout Leave-one-out’ (THL) method. In this method, the training sets are used to select the SNPs, the test mixtures are created from the holdout sets, and then the test mixtures are analyzed by MSA with the combined training and holdout sets for the baseline. However, when the test mixtures are analyzed in THL, a leave-one-out (LOO) rule prevents repeat use of genotypes from the holdout set in both the mixture and the baseline, which accounts for the error-enhanced divergence inherent in baselines.

The program ONCOR (Anderson et al. 2008) has a simulation feature useful for SNP baseline evaluation, i.e., step 4 of the development of SNP-based baselines. Importantly, the modified maximum likelihood estimation algorithm in ONCOR avoids the over-optimism of SPAM and GMA because it uses the LOO rule. However, ONCOR does not allow the user to perform THL in which mixture samples are simulated from only the holdout sets of the baseline and then evaluated with the full baseline. Instead, ONCOR simulates mixtures from the full baseline provided and then evaluates them

without a distinction between training and holdout sets. ONCOR is also restricted to the conditional maximum likelihood method, which provides frequentist confidence intervals for mixture proportions through bootstrap resampling, whereas Bayesian methods could provide more easily understood probability interval statements for all unknowns. Lastly, but importantly, no computer program that simulates mixture genotypes is publicly available to objectively evaluate the accuracy and precision of a baseline that includes haploid data. Analyses that use phenotypic data (e.g. Nolte and Sheets 2005) would also benefit from a method that accommodates haploid data because those data are evaluated in the same manner as haploid data within the framework of MSA.

We constructed an alternative system to develop and evaluate SNP baselines that reduces ascertainment bias introduced during the discovery step, reduces the costs associated with the development step as well as bias introduced during the selection step (Garvin and Gharrett 2007), and either reduces or eliminates optimistic bias in baseline evaluation from error-enhanced divergence and high-grading. Our Eco-TILLING method essentially combines the discovery and selection steps (1 and 3) of the baseline development into a single step. Our ascertainment panel consisted of 480 individuals that represented 12 populations across a geographic range, which was used to survey each target DNA sequence. Ascertainment bias was reduced because 40 individuals per stock were surveyed for genetic variants compared to a handful with standard sequencing methods. Costs were reduced because only informative SNPs were subsequently developed into laboratory assays.

In this study, we use a portion of the chum salmon baseline that is being developed to assess a method we call ‘leave-ten-percent-out cross validation’ (LTO) as an alternative to genotype simulations used in the program ONCOR to evaluate the precision and accuracy of genetic baselines for MSA. LTO is derived from the K-fold cross validation method of classification statistics (Hastie et al. 2001). K-fold cross validation, introduced by Geisser (1975), is in widespread use among analysts concerned with classification applications such as gene expression microarray experiments (Slonim 2002), landscape ecology (Boyce et al. 2002), and clinical medicine (Hess et al. 2006); and it has been recommended for use in MSA applications for fisheries (Waples 2010). Our LTO method reduces or avoids optimism of baseline performance to estimate mixture proportions and make individual assignments, accommodates haploid and diploid data, and restricts the mixture genotypes to those observed for individuals in actual samples from separate baseline stocks rather than simulating them assuming Hardy-Weinberg and Linkage Equilibrium (HWLE). In addition, because our LTO method uses full multi-locus genotypes rather than simulated ones, the analysis can be accomplished by either Bayesian or maximum likelihood (ML) estimation methods, which we use here for a direct comparison of the two.

Our baseline development method may have introduced some bias into our current chum salmon baseline evaluation because we used a subset of the same samples both to discover and high-grade SNP loci. Here we explore correction for bias with a Bayesian estimation method that shrinks observed allele frequencies in baseline stocks toward a better-anchored central value called the prior mean. The correction either

reduces or eliminates possible bias introduced with small sample sizes and stocks with low divergence discussed by Anderson *et al.* (2008), as well as any high-grading bias that we may have introduced with our baseline development method.

Lastly, we address the problem of predicting the number of SNP loci required to achieve specified baseline performance. Many baseline development projects are concerned with conversion from microsatellite-based baselines to SNP-based baselines. Attempts to quantify the number of SNPs that provide equivalent discrimination to microsatellite loci are generally based on either equivalent numbers of alleles (Kalinowski 2002) or comparisons of SNPs and microsatellites surveyed on the same individuals (Narum *et al.* 2008, Santure *et al.* 2010). Methods that compare individuals genotyped with both microsatellites and a sufficient number of informative SNPs will likely provide the more accurate comparison. We describe how to use the information from a baseline to generate increasing numbers of simulated SNP loci. The additional simulated loci are then evaluated with LTO, and the resulting data are analyzed by logistic regression to extrapolate the number of SNPs necessary to provide a given degree of accuracy. The simulated SNP loci are generated to realistically reproduce the increase in power from development of additional informative SNP loci by the same discovery methods used to develop the actual SNP loci, which in this case were the methods of Garvin and Gharrett (2007). This method can be adapted to any species for any number of loci and stocks.

Materials and Methods

Generation of test baselines and stock mixtures

Samples from 74 chum salmon stocks that range across the Pacific Rim were grouped into 25 continuous reporting groups (Figure 2.1, Table 2.1) based on geography and management areas delineated by ADF&G, NMFSABL, and the Department of Fisheries and Oceans, Canada (DFO). The stocks were further consolidated into 14 reporting groups determined by combining groups among which there were higher frequencies of misassignments. Each individual in the baseline was genotyped with 23 SNPs, representing 12 loci, and nine microsatellite loci (Table 2.2). The SNP loci were developed to maximize divergence among stocks; we used a single population to represent a geographic reporting group for both SNP discovery and to high-grade loci with Eco-TILLING (Garvin and Gharrett 2007). Importantly, the representative population for the reporting group was not always the same for all SNPs discovered, and less than half of the sample was used for the discovery and high-grading step.

For our SNP discovery efforts, we amplified targeted DNA sequences with pools of DNA from a regional representative stock and chose potentially informative loci according to estimated allele frequencies. Those loci were then evaluated with our LTO method with all of the baseline samples, which included the first portions of the sample used to choose the loci (the training set), the second portion of the sample (the holdout set), and complete samples from other populations within the geographic region that the discovery sample represented (additional holdout samples).

The mitochondrial SNPs were validated from previous RFLP work (Garvin et al. 2010) and other SNPs were reported in a method in which the phase of linked SNPs was determined empirically (Garvin and Gharrett 2010). This work is not meant as a report of the chum salmon baseline; rather, our interest here is to use this partial baseline to develop a method to evaluate genetic baselines. The full baseline will be published elsewhere.

We developed R code (www.R-project.org) to divide each baseline stock sample sequentially into ten equal parts (e.g. the first 10% were taken for Mixture 1, the second 10% for Mixture 2, etc.). We believe that the order of individuals in the sample was approximately random, although random order was not imperative. The individuals that composed one part from each stock were combined into a test mixture that included a total of 450 individuals, and the genetic information from the remaining nine parts was used as the baseline data in the mixture analyses performed with the programs BAYES (Pella and Masuda 2001) and SPAM (Debevec et al. 2000). Both the BAYES and SPAM analyses were repeated ten times; each time, a different one of the ten parts was used for the test mixture and the remaining nine parts served as the baseline. The LTO method guarantees that each individual is used in a test mixture once and in a test baseline nine times (except for 'remainder' individuals - below). The stock compositions of all test mixtures were identical.

Ideally all of the baseline sample sizes should be equal for this analysis so that each of the stocks would contribute equally to the mixture, which makes estimation more challenging than for 100% mixtures. In our baseline, the sample sizes of many stocks

differed and many were not evenly divisible by ten (Table 2.1). Therefore, the composition of the mixture was nearly proportional to baseline sample sizes. In addition, when we divided our baseline stock samples into ten equal parts, some individuals remained. Those individuals were added to each of the ten test baselines and not included in any of the test mixtures; every test baseline included these ‘remainder’ individuals. The sample sizes for the stocks in the ten test baselines that were used to resolve the ten test mixtures are approximately 90% as large as the original baseline sample sizes; and because we expect performance to improve with baseline sample size (Beacham et al. 2011), our evaluation may be conservative and slightly pessimistic when compared to analyses that use the complete baseline. The term ‘test dataset’ will be used to denote one of the ten ‘test mixtures’ and the associated ‘test baseline’.

Measurements of diversity

Many of the loci that we evaluated in this work are potentially informative for MSA because they demonstrate large divergence estimates among baseline populations. Several measurements of genetic diversity are often reported in the literature. We used GDA (Lewis and Zaykin 2001) to calculate F_{ST} (Weir and Cockerham 1984) locus by locus and overall, expected heterozygosity (H_e), and ϕ_{ST} , the haploid equivalent of F_{ST} (Excoffier et al. 1992) for the mitochondrial haplotype. Locus by locus and overall D_{EST} , or Jost’s D (Jost 2008) were calculated with the ‘adegenet’ package (Jombart 2008) in the R environment.

Bias may have been introduced into our baseline evaluation from several sources. Some of our baseline sample sizes are small (Table 2.1) and many populations demonstrate weak genetic structure, which can inflate divergence estimates, i.e. error-enhanced divergence (Anderson et al. 2008). Bias could have been introduced from our baseline development method because the training samples from regional representatives that were used to discover the SNPs were also used together with the holdout samples to evaluate them, i.e. high-grading bias. Bayesian methods for MSA revise the observed allele frequencies from the baseline genotypes by shrinking them toward a baseline central value among stocks (Pella and Masuda 2001), which reduces both types of bias. To demonstrate the shrinkage effects, we compared apparent diversity before and after the revision of the allele frequencies. Before revision, we calculated G_{ST} values for each locus with the methods of Nei and Chesser (1983) with the observed allele frequencies from the entire baseline; and then for each of the 10 baselines that were created during LTO. After revision, we performed the same calculations with the baseline posterior means of allele frequencies, computed as a weighted average of the original frequencies and the prior grand mean (Eq. 4 in Pella and Masuda 2001).

Baseline summary statistics

The results from BAYES and SPAM MSA for the ten test datasets provided a sample of 10 stock composition estimates for the mixtures of 74 baseline stocks. For BAYES, the posterior average for stock proportions from a mixture was the point estimate, and for SPAM, the conditional maximum likelihood estimate was chosen as the

point estimate. Regional compositions were obtained as sums over the point estimates for the individual stocks of the regional groups. Statistics computed from the experiment included the 10 point estimates of regional proportions and their means, variances, etc., as well as the observed bias and mean square errors from the true and known regional proportions.

An overall measure of bias and precision in combination is the mean squared error (*MSE*) of estimates of stock proportions, which is the average of squared errors from the true proportions under repeated sampling. The *MSE* also equals the sum of the squared bias plus the variance of a stock proportion estimator. An estimator with low *MSE* is desirable because this indicates low bias and low variance. An estimator with high *MSE* is undesirable and may have high bias, high variance, or both. Cochran (1963) noted that the effect of bias on accuracy is negligible if the absolute value of the bias is less than one-tenth of the standard deviation of the estimate; and that even with an absolute value of bias of up to one-fifth of the standard deviation, confidence statements would be only modestly affected.

The mean squared error for geographic region *g* (MSE_g) was calculated as the average of squared regional errors among the ten test datasets and is related to the variance of estimated proportions (s_g^2) and observed bias (b_g) by the equation,

$$MSE_g = \frac{1}{10} \sum_{i=1}^{10} (\hat{p}_{g,i} - p_g)^2 = \frac{1}{10} \sum_{i=1}^{10} (\hat{p}_{g,i} - \bar{p}_g)^2 + \frac{1}{10} \sum_{i=1}^{10} (\bar{p}_g - p_g)^2 = \frac{9}{10} s_g^2 + b_g^2,$$

where $\hat{p}_{g,i}$ is the estimated proportion for region *g* from the *i*th test dataset, p_g is the true and known proportion, \bar{p}_g is the average estimated proportion for region *g* among the ten

test datasets, $b_g = \bar{p}_g - p_g$ is the observed bias in the estimated proportion for region g , and s_g^2 is the sample variance of the estimated proportions for region g among the ten test datasets. Any of these statistics can be summed over regions for a summary value of the entire experiment.

Interpretation of summary statistics for the ten stock composition estimates from the test datasets requires caution. The ten stock composition estimates are statistically dependent because the ten test baselines used to analyze the ten test mixture samples overlap. Any pair of test baselines have in common approximately 89% of their individuals (eight out of nine parts of the original baseline samples are shared and the tenth part is assigned to the test mixture sample). Furthermore, each part of an original baseline sample plays two roles, once in a test mixture and nine times in a test baseline, which also induces statistical dependence among the ten stock composition estimates. Therefore, although the precision for any of the ten stock composition estimates (vectors of 74 stock proportions) from the test datasets can be computed from bootstrap resampling by SPAM and the posterior probability distribution from BAYES, precision of their overall average is unknown. Moreover, precision of the estimate of the proportion of the individuals in the mixture from any stock that are correctly identified to their source for any dataset could be similarly evaluated by SPAM and BAYES. However, precision of the overall average for proportion correctly identified across the 10 datasets is also unknown because of the statistical dependence.

BAYES LTO

The low information Dirichlet prior probability distribution for stock proportions (its weight during estimation counts as a single individual added to the mixture of 450 individuals) was set to equal proportions among the entire 74 stocks, even though the stocks were not equally represented in the mixtures or baselines. Three independent MCMC chains were run for each test mixture with different starting values for stock proportions. Each chain was started with 95% of the mixture contributed by one of the three major geographic regions: Asia, Western Alaska, or NE Pacific. The remaining 5% was split equally between the other two regions. Within the regions, all stocks contributed equally to these starting values.

The three MCMC chains were run with the BAYES program to obtain 400 000 samples of the unknowns (stock proportions and baseline genetic parameters) from their posterior distribution, and every 40th sample was saved (i.e. the thinning interval was set to 40). The first half of each chain was discarded as burn-in to remove dependence on starting values. The second halves of the three chains were combined to provide a total sample of 15 000 stock composition estimates from their posterior distribution. Gelman and Rubin (G & R) statistics were computed for each of the 25 reporting groups to determine if pooled samples from the three chains had converged to the posterior distribution for the regional composition (Gelman and Rubin 1992). In all analyses, the G & R statistics were less than 1.2, which is consistent with convergence of the chains.

For evaluation of the combined microsatellite and SNP baseline, we used the program BAYES to estimate the proportion contributed to the mixture by each of the 74

stocks. The regional composition was obtained as sums over the individual stocks of the regions. The BAYES program outputs posterior probabilities that an individual in the mixture came from the baseline stocks (the probabilities sum to 1). An individual can be assigned to one of the baseline stocks based on the highest source probability, or the proportion of times that the individual was assigned to each stock in the baseline can be reported. We used the latter output for this analysis.

SPAM LTO

We used the program SPAM (under the Rannala-Mountain model of baseline allele frequency distributions) to resolve the same ten test mixture samples based on their corresponding test baseline samples (Debevec et al. 2000). The 74 stocks were grouped into the same 25 and 14 reporting groups as were used for the BAYES analysis (Table 2.1).

SPAM simulation

Performance of the simulation mode in SPAM was also examined with this baseline. Although past practice has been to create 100% mixtures of simulated multi-locus genotypes, we wanted to compare mixtures of near-to-equal proportions of simulated multi-locus genotypes with corresponding mixtures of naturally-occurring multi-locus genotypes.

Because SPAM uses the same EM algorithm to analyze the mixtures, the only difference between SPAM simulations and SPAM LTO is the creation of the mixtures. If reduced bias and variance are erroneously introduced during the simulation of the multi-locus genotypes in the mixtures, then we should see different results between SPAM LTO and SPAM simulation. Therefore, the same 10 test baselines were used in SPAM's 'simulation' mode with the same 25 reporting groups as was done for the BAYES LTO and SPAM LTO analyses (Table 2.1). Mixture proportions for the simulations were set equal to the same stock proportions created when the baselines and mixtures were created with the LTO method. A sample size of 450 individuals was used to match our LTO method and the number of resamplings was set to 1000.

Simulated SNP loci: how many SNPs would equal the discriminatory power of the combined SNP and microsatellite baseline

To generate "new" simulated SNP loci, we used the same previously-described survey data on 74 chum salmon stocks that had 23 SNPs assayed per individual, which represented 12 loci (11 nuclear and one mitochondrial) to determine the discriminatory power of our baseline. We created new simulated loci for the individuals of the baseline samples by randomly drawing, with replacement, additional loci from the 11 empirically genotyped nuclear SNP loci for which data were available. We did not include the mitochondrial data for the creation of the simulated loci because the genetic material in the mitochondrion of an individual behaves as a single locus so generation of more than one "simulated" mitochondrial locus would not be biologically meaningful.

The method of drawing the SNP genotypes at these simulated loci guarantees that the loci are independent of the original loci and each other. A locus was randomly drawn from the 11 nuclear SNP loci and corresponding single-locus genotypes were generated for each original baseline individual by drawing a pair of alleles without replacement between the two draws (rather than sampling single-locus genotypes or full multi-locus genotypes) from the sample of its SNP baseline stock. Multi-locus SNP genotypes for individuals were generated by parallel independent sampling for additional simulated loci followed by concatenation of the outcomes. This strategy preserves the genetic information from our original, informative loci but adds the multi-locus variability that would be expected at more unlinked loci. The simulated SNP loci were appended to the original 11 SNP loci and the mitochondrial locus to create seven extended SNP sets with 20, 30, 40, 50, 60, 70 and 80 SNP loci. For example, eight new SNP loci were created and added to the 11 original SNP loci and the mitochondrial locus to create 20 locus genotypes. For the 30 SNP locus genotypes, 18 new SNP loci were created rather than simply adding 10 new SNP loci to the 20 SNP loci of the first round. After the simulated SNP genotypes for the original baseline individuals were generated for these extended SNP sets, the baseline samples were divided into ten equal test datasets and analyzed with BAYES LTO.

For this analysis, we calculated the proportion of the 450 individuals in the mixture that were correctly assigned to a reporting group of origin by the maximum *a posteriori* (MAP) rule (Pella and Masuda 2005). The MAP rule assigns an entire individual to the stock for which its posterior source probability is highest. The simple

MAP rule has the lowest misclassification rate if costs of misclassification are equal for stocks, and provides a good benchmark for improvement. The assignment of individuals to single stocks as whole units by the MAP rule contrasts with their fractional assignment to several stocks by the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) used in conditional maximum likelihood estimation by the programs SPAM, GMA, and ONCOR. The proportion of individuals correctly assigned to reporting group was calculated for each test mixture and the corresponding average of the ten test mixtures from the LTO was then calculated. For instances in which identical posterior source probabilities were calculated for an individual belonging to either of two reporting groups, we randomly assigned the individual to one of the groups. The number of identical probabilities that occurred was inconsequential and ranged from zero to five for a single test mixture. These accounted for only 0.3% of the total assignments for the ten test mixtures, and was not related to the number of loci (data not shown).

To determine the number of SNP loci needed to obtain a high degree of correct assignment, we used logistic regression to estimate the correct proportional assignment to reporting group (θ) for a given number of SNPs (X) based on the empirical performance of the extended SNP sets for 20 to 80 SNPs. The log odds regression model relating θ and X is:

$$\text{Ln}\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_i X_i,$$

where θ_i is the probability of correct assignment of a random individual to its reporting group, X_i is the number of SNP loci ($X_1 = 20, X_2 = 30, \dots, X_7 = 80$), β_0 is the intercept

parameter where $\theta_0 = \frac{e^{\beta_0}}{(1 + e^{\beta_0})}$, which is the probability of correct assignment when $X_0 = 0$ (i.e., no SNP loci are used), and β_1 is the slope, which equals the increase in log odds, i.e., $\text{Ln}\left(\frac{\theta}{1-\theta}\right)$, for a unit increase in the number of SNPs (Hosmer and Lemeshow 2000). Equivalently, the effect of a unit level increase in the number of SNPs is a multiplicative increase in the odds of correct assignment by a factor of e^{β_1} . We fit the model to the counts of correctly and incorrectly assigned individuals as related to the number of SNPs by the maximum likelihood method. The fit was performed in each of the ten cross validation datasets with the ‘glm’ function in the R environment with the family set to ‘binomial’, which provided estimates of β_0 and β_1 . Finally, the estimates from those parameters were used to calculate the number of SNPs necessary to achieve 90% correct assignment ($X^{90\%}$) and 95% correct assignment ($X^{95\%}$) for each of the ten datasets.

Results

Evaluation of combined microsatellite and SNP baseline

We estimated the stock composition and standard errors for each of the ten test datasets with the programs BAYES and SPAM. Ideally, the highest resolution with the largest number of reporting groups is desired, but this goal must be balanced with acceptable accuracy and precision of the group estimates, which can decline if divergence is weak among some groups. Stock composition estimates for multiple reporting groups from several different statistical analyses can be confusing to portray graphically. The

focus of this study is to compare MSA methods as well as to evaluate a genetic baseline. Therefore in order to present the results clearly, we display them in two formats.

The first display presents the mean stock composition for each analysis method (BAYES LTO, SPAM LTO, and SPAM Simulations) and compares them to the true value, which we know because we created the mixtures (Figure 2.2a). This provides comparisons among methods. Our second displays show baseline performances for each of the methods separately (Figures 2.2b, 2.2c, and 2.2d) with the mean stock group composition estimates paired with their true values and a reference line provided for perfect accuracy as in Anderson (2008). This provides a simple visual standard to evaluate the bias (i.e. the diagonal lines represents “100%” accuracy). Mean stock composition estimates are provided for both 25- (Figure 2.2) and 14-reporting group analyses (Figure 2.3).

The stocks that composed the 14 reporting groups were identified from misassignments among those that were poorly resolved when 25 regions were used. The poorly resolved stocks were combined into larger regions. The coarser aggregations may reduce misassignments to true regions but may not correspond with biological or management goals and may cause some optimism in predicted performance. Some of the 25 reporting groups show highly accurate levels of assignment (the filled circles are directly on the dotted line) whereas others do not (e.g. Lower Kuskokwim, which deviates from the perfect accuracy line; Figures 2.2b, 2.2c, 2.2d). The misassignments would likely be reduced by increasing the sample sizes for some stocks (Table 2.1), by

including more informative loci that show divergence among those stocks, or both (Beacham et al. 2011).

Evaluation of the bias and accuracy of estimates with BAYES and SPAM

Accuracy can be evaluated by estimating bias, either as the persistent or systematic error in estimates of stock proportions under potential repeat baseline and mixture sampling or by the corresponding misclassification rates of mixture individuals from particular baseline stocks. Misclassification rates are clearer in detecting shortcomings of the assessment of the mixture composition. Any MSA method includes assigning, fractionally or whole, individuals to their sources. Compensating errors in assignments can be obscured when the stock proportions are computed, but are exposed with misclassification rates. Precision can be measured by the variability in estimates of stock proportions among potential repeat baseline and mixture samples. Bias and precision can be evaluated for the assignment to a particular stock, group of stocks, or the entire baseline.

We quantified the overall accuracy of each of the three methods (BAYES LTO, SPAM LTO, and SPAM Simulation) by calculating the absolute difference between the mean of the ten test estimates of a regional proportion and the true value for each of the 25 reporting groups (Figure 2.4). A few regions had comparatively high estimates of bias in both the BAYES LTO and SPAM LTO analyses, but the SPAM LTO method had nearly 2-fold greater overall sum of absolute bias terms across the 25 reporting groups

$(\sum_{g=1}^{25} |b_g|)$ than BAYES LTO. In addition, the bias in the Lower Kuskokwim reporting group was much higher with SPAM LTO than BAYES LTO. We also included the size of each baseline sample because a positive association between bias and sample size was present in some reporting groups (e.g. Lower Yukon, Lower Kuskokwim, and Behm Canal), which was not anticipated and may be coincidental. As expected the simulated mixtures that were evaluated with SPAM consistently demonstrated low bias across regions when compared to the SPAM LTO method.

We evaluated the apparent precision of the three methods by calculating the standard deviation of the ten estimates for each of the 25 reporting groups (Figure 2.5). (Recall that the ten estimates for any method are not independent, which is why we say apparent precision). Many of the reporting groups that had high estimated bias also had high standard deviation with both the BAYES and SPAM LTO methods. However, BAYES LTO had higher standard deviation than SPAM LTO for the Lower Kuskokwim and Behm Canal reporting groups. The SPAM Simulation had very low standard deviation across all groups.

Finally, we combined the bias and variance estimates into the mean squared error (we report the square root of this value to provide equivalent measures across accuracy and precision estimates) (Figure 2.6). The BAYES and SPAM LTO methods were similar and variable among groupings; whereas the overly-optimistic SPAM Simulation showed a consistently low mean error over groupings. For an overall comparison, we summed the mean error over all regions for both 25 and 14 reporting groups (Table 2.3). The sum was lower with SPAM LTO for the 25 regional groups and lower with BAYES

LTO for the 14 regional groups. The SPAM Simulation showed the smallest sum for both 25 and 14 regional groups. Finally, the sum was smaller for all three methods when the 25 groups are reduced to 14.

Evaluation of potential bias in baseline

Bias could have been introduced into our evaluation because our baseline sample sizes were reduced by up to 10% when we removed individuals to create mixtures. The smaller baseline samples add uncertainty to the MSA, which would be compensated by opposing error-enhanced divergence. Bias could also have been introduced during our baseline development because we used some of the same samples to choose SNPs and to evaluate them (high-grading bias), although in most cases the overlap was minimal and our large sample sizes for the ascertainment panels likely offset some bias. Still, in order to determine if divergence estimates were artificially inflated, we compared locus-by-locus G_{ST} values for the original baseline with corresponding mean G_{ST} values computed for the ten BAYES LTO baselines (Figure 2.7). For some loci the G_{ST} values were indeed inflated, which suggests that bias was introduced from the smaller sample sizes that resulted from reduction of the baseline by 10% to create the mixtures. We compared the locus-by-locus G_{ST} values for the original baseline and the ten BAYES LTO baselines with allele frequencies that were shrunk toward their prior means from Eq. 4 in Pella and Masuda (2001) (Figure 2.7). For all diploid loci, the G_{ST} values that were calculated with the shrunken allele frequencies were smaller than the values calculated with the observed allele frequencies. By inference, the reduction in apparent divergence due to the

shrinkage computations in the Bayesian method corrects to some degree for the aforementioned bias.

How many SNP loci are needed to exceed or equal the combined SNP and microsatellite baseline?

We calculated the mean proportion of individuals correctly assigned by the MAP rule to the 25 reporting groups from the ten test data sets. This was done for 20, 30, 40, 50, 60, 70, and 80 extended SNP locus genotypes ('Empirical Data', Figure 2.8) and for the original combined microsatellite and SNP baseline (horizontal dashed line, Figure 2.8). A reference line is provided for a hypothetical 90% correct assignment rule for all individuals in the baseline (horizontal solid line). Between 50 and 60 informative SNP loci appear necessary to equal the combined SNP and microsatellite baseline. The mean of the proportion of correctly assigned individuals may be even higher with coarser scale of geographic regions.

The empirical relationship between the proportion correctly assigned and the number of SNP loci, however, did not appear to be asymptotic in the range we reported but predicted further improvement if even more SNP loci were used for MSA. Therefore, we fit an asymptotic curve with logistic regression to the empirical data for each of the ten extended datasets with 20, 30, 40, 50, 60, 70, and 80 SNP loci. The mean value for the slope and intercept parameters from these ten fitted models was used to estimate the relationship between the number of SNPs and the proportion of correctly assigned individuals (Figure 2.8). Because a correct assignment accuracy of 90% or greater is

routinely sought for MSA, we calculated the number of SNP loci needed to achieve 90% accuracy to region from the fitted curve of each extended dataset, which averaged 125 SNPs with a standard error of ± 5.4 . According to our model, 95% assignment accuracy could be achieved with 158 SNPs with a standard error of ± 11.4 SNPs with this set of markers and baseline. However, this number is the mean percent correct assignment over all reporting groups. Specific reporting groups may require more or fewer SNPs depending on the divergence among those stocks.

Discussion

Mixed stock analysis is an important tool for the conservation and management of numerous species: it uses information from DNA variation simply, can be applied to any organism, and provides a means with which to identify the origin of individuals sampled from mixtures with non-lethal sampling. Technological advances will likely continue to provide a growing wealth of genetic data as well as the ability to generate millions of genotypes rapidly and inexpensively (Ragoussis 2009, Larson-Cook et al. 2011). As a result, MSA will likely be increasingly used to manage and monitor an ever-wider array of species and stocks. Furthermore, the new capacity for economical, large-scale laboratory determinations of individual genotypes will enable an increase of both baseline and mixture samples, which should improve the accuracy and precision of MSA and provide more data to biologists. The numbers of individuals in baselines and mixtures are more likely to be limited by the cost to sample them, especially from source stocks but also from catches, rather than laboratory costs of genotyping.

The LTO method reduces or eliminates the optimism in assessments of baselines for MSA that are obtained with the simulation methods of GMA and SPAM. ONCOR is the primary alternative for more objective assessments of baselines, but it uses only diploid data. The LTO method can incorporate diploid data as well as haploid and phenotypic data, both of which can be informative. When the baseline samples are composed of the multi-locus genotypes of individuals from the source stocks, the LTO method provides several additional advantages to the alternative methods and computer programs for baseline assessment.

First, the LTO method uses the actual multi-locus genotypes from the available baseline samples to create both the test mixtures and their associated test baselines. Alternative methods commonly simulate test mixture individuals from allele frequencies in the actual baseline samples. The simulators usually include the Hardy-Weinberg and linkage equilibrium (HWLE) assumptions for the separate stocks, which could cause some distrust in the assessments when these conditions are suspect. Although the Bayesian and maximum likelihood estimation methods that we used in LTO include the HWLE assumptions, their validity is not necessary to evaluate the performance of baselines if the same baseline samples and estimation methods will be used in the future analyses. Furthermore, the use of full multi-locus genotypes makes double cross-validation with either STH or THL reasonably easy to perform if the available baseline samples are divided into training and holdout sets. Sorting the training and holdout sets among the ten baseline and mixture datasets is simple and practical when done with a computer. The training samples used to explore for SNPs can easily be kept from

mixtures so that the mixtures can be composed of only holdout individuals. The training and holdout sets can be combined in the test baselines so that the test baselines and mixtures never share individuals. Although we did not completely follow this recipe, we encourage its use in LTO for baseline evaluation in order to remove any potential for overly-optimistic assessments.

Second, each individual of the actual baseline occurs at most in one test mixture (except for the remainder individuals). Therefore, the assignments of baseline individuals are known, which can provide useful information that is unavailable with simulation methods. For example, an individual may be misassigned due to missing data from multiple loci, which would be easily identified by locating that individual in the database. Also, the destination of misassigned individuals may identify useful reporting groups that may not have been evident from simple geographic or management analyses or could reveal migration among stocks.

Third, and last, the LTO method can use Bayesian methods as well as maximum likelihood methods for the evaluation of mixture composition and assignment of individuals to their source. The Bayesian computations recognize the uncertainty in the allele frequencies at the loci and their potential to inflate the divergence among stocks. The conditional maximum likelihood method treats the allele frequencies as known and equal to the observed values in the baseline and thereby treats the divergence among stocks as known. Unlike maximum likelihood methods that use bootstrap confidence intervals to describe uncertainty in the mixture composition, Bayesian methods can easily provide posterior probability intervals for all the unknowns underlying the MSA.

Bayesian probability statements about unknowns are simple and easily understood, whereas frequentist confidence intervals are complex and can easily be misinterpreted. In addition, the methods in the program BAYES use allele frequency estimates for the baseline stocks that are shrunk toward the overall mean, which reduces bias that may result from small sample sizes, error-enhanced divergence among stocks, and bias from high-grading SNP loci. Further, this bias also may be reduced when stocks are aggregated into regional groups, which was discussed previously (Anderson et al. 2008). The Bayesian method developed by Pella and Masuda (2001) also uses information from the mixture genotypes to update the baseline during the analysis, whereas standard conditional maximum likelihood methods do not. The reason we could use Bayesian methods with their high computational cost is that only one or two repeat partitions of the baseline samples were necessary. Although we anticipate that a few repeat partitions will satisfy most user's needs for baseline evaluations, future experiences will be telling. If many repeat partitions were deemed necessary, the LTO evaluation would have to be limited to maximum likelihood estimation methods for the present.

In addition to genetic data, spatial, temporal, morphometric, and phenotypic data for individuals can be used to increase the ability to distinguish among stock sources (Barbee and Swearer 2007, Gomez-Diaz and Gonzales-Solis 2007, Nolte and Sheets 2005, Reich et al. 2008). The standard mathematical mixture model for MSA with genetic data used in BAYES, SPAM, and ONCOR does not include components for geographic information corresponding to the individuals in the mixture. For example, if chum salmon stocks differ in their geographical and temporal distributions in the Bering

Sea, the addition of the appropriate components to the mixture model like Reich and Bondell (2011) would better use this potentially powerful information. Enhancement in genetic data collection and mathematical methods for MSA has become critical as more users exploit limited natural resources and as conservation concerns arise with global climate change, ocean acidification, and risk of extirpation for many species (Barnosky et al. 2011).

The main limitations of LTO are that the number of replicate mixture samples of a specified stock composition is ten rather than much greater numbers available with simulations with SPAM, GMA, or ONCOR; and currently the method is less automated and requires more effort on the researchers' part, primarily to format the data, a task that should eventually be eliminated with additional computer code. The limited number of replicates allows sampling variation in performance statistics that could be reduced or eliminated by a larger number of replicates but also makes the amount of computation tractable. Theoretically, the number of replicates could be made unlimited by randomly partitioning the baseline samples into ten equal parts again and again, followed by the LTO computations. However, the amount of computation required, especially for BAYES, makes this approach impractical for now, even though it may seem worthwhile. In addition, the balance in our LTO method would be lost; instead of each individual occurring once in a test mixture sample and nine times in test baselines, its frequencies in test mixtures and baselines would be random.

To reduce the human effort we chose to evaluate the baseline with mixtures that included all baseline stocks rather than 100% mixtures, and considered a single mixture

sample size of 450 individuals rather than mixtures with ranges of possible sample sizes to achieve research goals. Satisfactory performance with the mixture composition that we examined implies satisfactory performance would occur with 100% mixtures. The sample size of the mixture is sufficiently large to ensure that the stock composition of any future sample of such size would be near that of the actual mixture regardless of its unknown value (Thompson 1992, Pella and Geiger 2009), and most of the remaining uncertainty in estimates of the composition of the stock mixture would be due to the limited discriminatory power of the genetic information.

An evaluation of the accuracy and precision of BAYES and SPAM revealed that the BAYES method had lower estimated bias but higher standard deviation than SPAM at geographical regions for which stocks are difficult to assign such as those from coastal western Alaska. Chum salmon from this geographic region are difficult to assign correctly but are of special interest for management and conservation because they are important for subsistence and commercial fisheries by rural Alaskan communities (Wolfe and Spaeder 2009). The inability to accurately assign individuals to this area may reflect their recent colonization of the area after the Last Glacial Maximum (Seeb and Crane 1999, Wilmot et al. 1994, chapter 3 of this thesis), high levels of gene flow (Olsen et al. 2010), or both. As was mentioned previously, if bias is less than one-fifth of the variance, confidence levels should be only modestly affected, which was the case for nine of the 25 reporting groups with BAYES LTO, but only three of the reporting groups with SPAM LTO. However, for either method, the amount of bias was at most 1%-2% and of limited practical concern for most applications, although for stocks of regions for which

source identification is especially challenging, such as the coastal western Alaskan stocks, this may not apply.

The development of genetic baselines can be costly in both time and resources if it is open-ended; that is, the number of markers needed to provide accurate composition estimates of mixtures or assignments of individuals is unknown prior to baseline development. Open-ended collection of genetic baselines has been the standard practice, and the number of SNPs needed has depended on the target species of interest, the amount of divergence among stocks, and the stocks that were included in the analysis (Smith et al. 2005, Elfstrom et al. 2006, Campbell and Narum 2008, Griffiths et al 2010, Campbell et al. 2012). We provide a method to estimate the number of genetic markers that will be needed for MSA given an initial small set of informative markers. We showed that, for chum salmon, about 60 informative SNPs would be equivalent to the nine microsatellite and 12 informative SNP loci in the baseline. Furthermore, a logistic regression analysis predicted that a baseline with about 125 informative SNPs would be needed to assign stocks with more than 90% accuracy to a group. Our method allows managers and scientists to place a direct cost on the accuracy of this chum salmon genetic baseline, and provides an estimate for SNP development in similar species.

Acknowledgements

We would like to thank Sharon Hall for her tireless work in the lab and isolating the DNA. We much appreciated the thorough and helpful readings of an earlier draft by Robin Waples and two anonymous reviewers. We would also like to thank several funding agencies for providing stipend and laboratory support: The Rasmuson Foundation, the Arctic–Yukon–Kuskokwim Sustainable Salmon Initiative (www.aykssi.org) awarded through the Bering Sea Fishermen’s Association, the University of Alaska Experimental Program to Stimulate Competitive Research (EPSCoR), and the U.S. National Oceanic and Atmospheric Administration (NOAA) Alaska Fisheries Science Center (AFSC). This work was also supported by a grant of the HPC resources from the Arctic Region Supercomputing Center. The findings and conclusions presented by the authors, however, are their own and do not necessarily reflect the views or positions of the reviewers, funding agencies, or the University of Alaska Fairbanks, School of Fisheries and Ocean Sciences.

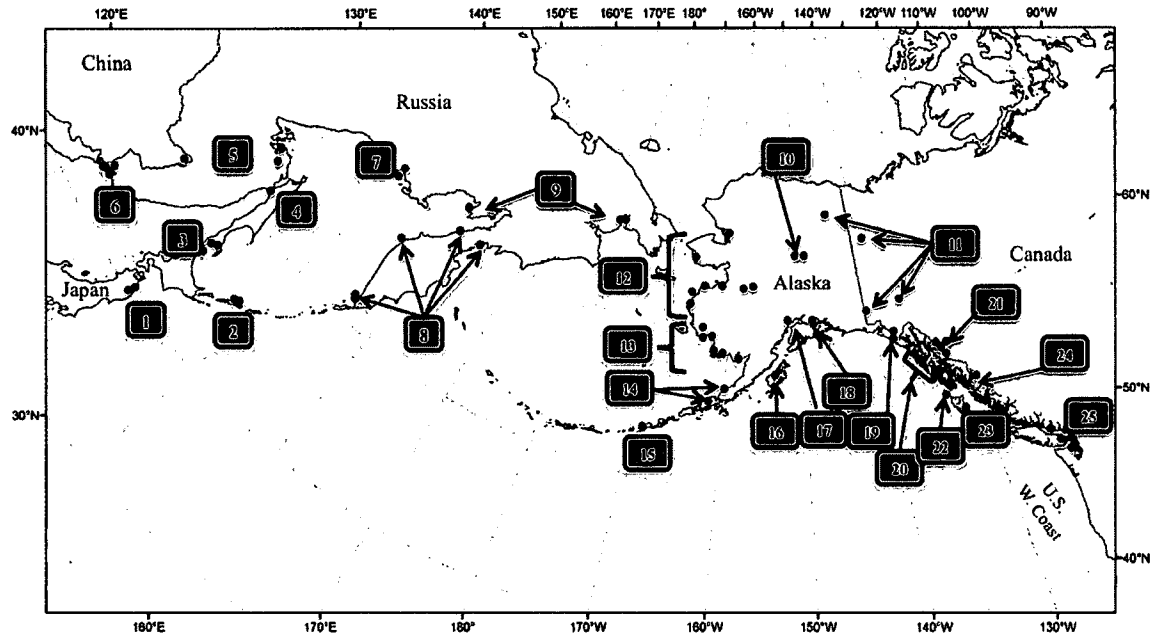


Figure 2.1. Area of study. Geographical sampling locations (dots) of the 74 stocks used for this study and the 25 regions (numbers) used for MSA. For group nine, two stocks that were geographically distant were genetically similar and were clustered as a single group.

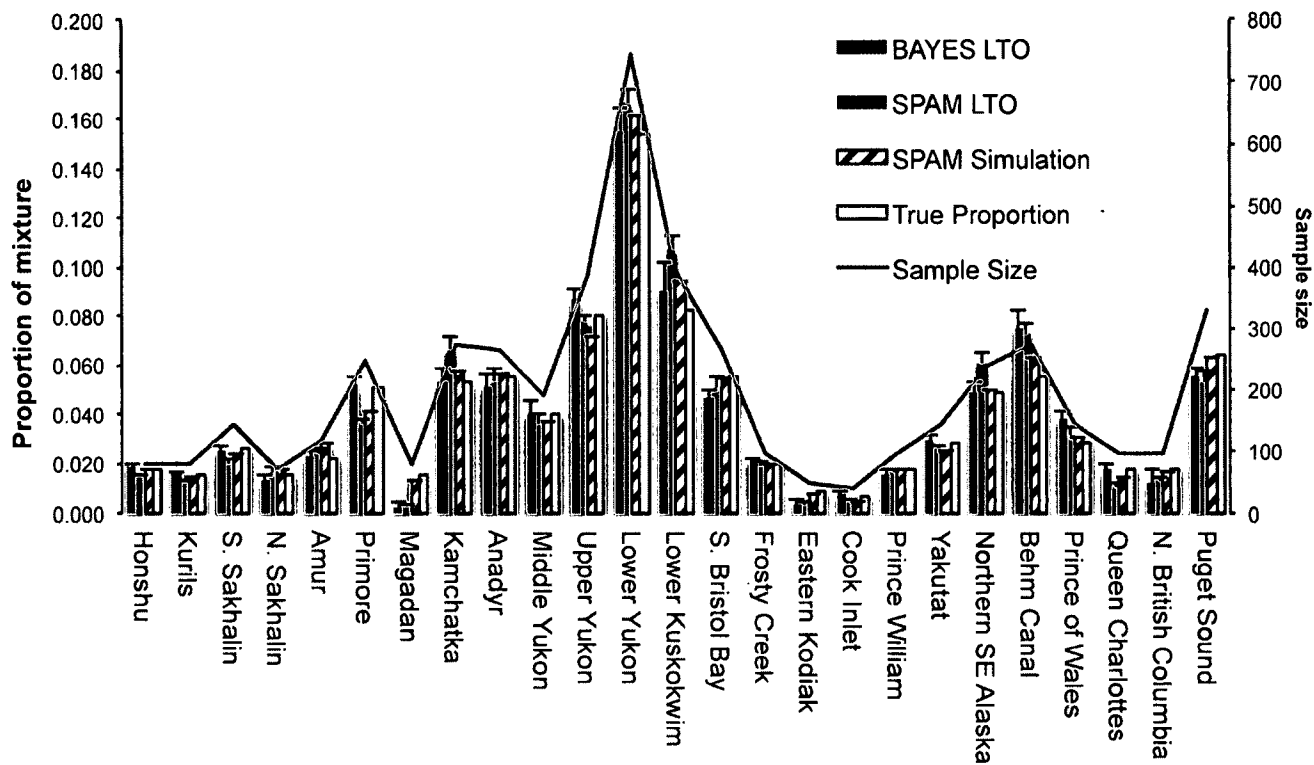


Figure 2.2a. Mean stock composition estimates for 25 reporting groups. Values were estimated by both BAYES and SPAM, and their standard errors (projecting whiskers above the bars) were calculated from the ten test datasets created from the combined microsatellite and SNP baseline. Filled bars indicate estimated reporting group proportions; and the unfilled bar for each region shows the true proportion of the mixtures created. The black bar is the average estimate of the ten test datasets computed with BAYES; the gray bar is the average estimate of the ten test datasets computed with SPAM; and the hatched bar is the average estimate computed from ten test datasets of SPAM Simulations. The gray continuous line is associated with the secondary y-axis and gives the sample size for each group.

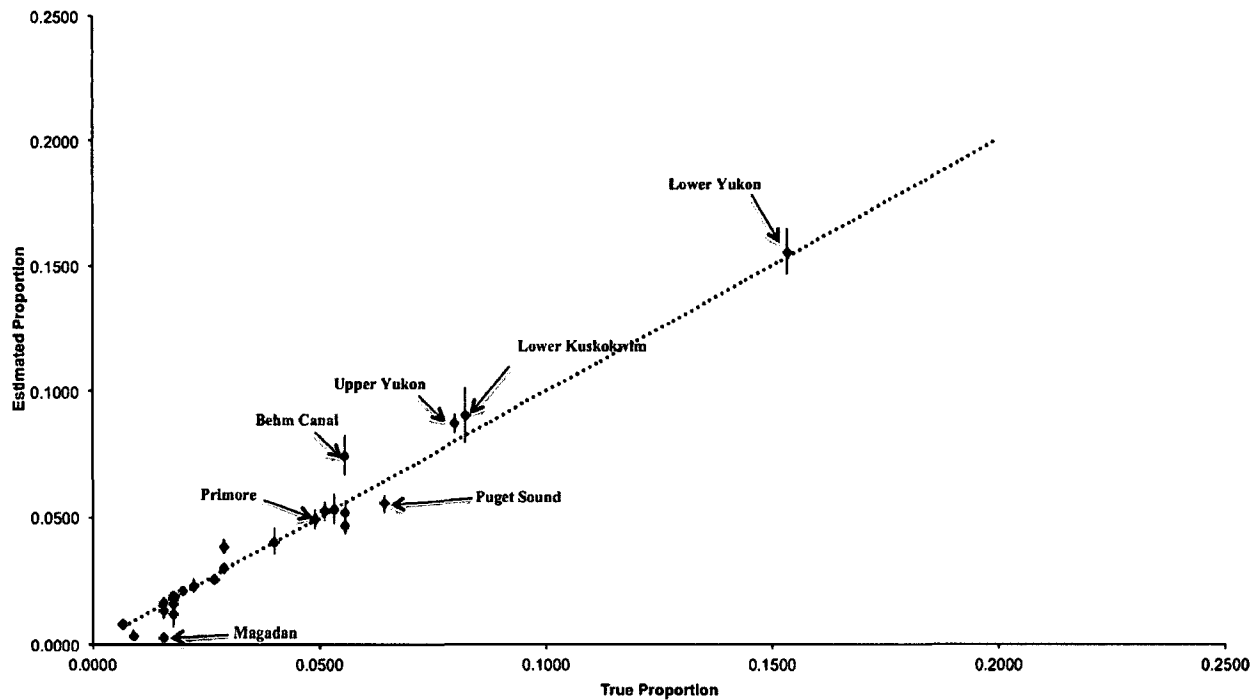


Figure 2.2b. BAYES LTO for 25 reporting groups. Comparison of the mean stock composition estimates of mixtures for 25 reporting groups versus the true composition of the mixture with BAYES LTO. The black diagonal line represents the relationship between a perfectly accurate estimate and the true value. Each filled circle represents the average proportion for one of the 25 reporting groups and the standard error of the estimated proportions is indicated by the whiskers for each circle. Names of groups whose averages included high bias for at least one of the ten mixture samples are indicated with arrows.

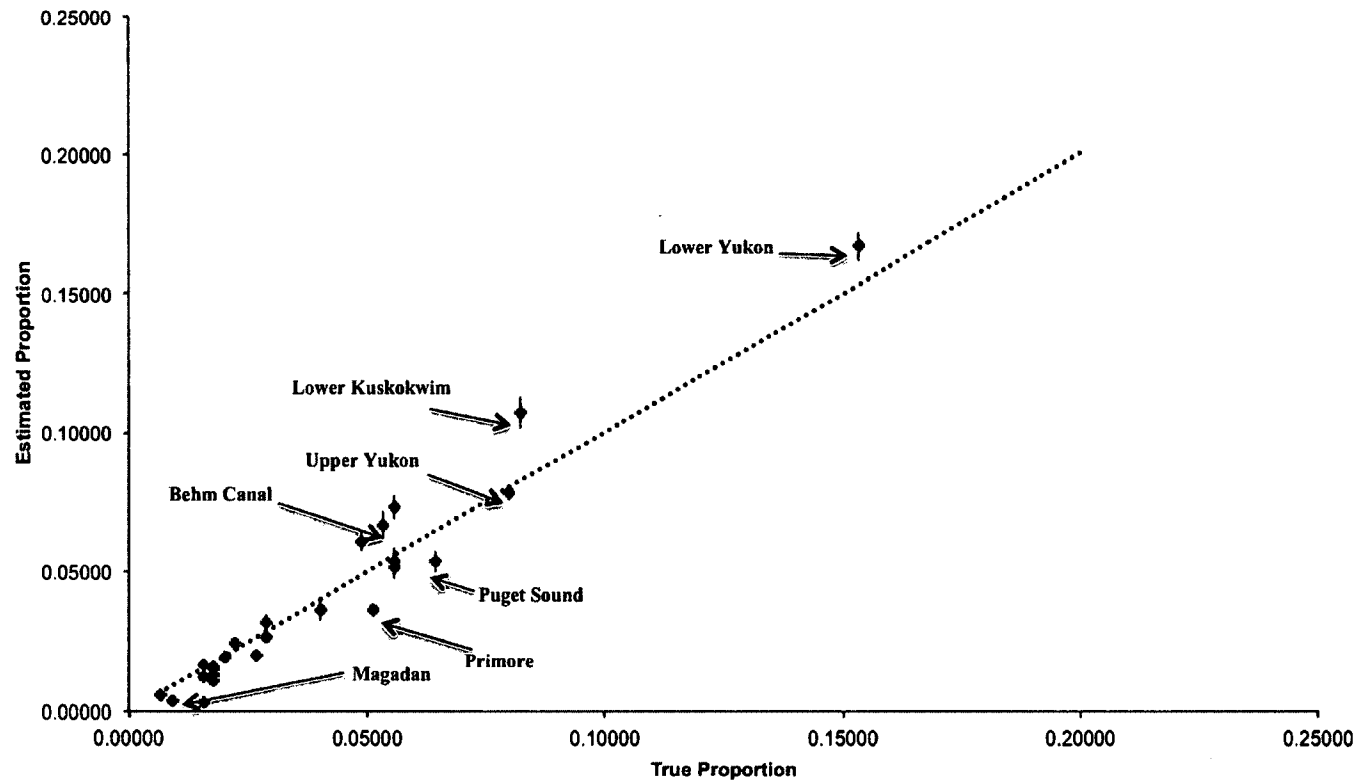


Figure 2.2c. SPAM LTO for 25 reporting groups. Comparison of the mean stock composition estimates of mixtures for 25 reporting groups versus the true composition of the mixture with SPAM LTO. The black diagonal line represents the relationship between a perfectly accurate estimate and the true value. Each filled circle represents the average proportion for one of the 25 reporting groups and the standard error of the estimated proportions is indicated by the whiskers (which are very short) for each circle. Names of groups whose averages included high bias for at least one of the ten mixture samples are indicated with arrows.

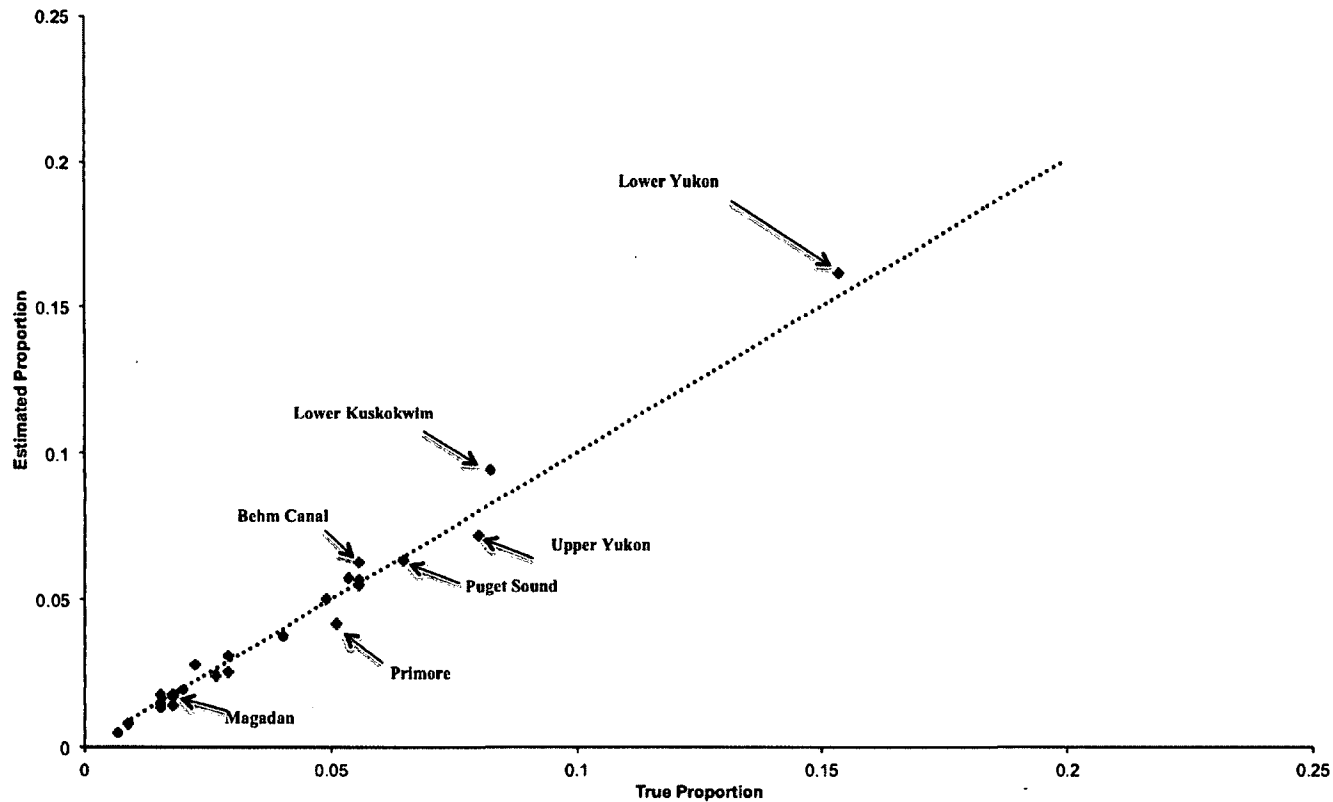


Figure 2.2d. SPAM simulations for 25 reporting groups. Comparison of the mean stock composition estimates of mixtures for 25 reporting groups versus the true composition of the mixture with SPAM Simulations. The black diagonal line represents the relationship between a perfectly accurate estimate and the true value. Each filled circle represents the average proportion for one of the 25 reporting groups and the standard error of the estimated proportions is indicated by the whiskers for each circle. Names of groups whose averages included high bias for at least one of the ten mixture samples are indicated with arrows.

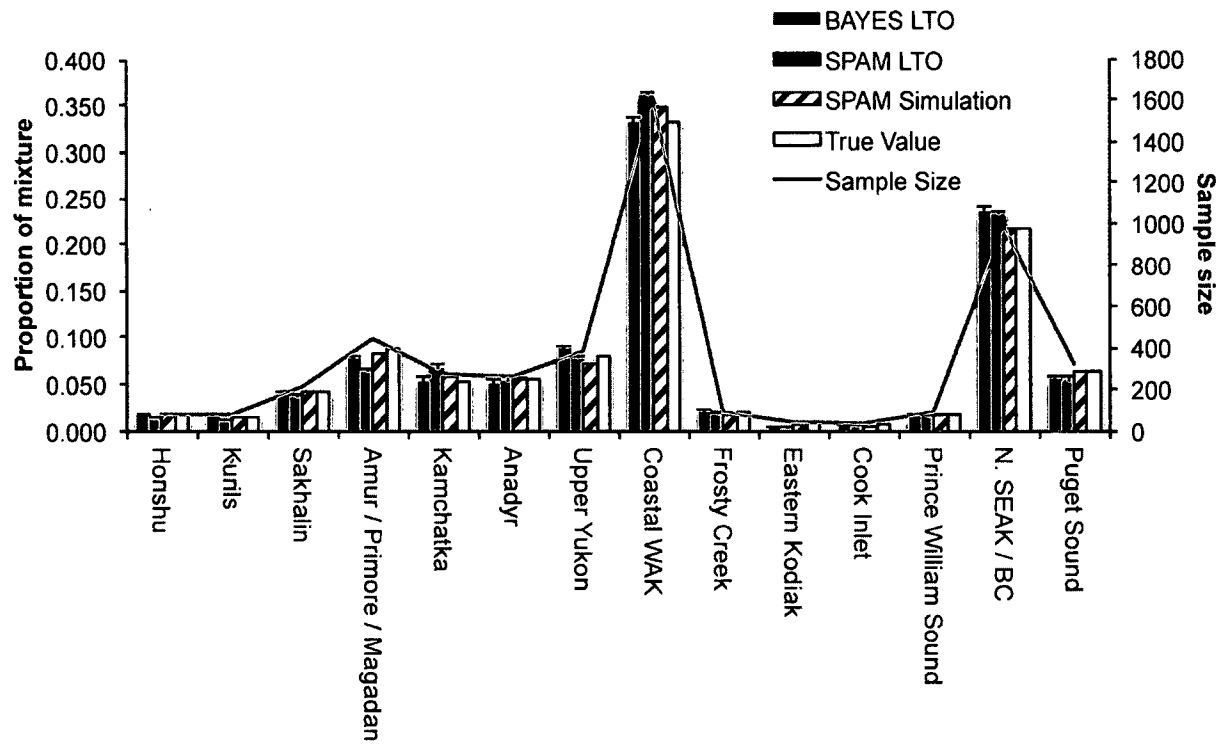


Figure 2.3a. Mean stock mixture estimates for 14 reporting groups. Filled bars indicate estimated proportions in the mixture and the unfilled bar in each set shows the true group proportion, of the mixtures created; the black bar represents the average estimate of the ten test datasets computed with BAYES; the gray bar is the average estimate of the ten test datasets computed with SPAM; and the gray bar with black outline is the average estimate computed from ten test datasets of SPAM simulations. Whiskers indicate standard errors. The gray continuous line is associated with the secondary y-axis and gives the sample size for each group.

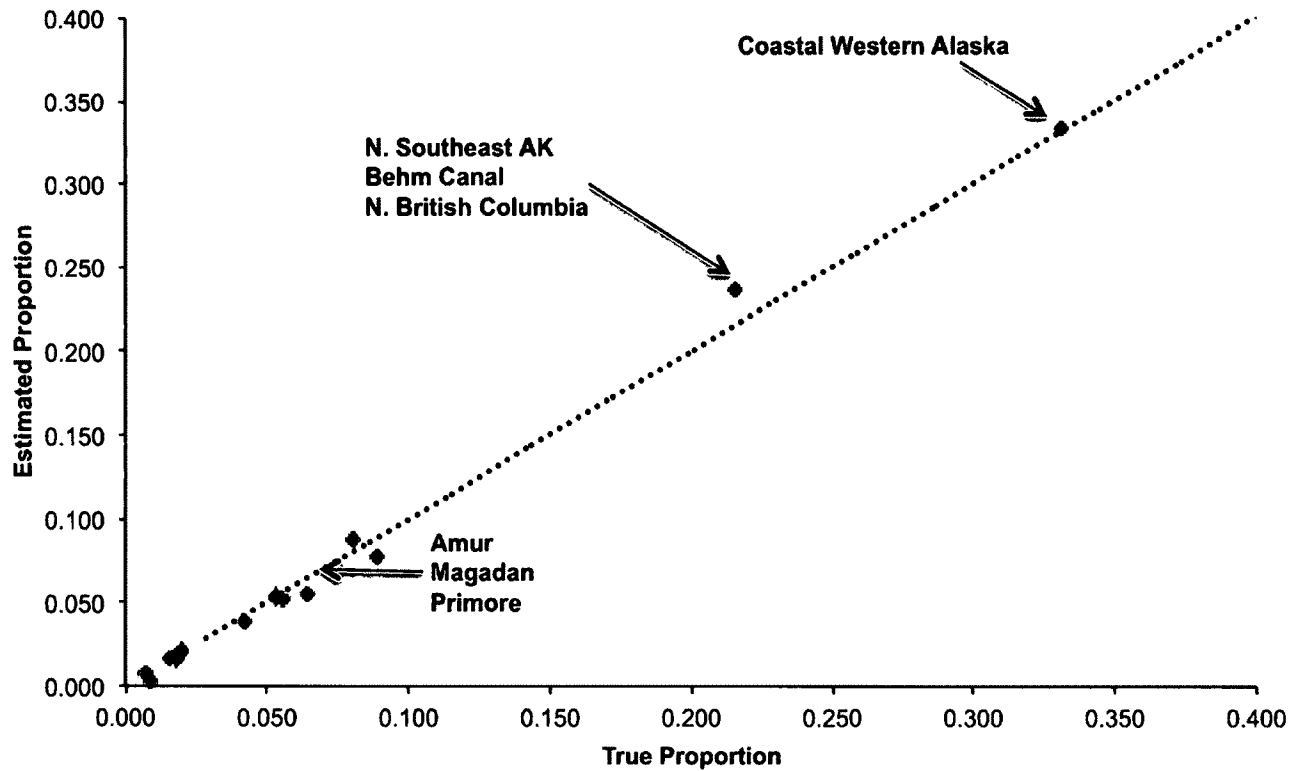


Figure 2.3b. BAYES LTO for 14 reporting groups. The mean stock composition estimates of mixtures for 14 reporting groups versus the true composition of the mixture with BAYES LTO. The black diagonal line represents the relationship between a perfectly accurate estimate and the true value. Each filled circle represents the average proportion for one of the 14 reporting groups and the standard error of the estimated proportions is indicated by the whiskers for each circle. Names of groups whose averages included high bias for at least one of the ten mixture samples are indicated with arrows.

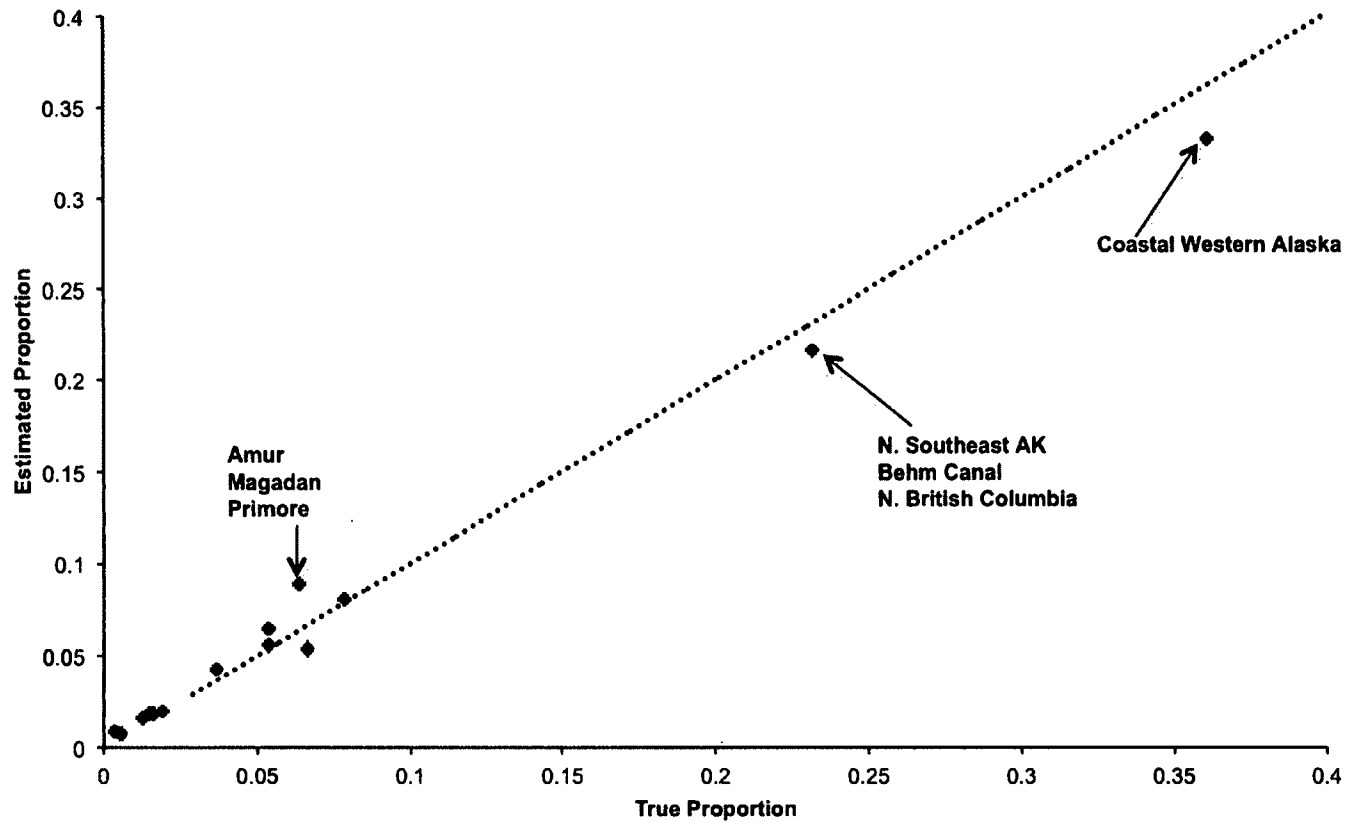


Figure 2.3c. SPAM LTO for 14 reporting groups. The mean stock composition estimates of mixtures for 14 reporting groups versus the true composition of the mixture with SPAM LTO. The black diagonal line represents the relationship between a perfectly accurate estimate and the true value. Each filled circle represents the average proportion for one of the 14 reporting groups and the standard error of the estimated proportions is indicated by the whiskers for each circle. Names of groups whose averages included high bias for at least one of the ten mixture samples are indicated with arrows.

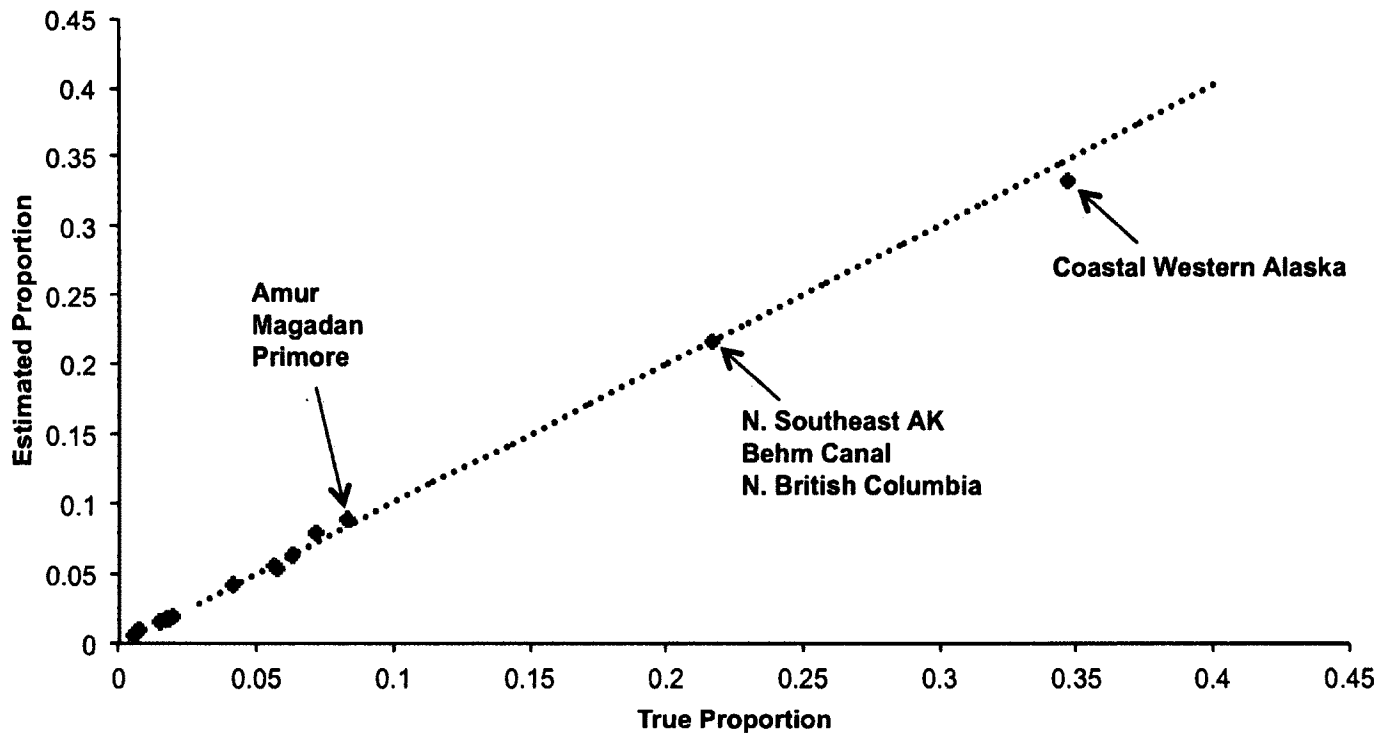


Figure 2.3d. SPAM simulations for 14 reporting groups. The mean stock composition estimates of mixtures for 14 reporting groups versus the true composition of the mixture with SPAM Simulations. The black diagonal line represents the relationship between a perfectly accurate estimate and the true value. Each filled circle represents the average proportion for one of the 14 reporting groups and the standard error of the estimated proportions is indicated by the whiskers (which are very short) for each circle. Names of groups whose averages included high bias for at least one of the ten mixture samples are indicated with arrows.

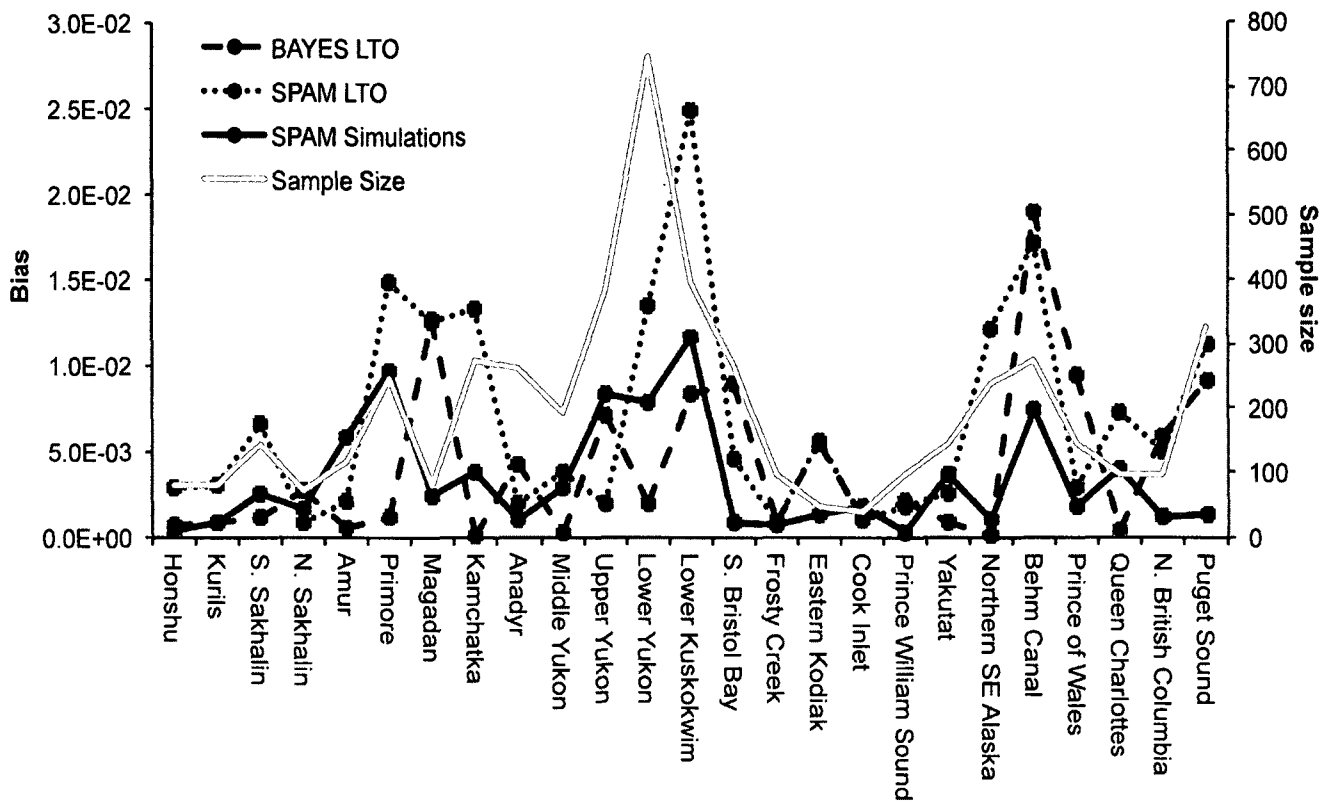


Figure 2.4. Bias in stock proportion estimates. Bias was calculated as absolute difference between the mean and the true value for each of the 25 reporting groups. Values are given for the BAYES LTO and the SPAM LTO as well as the SPAM simulation. The gray continuous line is associated with the secondary y-axis and gives the sample size for each group.

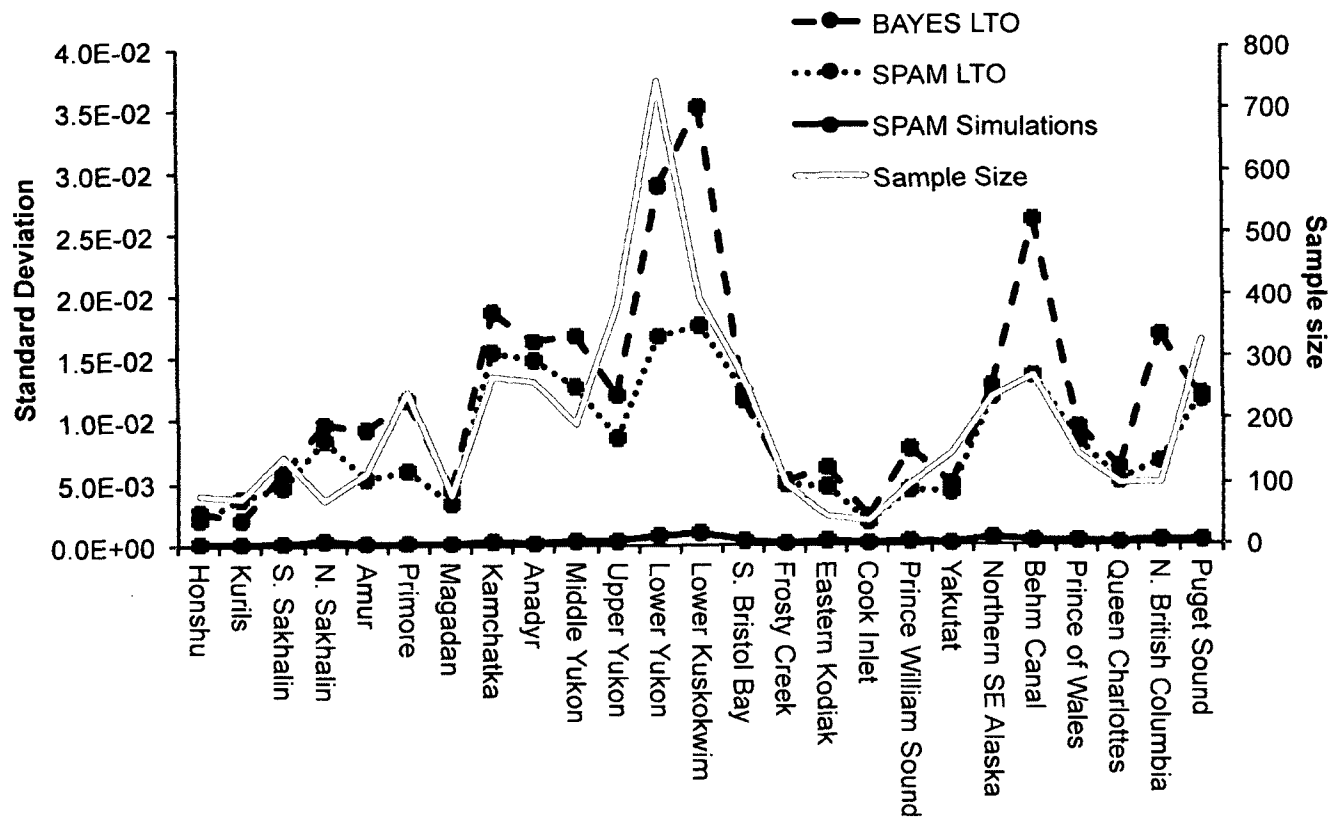


Figure 2.5. Standard deviation of stock proportion estimates. Values are given for the BAYES LTO and the SPAM LTO as well as the SPAM simulation. The gray continuous line is associated with the secondary y-axis and gives the sample size for each group.

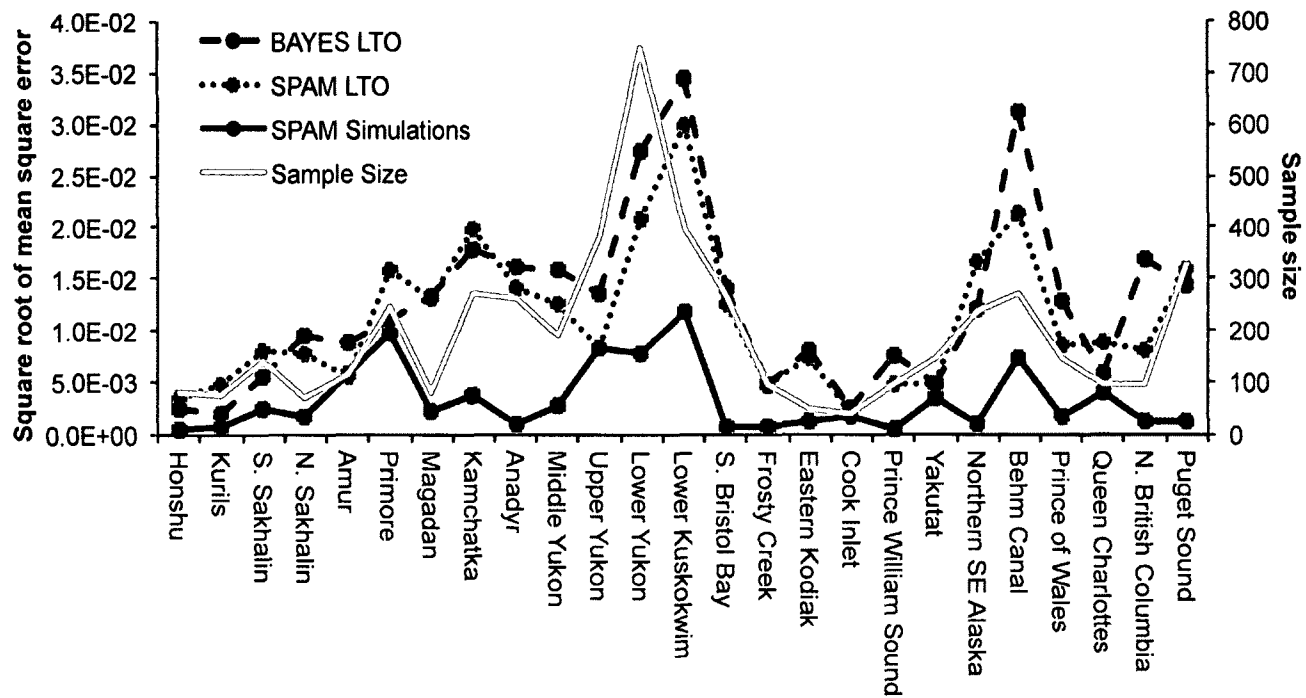


Figure 2.6. Mean squared error^{1/2} of the stock proportion estimates. Values are given for the BAYES LTO and the SPAM LTO as well as the SPAM simulation for each of the 25 reporting groups. The gray continuous line is associated with the secondary y-axis and gives the sample size for each group.

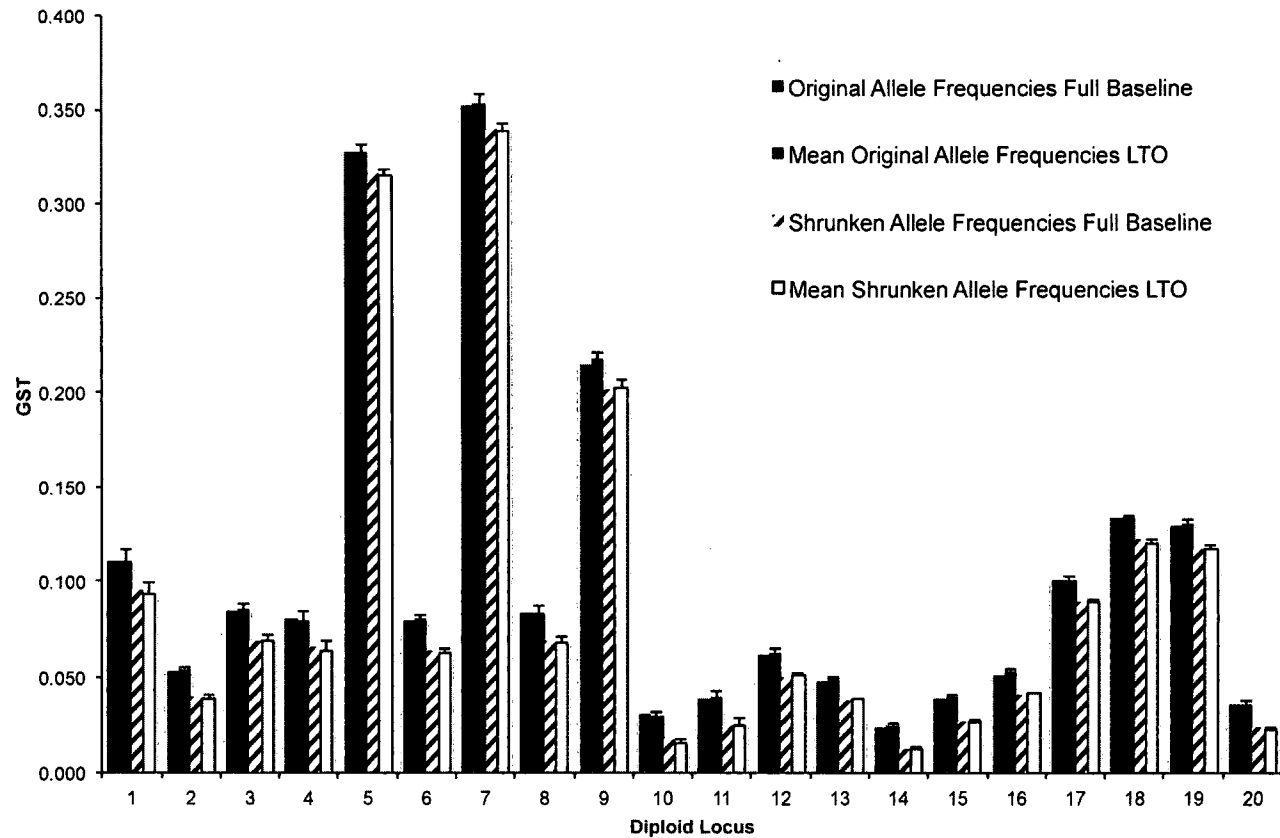


Figure 2.7. G_{ST} values to assess possible introduction of bias. G_{ST} values were calculated with the observed allele frequencies from the original full baseline and then for each of the 10 LTO baselines. The calculations were repeated with allele frequencies that were shrunk toward their prior grand mean according to Eq. 4 in Pella and Masuda (2001).

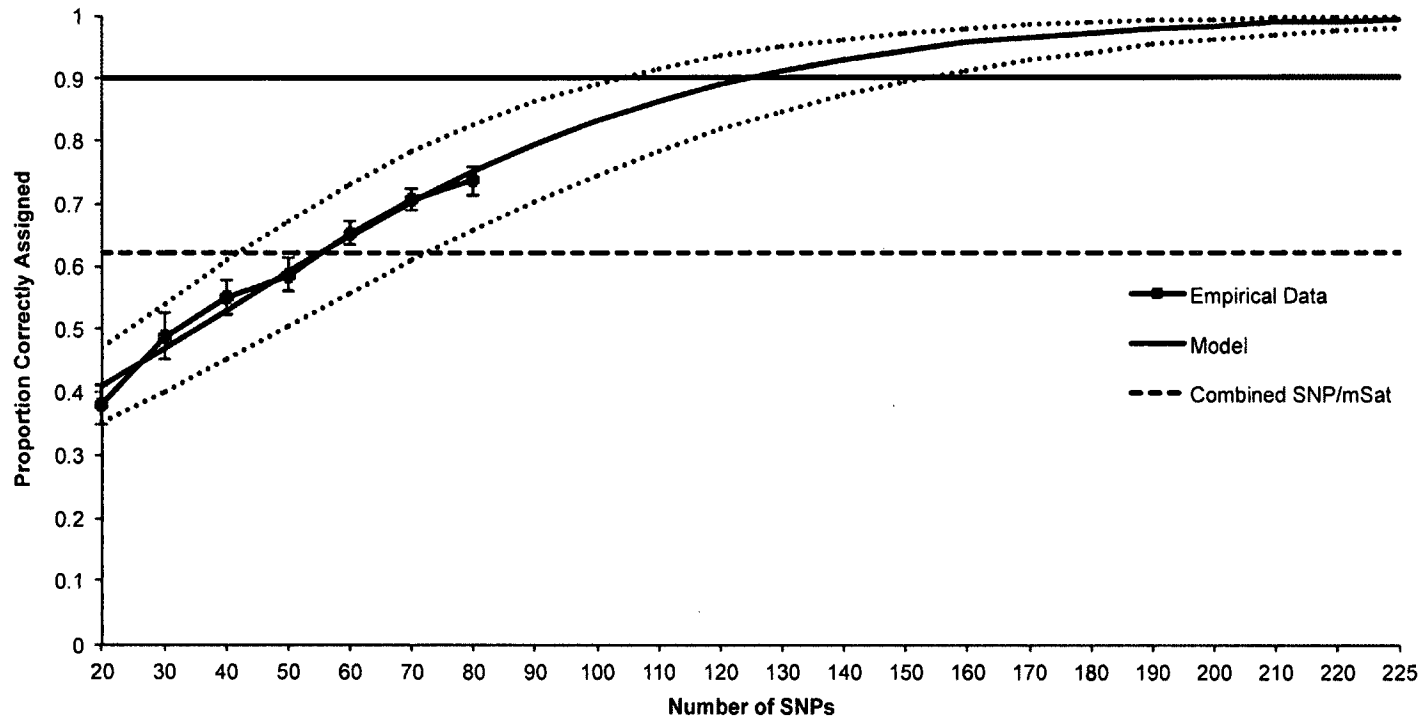


Figure 2.8. Mean proportion of individuals in the mixture correctly assigned. Assignment was to their reporting group (25 reporting groups) of origin with the MAP rule as related to increasing numbers of simulated SNP loci. Empirical data are shown with the solid line and black squares. The gray line represents a logistic function fit to the data with maximum likelihood and the dotted gray lines are the 2.5% and 97.5% confidence intervals. The solid black horizontal line represents the 0.9 proportion correctly assigned and the dotted black horizontal line is the proportion correctly assigned with the combined SNP, microsatellite, and mitochondrial DNA data (21 loci). Each extended set of loci includes the original 11 SNPs and mitochondrial DNA together with the simulated SNP loci.

Table 2.1

Stocks used in the study and their sources³. The 14- and 25- regional groupings and the baseline sample sizes are provided.

| Stock # | Stock Name | Date | Lat | Long | 25 Group # | Group Name | 14 Group # | Sample Size | Source |
|---------|----------------|-----------|-------|---------|------------|--------------|------------|-------------|---------|
| 1 | Tsugarishi | 1991 | 39.20 | 141.80 | 1 | Honshu | 1 | 40 | KITU |
| 2 | Katagishi | 1991 | 39.60 | 142.00 | | | | 40 | KITU |
| 3 | Reidovaya | 2006 | 45.38 | 147.98 | 2 | Kurils | 2 | 30 | LGG |
| 4 | Sopochnoe Lake | 2004 | 45.32 | 148.41 | | | | 48 | LGG |
| 5 | Naiba | 1995/1996 | 47.45 | 142.76 | 3 | S. Sakhalin | 3 | 47 | RAS |
| 6 | Okhotsk | 2003 | 46.87 | 143.17 | | | | 23 | RAS |
| 7 | Taranai | 2003 | 46.63 | 142.43 | | | | 25 | LGG |
| 8 | Udararnitsa | 1994 | 46.80 | 143.30 | | | | 48 | RAS |
| 9 | Tym | 1995/2003 | 51.26 | 142.71 | 4 | N. Sakhalin | | 71 | RAS/ABL |
| 10 | Heilong | 1994 | 48.38 | 134.38 | 5 | Amur | 4 | 44 | X.Luan |
| 11 | Amur Early | 2003 | 52.93 | 141.17 | | | | 45 | RAS |
| 12 | Amur Late | 2003 | 52.93 | 141.17 | | | | 27 | RAS |
| 13 | Anuyi | 2002 | 49.32 | 136.47 | 6 | Primore | | 46 | RAS |
| 14 | Barabashevka | 1994/1995 | 43.11 | 131.64 | | | | 50 | RAS |
| 15 | Narva | 1995/2005 | 42.99 | 131.49 | | | | 67 | RAS |
| 16 | Ryazakanovka | 1994/1995 | 43.16 | 132.11 | | | | 63 | RAS |
| 17 | Suifen | 1994 | 43.34 | 131.82 | | | | 20 | RAS |
| 18 | Ola | 1999 | 59.60 | 151.27 | 7 | Magadan | | 43 | RAS |
| 19 | Taui | 1999 | 59.39 | 149.14 | | | | 37 | RAS |
| 20 | Hailula | 2003 | 58.20 | 162.03 | 8 | Kamchatka | 5 | 47 | RAS |
| 21 | Ossoro | 1996 | 59.18 | 163.15 | | | | 48 | TNRO |
| 22 | Hairsova | 1990/1993 | 57.09 | 156.52 | | | | 96 | RAS |
| 23 | Kol | 2003 | 53.81 | 155.94 | | | | 47 | RAS |
| 24 | Utka | 2002 | 53.15 | 156.08 | | | | 35 | RAS |
| 25 | Oclan | 1993 | 62.77 | 164.33 | 9 | Anadyr | 6 | 73 | RAS |
| 26 | Anadyr | 1991 | 64.90 | 176.22 | | | | 111 | RAS |
| 27 | Kanchalon | 1991 | 65.12 | 176.53 | | | | 79 | ABL |
| 33 | Saicha | 1994 | 64.47 | -146.98 | 10 | Middle Yukon | 7 | 96 | ADF&G |
| 34 | Toklat | 1994 | 64.45 | -150.31 | | | | 96 | ADF&G |
| 28 | FishBranch | 1992 | 66.45 | -138.58 | 11 | Upper Yukon | 8 | 96 | ADF&G |
| 29 | Kluane | 1992 | 61.88 | -139.72 | | | | 96 | USFWS |
| 30 | Sheenjek | 1988/1989 | 66.74 | -144.57 | | | | 96 | ADF&G |
| 31 | Teslin | 1992 | 61.57 | -134.90 | | | | 96 | USFWS |
| 32 | Kobuk | 2000 | 66.92 | -160.81 | 12 | Lower Yukon | 7 | 96 | ADF&G |
| 35 | Agiapuk | | 65.17 | -165.68 | | | | 96 | DFO |
| 36 | Pilgrim | 2004 | 65.16 | -165.22 | | | | 96 | KWRK |
| 37 | Snake | 2004 | 64.50 | -165.41 | | | | 96 | KWRK |

Table 2.1 continued

| Stock # | Stock Name | Date | Lat | Long | 25 Group # | Group Name | 14 Group # | Sample Size | Source |
|---------|---------------|----------------|-------|---------|------------|------------------------|------------|-------------|--------|
| 38 | Pikmitalik | 2004 | 63.27 | -162.60 | | | 7 | 96 | KWRK |
| 39 | Atchulingak | 1989 | 61.96 | -162.83 | | | | 96 | USFWS |
| 40 | Anvik | 1989 | 62.68 | -160.20 | | | | 75 | USFWS |
| 41 | Kaltag | 1992 | 64.33 | -158.72 | | | | 48 | USFWS |
| 42 | Nulato | 2003 | 64.71 | -158.14 | | | | 48 | USFWS |
| 43 | Kanektok | 1989 | 59.75 | -161.93 | 13 | Lower Kuskokwim | | 75 | ADF&G |
| 44 | Kasigluk | 1990 | 60.85 | -161.23 | | | | 73 | ADF&G |
| 45 | Kwethluk | 1989 | 60.81 | -161.45 | | | | 77 | ADF&G |
| 46 | Goodnews | 1989 | 59.13 | -161.48 | | | | 96 | ADF&G |
| 47 | Nushagak | 1988 | 58.80 | -158.63 | | | | 75 | ADF&G |
| 48 | Bigcreek | 1988/2000 | 58.29 | -157.53 | 14 | S. Bristol Bay | | 96 | ADF&G |
| 49 | Gertrude | 1987/1999 | 58.17 | -156.21 | | | | 96 | ADF&G |
| 50 | Meshik | 1989 | 56.81 | -158.66 | | | | 75 | ADF&G |
| 51 | Frosty | 2000 | 55.07 | -162.81 | 15 | Frosty | 9 | 96 | ADF&G |
| 52 | Kizyuak | 1989 | 57.82 | -152.80 | 16 | Kodiak | 10 | 48 | ADF&G |
| 53 | LittleSu | 1990 | 61.25 | -150.29 | 17 | Cook Inlet | 11 | 39 | ABL |
| 54 | Olsen | 1992/1997 | 60.76 | -146.17 | 18 | Prince William Sour | 12 | 96 | ABL |
| 55 | Alsek | 2000 | 59.13 | -138.62 | 19 | Yakutat | 13 | 96 | ABL |
| 56 | EAlsek | 2006 | 59.11 | -138.52 | | | | 48 | UAFSOS |
| 57 | Green's Creek | 1995 | 58.10 | -134.76 | 20 | Northern SE Alaska | | 96 | ABL |
| 58 | Herman Creek | 1987/1990/2008 | 59.42 | -136.10 | | | | 96 | ABL |
| 59 | Taku | 2000 | 58.43 | -133.98 | | | | 45 | ABL |
| 60 | Blossom | 1986 | 55.40 | -130.61 | 21 | Behm Canal | | 48 | ABL |
| 61 | Marten | 1986 | 55.16 | -130.53 | | | | 48 | ABL |
| 62 | Portage Creek | 1986/1988 | 55.77 | -131.04 | | | | 96 | ABL |
| 63 | Wilson | 1986 | 55.40 | -130.61 | | | | 40 | ABL |
| 64 | Herman River | 1986 | 55.99 | -131.27 | | | | 40 | ABL |
| 65 | Karta | 1986 | 55.56 | -132.57 | 22 | Prince of Wales Island | | 48 | ABL |
| 66 | Old Tom Creek | 1986/1988 | 55.40 | -132.40 | | | | 96 | ABL |
| 67 | Bag Harbor | 1989 | 52.35 | -131.36 | 23 | QCI | | 48 | ABL |
| 68 | Tasu | 1989 | 52.87 | -132.08 | | | | 48 | ABL |
| 69 | Klownick | 1989 | 52.38 | -126.75 | 24 | N. British Columbia | | 48 | ABL |
| 70 | Neekas | 1989 | 52.47 | -128.17 | | | | 48 | ABL |
| 71 | Grant | 1998 | 48.27 | -122.02 | 25 | Puget Sound | 14 | 96 | WDFG |
| 72 | Kennedy | 1996 | 47.10 | -123.09 | | | | 96 | WDFG |
| 73 | Johns | 2003 | 47.24 | -123.04 | | | | 96 | WDFG |
| 74 | Quilcene | 1997 | 47.80 | -122.86 | | | | 40 | ABL |

³ ADF&G – Alaska Department of Fish & Game, DFO – Department of Fisheries Oceans, Canada, KWRK – Kawerek, LGG – Laboratory of Genetic Identification, Institute of General Genetics, NMFSABL – National Marine Fisheries Service Auke Bay Labs, TNRO – Kamchatka TINRO, UAF – University of Alaska Fairbanks, USFWS – U.S. Fish and Wildlife Service, WDF&W – Washington Department of Fish and Wildlife.

Table 2.2

Measures of genetic diversity. Values are given for the 12 SNP and nine microsatellite loci analyzed for all individuals in the 74 chum salmon stocks in this work. F_{ST} is Weir and Cockerham's θ (Weir and Cockerham 1984), D_{EST} is Jost's D (Jost 2008), and H_e is the expected heterozygosity.

| Locus | # SNPs | Type | F_{ST} | D_{EST} | H_e |
|------------|--------|------|----------|-----------|-------|
| VT | 1 | SNP | 0.086 | 0.101 | 0.491 |
| IN | 2 | SNP | 0.053 | 0.015 | 0.228 |
| SP | 1 | SNP | 0.076 | 0.074 | 0.484 |
| RH | 1 | SNP | 0.081 | 0.012 | 0.129 |
| VR | 3 | SNP | 0.156 | 0.308 | 0.730 |
| IS | 2 | SNP | 0.081 | 0.120 | 0.622 |
| ER | 1 | SNP | 0.388 | 0.225 | 0.435 |
| PL | 1 | SNP | 0.096 | 0.034 | 0.276 |
| RF | 1 | SNP | 0.218 | 0.071 | 0.303 |
| CL | 1 | SNP | 0.034 | 0.020 | 0.408 |
| PER | 1 | SNP | 0.040 | 0.004 | 0.082 |
| MT | 8 | SNP | 0.345 | N/A | N/A |
| One104 | N/A | mSat | 0.027 | 0.346 | 0.951 |
| One102 | N/A | mSat | 0.011 | 0.127 | 0.922 |
| Ots68 | N/A | mSat | 0.019 | 0.298 | 0.956 |
| Ssa419 | N/A | mSat | 0.028 | 0.157 | 0.872 |
| One114 | N/A | mSat | 0.017 | 0.178 | 0.933 |
| Omy1011 | N/A | mSat | 0.026 | 0.272 | 0.937 |
| One101 | N/A | mSat | 0.060 | 0.384 | 0.908 |
| Oki100 | N/A | mSat | 0.044 | 0.313 | 0.905 |
| Ots103 | N/A | mSat | 0.022 | 0.378 | 0.965 |
| mSats Only | | | 0.028 | 0.273 | 0.928 |
| SNPs only | | | 0.157 | 0.089 | 0.399 |
| Overall | | | 0.073 | 0.172 | 0.627 |

Table 2.3

Sum of mean error^{1/2} for three methods; BAYES LTO, SPAM LTO and SPAM simulation.

| Regions | BAYES LTO | SPAM LTO | SPAM Sim |
|----------------|------------------|-----------------|-----------------|
| 25 | 0.314 | 0.279 | 0.085 |
| 14 | 0.157 | 0.174 | 0.044 |

References

- Anderson E. 2010. Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Mol. Ecol.* 10(4): 701-710.
- Anderson E., Waples R.S., and Kalinowski S.T. 2008. An improved method for estimating the accuracy of genetic stock identification. *Can. J. Fish. Aquat. Sci.* 65: 1475-1486.
- Barbee N.C., and Swearer S.E. 2007. Characterizing natal source population signatures in the diadromous fish *Galaxias maculatus*, using embryonic otolith chemistry. *Mar. Ecol. Prog. Ser.* 343: 273-282.
- Barnosky A.D., Matzke N., Tomiya S., *et al.* 2011. Has the Earth's sixth mass extinction already arrived? *Nature* 471(7336): 51-57.
- Beacham T.D., Candy J.R., Le K.D., and Wetklo M. 2009. Population structure of chum salmon (*Oncorhynchus keta*) across the Pacific Rim, determined from microsatellite analysis. *Fish. Bull.* 107: 244-260.
- Beacham T.D., McIntosh B., and Wallace C.G. 2011. A comparison of polymorphism of genetic markers and population sample sizes required for mixed-stock analysis of sockeye salmon (*Oncorhynchus nerka*) in British Columbia. *Can. J. Fish. Aquat. Sci.* 68: 550-562.
- Billington N. 2003. Mitochondrial DNA. In *Population Genetics: Principles and Applications for Fisheries Scientists*. Edited by Hallerman E.M. American Fisheries Society, Bethesda.

- Bowen B.W., Grant W.S., Hillis-Starr Z., *et al.* 2007. Mixed-stock analysis reveals the migrations of juvenile hawksbill turtles (*Eretmochelys imbricata*) in the Caribbean Sea. *Mol. Ecol.* 16(1): 49-60.
- Boyce M.S., Vernier P.R., Nielsen S.E., and Schmiegelow F.K.A. 2002. Evaluating resource selection functions. *Ecol. Model.* 157: 281-300.
- Brandt A.L., Ishida Y., Georgiadis N.J., and Roca A.L. 2012. Forest elephant mitochondrial genomes reveal that elephantid diversification in Africa tracked climate transitions. *Mol. Ecol.* 21(5): 1175-1189.
- Campbell N.R., Amish S.J., Pritchard V.L., *et al.* 2012. Development and evaluation of 200 novel SNP assays for population genetic studies of westslope cutthroat trout and genetic identification of related taxa. *Mol. Ecol. Resour.* 12(5): 942-949.
- Campbell N.R., and Narum S.R. 2008. Identification of novel single-nucleotide polymorphisms in chinook salmon and variation among life history types. *Trans. Amer. Fish. Soc.* 137: 96-106.
- Cochran W. 1963. *Sampling techniques*. Wiley and Sons, New York.
- Debevec E., Gates R., Masuda M., *et al.* 2000. SPAM (version 3.2): Statistics Program for Analyzing Mixtures. *J. Hered.* 91: 509-510.
- Dempster A.P., Laird N.M., and Rubin D.B. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. *J. R. Stat. Soc.: Series B (Methodology)* 39: 1-38.

- Elfstrom C., Smith C., and Seeb J. 2006. Thirty-two single nucleotide polymorphism markers for high-throughput genotyping of sockeye salmon. *Mol. Ecol. Notes* 6: 1255-1259.
- Excoffier L., Smouse P.E., and Quattro J.M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131: 479-491.
- Foote A.D., Morin P.A., Durban J.W., *et al.* 2010. Positive selection on the killer whale mitogenome. *Biol. Lett.* 7(1): 116-118.
- Fournier D.A., Beacham T.D., Riddell B.E., and Busack C.A. 1984. Estimating stock composition in mixed stock fisheries using morphometric, meristic, and electrophoretic characteristics. *Can. J. Fish. Aquat. Sci.* 41(3): 400-408.
- Garvin M.R., Bielwaski J.P., and Gharrett A.J. 2011. Positive Darwinian selection in the piston that powers proton pumps in Complex I of the mitochondria of Pacific Salmon. *PLoS One* 6(9): e24127.
- Garvin M.R., and Gharrett A.J. 2010. Application of SNP markers to chum salmon (*Oncorhynchus keta*): Discovery, genotyping, and linkage phase resolution. *J. Fish Biol.* 77(9): 2137-2162.
- Garvin M.R., and Gharrett A.J. 2007. DEco-TILLING: An inexpensive method for SNP discovery that reduces ascertainment bias. *Mol. Ecol. Notes* 7: 735-746.
- Garvin M.R., Saitoh K., Brykov V., Churikov D., and Gharrett A.J. 2010. Single nucleotide polymorphisms in chum salmon (*Oncorhynchus keta*) mitochondrial DNA derived from restriction site haplotype information. *Genome* 53: 501-507.

- Geisser S. 1975. The predictive sample reuse method with applications. *J. Amer. Stat. Assoc.* 70(350): 320-328.
- Gelman A., and Rubin D.B. 1992. Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7: 457-511.
- Gisclair B.R. 2009. Salmon bycatch management in the Bering Sea walleye pollock fishery: Threats and opportunities for western Alaska. *Am. Fish. Soc. Symp.* 70: 799-816.
- Gomez-Diaz E., and Gonzales-Solis J. 2007. Geographic assignment of seabirds to their origin: combining morphological, genetic, and biogeochemical analyses. *Ecol. Appl.* 17(5): 1484-1498.
- Griffiths A.M., Machado-Schiaffino G., Dillane E., *et al.* 2010. Genetic stock identification of Atlantic salmon (*Salmo salar*) populations in the southern part of the European range. *BMC Genet.* 11: 31.
- Grossman L.I., Wildman D.E., Schmidt T.R., and Goodman M. 2004. Accelerated evolution of the electron transport chain in anthropoid primates. *Trends Genet.* 20(11): 578-585.
- Habicht C., Seeb L.W., Myers K.W., Farley E.V., and Seeb J.E. 2010. Summer–fall distribution of stocks of immature sockeye salmon in the Bering Sea as revealed by single-nucleotide polymorphisms. *Trans. Amer. Fish. Soc.* 139(4): 1171-1191.
- Hastie T., Tibshirani R., and Friedman J. 2001. *Elements of statistical learning: Data mining, inference and prediction.* Springer-Verlag, New York.

- Hess K.R., Anderson K., Symmans W.F., *et al.* 2006. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J. Clin. Oncol.* 24(26): 4236-4244.
- Hosmer D.W., and Lemeshow S. 2000. *Applied Logistic Regression*. John Wiley & Sons, Inc.
- Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24: 1403-1405.
- Jost L. 2008. G_{ST} and its relatives do not measure differentiation. *Mol. Ecol.* 17(18): 4015-4026.
- Kalinowski S.T. 2002. How many alleles per locus should be used to estimate genetic distances? *Heredity* 88: 62-65.
- Kalinowski S.T. 2003. *Genetic Mixture Analysis 1*. Department of Ecology, Montana State University. http://www.montana.edu/kalinowski/GMA/GMA_Home.htm
- Larson-Cook K, Zon EV, Rai S, Rusch T 2011. Microplate replacement for HT screening : Technology miniaturizes reactions in highly automated inline platform. *Genet. Eng. Biotech. N.* 31: 40-41.
- Lewis P.O., and Zaykin D. 2001. *Genetic Data Analysis: Computer program for the analysis of allelic data*. Version 1.0 (d16c).
<http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php>

- Manel S., Gaggiotti O.E., and Waples R.S. 2005. Assignment methods: matching biological questions with appropriate techniques. *Trends Ecol. and Evol.* 30(3): 136-142.
- Moriya S., Shunpei S., Azumaya T., Suzuki O., and Urawa S. 2006. Genetic stock identification of chum salmon in the Bering sea and north Pacific ocean using mitochondrial DNA microarray. *Mar. Biotechnol.* 9: 179-191.
- Narum S.R., Banks M., Beacham T.D., *et al.* 2008. Differentiating salmon populations at broad and fine geographic scales with microsatellites and single nucleotide polymorphisms. *Mol. Ecol.* 17: 3464-3477.
- Negrini R., Nicoloso L., Crepaldi P., *et al.* 2008. Assessing SNP markers for assigning individuals to cattle populations. *Anim. Genet.* 40: 18-26.
- Nei M., and Chesser R.K. 1983. Estimation of fixation indices and gene diversities. *Annal. Hum. Genet.* 47: 253-259.
- Nolte A.W., and Sheets H.D. 2005. Shape based assignment tests suggest transgressive phenotypes in natural sculpin hybrids (Teleostei, Scorpaeniformes, Cottidae). *Front. Zool.* 2: 11.
- Olsen J.B., Crane P.A., Flannery B.G., *et al.* 2010. Comparative landscape genetic analysis of three Pacific salmon species from subarctic North America. *Conserv. Genet.* 12(1): 223-241.
- Pella J., and Masuda M. 2001. Bayesian methods for analysis of stock mixtures from genetic markers. *Fish. Bull.* 99: 151-167.

- Pella J., and Masuda M. 2005. Classical discriminant analysis, classification of individuals, and source population composition of mixtures. In *Stock Identification Methods: applications in fisheries science*. Edited by Cadrin S., Friedland K., Waldman J. Academic Press, New York. pp. 517-552.
- Pella J., and Milner G. 1987. Use of genetic markers in stock composition analysis. In *Population Genetics and Fishery Management* Edited by Ryman N., Utter F. Washington Sea Grant, Seattle and London. pp. 247-275.
- Pella J.J., and Geiger H.J. 2009. Sampling considerations for estimating geographic origins of Chinook salmon bycatch in the Bering Sea pollock fishery, Special Publication No. SP 09-08, pp. 1-58. Alaska Department of Fish and Game, Anchorage.
- Ragoussis J. 2009. Genotyping technologies for genetic research. *Ann. Rev. Hum. Genet.* 10: 117-133.
- Reich B.J., and Bondell H.D. 2011. A Spatial dirichlet process mixture model for clustering population genetics data. *Biometrics* 67(2): 381-390.
- Reich D., Price A.L., and Patterson N. 2008. Principal component analysis of genetic data. *Nat. Genet.* 40(5): 491-492.
- Santure A.W., Stapley J., Ball A.D., *et al.* 2010. On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Mol. Ecol.* 19(7): 1439-1451.

- Scott G.R., Schulte P.M., Egginton S., *et al.* 2010. Molecular evolution of cytochrome *c* oxidase underlies high-altitude adaptation in the bar-headed goose. *Mol. Biol. Evol.* 28(1): 351-363.
- Seeb L.W., and Crane P.A. 1999. Genetic heterogeneity in chum salmon in western Alaska, the contact zone between northern and southern lineages. *Trans. Amer. Fish. Soc.* 128: 58-87.
- Seeb L.W., Templin W.D., Sato S., *et al.* 2011. Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. *Mol. Ecol. Resour.* 11: 195-217.
- Slonim D.K. 2002. From patterns to pathways: gene expression data analysis comes of age. *Nat. Genet.* 32: 502-508.
- Smith C.T., Elfstrom C.M., Seeb L.W., and Seeb J.E. 2005. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. *Mol. Ecol.* 14: 4193-4203.
- Templin W.D., Seeb J.E., Jasper J.R., Barclay A.W., and Seeb L.W. 2011. Genetic differentiation of Alaska Chinook salmon: the missing link for migratory studies. *Mol. Ecol. Resour.* 11: 226-246.
- Thompson S.K. 1992. *Sampling*. John Wiley & Sons, New York.
- Waples R.S. 2010. High-grading bias: subtle problems with assessing power of selected subsets of loci for population assignment. *Mol. Ecol.* 19: 2599-2601.

- Wasser S.K., Shedlock A.M., Comstock K., *et al.* 2004. Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. *Proc. Nat. Acad. Sci, USA* 101(41): 14847-14852.
- Weir B.S., and Cockerham C.C. 1984. Estimating F -statistics for the analysis of population structure. *Evolution* 38: 1358-1370.
- Wilmot R.L., Everett R.J., Spearman W.J., *et al.* 1994. Genetic stock structure of western Alaskan chum salmon and a comparison with Russian far east stocks. *Can. J. Fish. Aquat. Sci.* 51(Suppl. 1): 84-94.
- Wolfe R.J., and Spaeder J. 2009. People and salmon of the Yukon and Kuskokwim drainages and Norton Sound Alaska: Fishery harvests, culture change, and local knowledge systems. *Am. Fish. Soc. Symp.* 70: 349-379.

Chapter 3

Recent Physical Connections among Now-Divided Drainages May Explain Weak Genetic Structure in Western Alaskan Chum Salmon (*Oncorhynchus keta*) Populations⁴

⁴ Michael R. Garvin, Christine M. Kondzela, Bruce Finney, Patrick C. Martin, Jeffrey R. Guyon, William D. Templin, Nick DeCovich, Sara Gilk-Baumer, and Anthony J. Gharrett. *Molecular Ecology* (anticipated)

Abstract

Variability of western Alaskan chum salmon runs and some recent declines have prompted efforts to understand the causes, which for many populations, requires knowledge of the origins of samples caught at sea. However, regional assignments based on genotypic data are difficult because the genetic structure among populations from much of western Alaska is weak. The weak structure has been attributed to high levels of present-day gene flow among populations that may be thousands of kilometers apart. We used genotypes from microsatellite and single nucleotide polymorphism loci to investigate alternative explanations for the current genetic structure of chum salmon populations from western Alaska. We also estimated current levels of gene flow among Kuskokwim River populations. Our results suggest that the weak genetic structure is best explained by physical connections that occurred after the Holocene Maximum among the Yukon, Kuskokwim, and Nushagak drainages, which allowed gene flow to occur among now distant populations.

Introduction

Historical abundances of Pacific salmon that spawn in coastal western Alaskan drainages have been highly variable (Linderman & Bergstrom 2009; Wolfe & Spaeder 2009).

Abrupt, widespread declines have stimulated efforts to understand the factors underlying the reductions, even though some populations have recently rebounded. Potential causes for variable abundances include climate perturbations in both marine and fresh water ecosystems and incidental bycatch by the Bering Sea groundfish and western Alaskan salmon fisheries (Beamish & Bouillon 1993; Kruse 1998; Seeb et al. 2004; Gisclair 2009). Because several of the potential causes for declines include processes in the marine environment, it is essential that the origin of fish that are sampled at sea be ascertained.

The primary approach used to assign individuals to their population of origin is mixed stock analysis (MSA) (Fournier et al. 1984), which provides a probabilistic assessment of the origin of the samples by comparing their genotypes to a reference baseline of genotypes. Chum salmon (*Oncorhynchus keta*), in general, have lower levels of divergence compared to other Pacific salmon species such as Chinook (*O. tshawytscha*), sockeye (*O. nerka*), coho (*O. kisutch*), and steelhead (*O. mykiss*) (Quinn 2005; Olsen et al. 2010; Seeb et al. 2011; Templin et al. 2011). The summer-run chum salmon populations from the large geographic area that includes Kotzebue Sound, Norton Sound, the Lower Yukon River, the Kuskokwim River, and Bristol Bay (Figure 3.1a) show much lower divergence than populations from the remainder of the chum salmon range (Beacham et al. 2009a; Seeb et al. 2011), possibly because of their abundance and

recent colonization of western Alaska after the Last Glacial Maximum (LGM) (Wilmot et al. 1994). A subset of these populations that we will call ‘coastal southwestern Alaskan’ demonstrate genetic divergence that is 5-10 times lower than the rest of western Alaska and has created difficulties for analyses that rely on MSA (Beacham et al. 2009b; Seeb et al. 2011). These drainages span a geographic area of approximately 350 000 km², which includes southern Norton Sound, the Lower Yukon River, the Kuskokwim River, and northern Bristol Bay (roughly the size of Washington, Oregon, and Idaho combined) and each area has very different conservation and management goals (Wolfe & Spaeder 2009).

The weak genetic structure of these broadly-distributed populations could result from several causes. Balancing selection (convergent evolution) could produce the genetic similarity among populations, but is unlikely because the size of the geographic region is large and includes several different ecosystems (Olsen et al. 2010); so one would not expect allele frequencies to be maintained among populations that are located in different environmental regimes. Moreover, one would expect only a few loci to be affected, but large numbers of nuclear markers indicate that the structure of these populations is weak (Seeb et al. 2004; Beacham et al. 2009b; Seeb et al. 2011). It is possible that present-day gene flow among coastal southwestern Alaskan populations has maintained the reduced genetic divergence (Utter et al. 2009; Olsen et al. 2010), although some of the populations that have the smallest divergence are thousands of kilometers apart. Divergent populations of Pacific salmon generally demonstrate an isolation-by-distance (IBD) pattern “as the fish swims” rather than “as the crow flies”.

These dispersal patterns are consistent with the concept that populations that are in close proximity show less divergence than more distant ones because proximal populations exchange more migrants. Therefore, weak genetic structure would be observed among geographically distant populations only if substantial gene flow occurred serially between neighboring populations for many generations or if long distance dispersals were common.

One untested hypothesis is that presently observed patterns of genetic divergence reflect physical connections among coastal southwestern Alaskan river systems that allowed gene flow to erode genetic divergence, but those connections no longer exist. Some of these physical connections could have resulted in populations that are currently thousands of kilometers apart to being within tens of kilometers of each other and would have provided short corridors for gene flow. Geological data indicate that connections occurred between the Lower Yukon and Kuskokwim rivers at least twice (Creager & McManus 1967; Shepard & Wanless 1971). The first was near the village of Kalskag at the existing Yukon-Kuskokwim portage (Dougan 2010); and the second was through what is now the Johnson River, which empties into the Lower Kuskokwim River downstream from Bethel (Figure 3.1a, 3.1b). The dates of these connections are unknown, but when the Yukon and Kuskokwim Rivers were connected they emptied through the Kuskokwim River embayment and gene flow would have been likely among populations on the lower stretches of these two rivers.

Historical connections between the Nushagak and the Kuskokwim rivers are also likely. From bathymetry data and known rates of sea level increases (Fairbanks 1989), it

has been deduced that approximately 11 500 years ago, the lower reaches of these drainages joined near what is now Port Moller and then emptied into the Bering Sea at the present location of the Bering Canyon on the shelf break (Hopkins 1967) (Figure 3.1a). In addition, the Mulchatna River (a tributary of the Nushagak River) may have established connections to the Stony River (a tributary of the Middle Kuskokwim River) near Telequana Lake (Figure 3.1c), but the timing of this event has not been established (Maddren 1910). In this study, we used data for western Alaskan chum salmon populations that were genotyped with 58 SNPs (Garvin & Gharrett 2007, 2010; Garvin et al. 2010; Seeb et al. 2011), and 12 microsatellites (Kondzela et al. In Preparation) to ask if present-day or historical connections among populations of coastal southwestern Alaskan chum salmon better explain their present day weak genetic divergence.

Materials and Methods

Populations and genotype data

We used data from populations of western Alaskan chum salmon that were genotyped at single nucleotide polymorphism (SNPs) (Garvin & Gharrett 2007, 2010; Garvin et al. 2010) and microsatellite loci (Kondzela et al. In Preparation) (Table 3.1, Table 3.2).

Thirty-five populations from western Alaska were available for analyses, but complete suites of data were not available for all populations. Therefore, we compiled one data set that consisted of 25 populations genotyped at 12 microsatellite and 58 SNP loci [25P70L], which represented the largest geographic range for chum salmon in western Alaska. We also compiled a second data set that focused on the coastal southwestern geographic region and was comprised of 21 populations genotyped at 50 SNP loci [21P50L]. Populations were divided into eight regional groups designated as: A – Kotzebue Sound, B – Norton Sound, C- Lower Yukon River, D – Middle Yukon River, E – Lower Kuskokwim River, F – Middle Kuskokwim River, G – Upper Kuskokwim River, and H – Bristol Bay (Figure 3.1a, Table 3.1).

Hierarchical G-test

Genetic divergence within and among regional groups of chum salmon populations was tested with log likelihood ratios (*G*-tests) that were calculated in ExcelTM (McDonald 2009). The advantage of the *G*-test is that tests can be summed across loci and different regional groups to identify significance at hierarchical levels. It has been demonstrated

with simulations that the G -test can have high type I error but high power (Ryman et al. 2006), which is largely because the G -statistic does not approximate a chi-square distribution for low numbers of expected alleles. Therefore, for the multi-allelic microsatellite loci, some alleles were combined. We determined the expected number of alleles at each locus for the smallest sample. If that was less than four, the alleles were binned with the next largest sized one. A G -statistic was calculated for each locus and significance was determined by summing over all loci for each regional group. Regional groups were determined based on the geographical locations from which the samples were taken and based on the present-day courses of the river systems (Figure 3.1). An approximate F -test can be constructed that can compare the extent of divergence among and within populations (Hawkins et al. 2002):

$$F_{df_{among}df_{within}} = \frac{\frac{G_{among}}{df_{among}}}{\frac{G_{within}}{df_{among}}}$$

where df = the degrees of freedom.

Measures of divergence

We used the program GDA (Lewis & Zaykin 2001) to estimate locus-by-locus values and values over all loci for allele numbers, expected and observed heterozygosities, θ for nuclear loci (analogous to F_{ST} ; Weir & Cockerham 1984) and Φ_{ST} for the mitochondrial variants (Excoffier et al. 1992). Locus-by-locus and overall Jost's D_{EST} was calculated with the function 'D_Jost', which is available in the 'adegenet' package (Jombart 2008) in the R environment. For comparative purposes, we also

calculated θ for each locus with data from only the coastal southwestern populations (θ_{CSW}).

Principal components analysis (PCA)

A PCA was performed on the allele frequency data from the [25P70L] data; all allele frequencies were arcsine-square root transformed prior to the analysis in SYSTAT (Sokal & Rohlf 1994). Microsatellite alleles were binned as described for the G -test. The loadings for the first three components were used to plot the PCA.

Trees

A neighbor-joining tree was constructed observed allele frequency data with Cavalli-Sforza and Edwards (1967) chord distances. A consensus tree was produced from 1000 neighbor joining trees generated by bootstrapping loci and combining them with the CONSENSUS package in the program PHYLIP (Felsenstein 2004). The trees were drawn with the software Dendroscope (Huson et al. 2007).

Outlier analysis

We used the program Arlequin 3.5 (Excoffier & Lischer 2010) to identify genetic markers that showed larger than expected F_{ST} values as compared to a null distribution based on expected heterozygosity. Previous studies that simulated loci under selection

among populations showed that Arlequin 3.5 can have high type I and type II errors (Narum & Hess 2011); however, the two alternative methods that identify outlier loci (FDIST2 and BAYESCAN (Beaumont & Nichols 1996; Foll & Gaggiotti 2008)) do not take into account the hierarchical structure of populations, which clearly exists for the populations in this study. For this analysis, we included data from 70 loci and all samples from coastal southwestern Alaska because they showed the weakest genetic structure with both the neighbor joining trees and the PCA.

Data from the Salmon and Tatlawiksuk rivers were excluded because information was unavailable for all 70 loci, and fall-run Upper Kuskokwim fish (Gilk et al. 2009) were excluded because gene flow is unlikely between those populations and summer-run fish because they have different spawning times. Fish from the Middle Yukon populations also spawn later than summer-run fish so they were excluded as well. However, the samples from the Takotna River from the Upper Kuskokwim drainage were included in the analysis because they clustered with the samples from the Middle Kuskokwim River and they are summer-run fish. We tested for outliers with data from (1) only SNPs, (2) only microsatellites, and (3) both marker types combined.

Isolation by distance

Because the isolation-by-distance (IBD) analysis assumes that drift rather than diversifying selection is responsible for divergence among populations, we eliminated loci that might potentially be under positive directional selection by removing outlier loci

from the dataset prior to the IBD analysis. Pairwise F_{ST} values were then calculated among populations with the software program GDA.

We used the program GENEPOP (Rousset 2008) to test for IBD among and within summer-run chum populations from the Yukon, Kuskokwim, and Nushagak drainages as they exist today and again assuming historical connections among those systems. Only one population from the Nushagak drainage was available for analysis with the [25P70L] data set, but samples from more populations were available for the [21P50L] data set (Table 3.1). Therefore, we tested for IBD between the Yukon and Kuskokwim drainages with the [25P70L] data and between the Kuskokwim and Nushagak with the [21P50L] data. In order to determine if the exclusion of 20 loci affected the IBD results for the test between the Kuskokwim and the Nushagak rivers, we tested for IBD between the Yukon and Kuskokwim drainages with the [21P50L] data to see if the results were similar to the analysis with the [25P70L] data.

GENEPOP estimates the relationship between the genetic and water distances (d) from:

$$\left(\frac{F_{ST}}{1 - F_{ST}} \right) \sim \hat{a} + \hat{b}d$$

where \hat{a} is the intercept and the slope, $\hat{b} = \frac{1}{4D_e\sigma^2}$, is inversely proportional to the effective linear density of individuals (D_e) and the mean-squared parent-offspring distance (σ^2). Significance was determined with a Spearman rank correlation coefficient. We analyzed

the same data with the Mantel test available with the ‘ade’ package in the R environment to determine the correlation coefficient because GENEPOP does not provide one.

The present-day great circle geographic distances were obtained by drawing the shortest distance between the sample location of each population through freshwater or marine environments with Google Earth™. The latitude and longitude coordinates identify the weir locations or sonar stations where the samples were taken, which are not necessarily the locations of the spawning populations. Therefore, we used the latitude and longitude coordinate for the centroid of each drainage provided by the sub-watershed delineation tool in the Riverscape Analysis Project (Whited et al. 2012).

The possible historical connection between the Yukon and Kuskokwim rivers and between the Kuskokwim and Nushagak rivers were drawn by following published reconstructed connections (Maddren 1910; Shepard & Wanless 1971). These historical connections were then drawn in Google Earth™ and the shortest geographic distances among populations through freshwater or marine environments were recalculated assuming that the present-day populations were located in the past where they are today.

Dispersal distance

The IBD analysis provides an estimate for the slope (\hat{b}) of the regression of the genetic distances on the geographic distances. Rearrangement of the terms in the equation give the relationship $\sigma \approx (4\hat{b}D_e)^{-1/2}$. The approximate 95% confidence interval of the geographic distance of parent-offspring pairs is 4σ for a normal

distribution, which provides an estimate of the dispersal distances of individuals within populations (e.g. Gharrett et al. 2012)

We calculated density with empirical data from the Kuskokwim River and the slope determined in GENEPOP with the [25P70L] data set. We used estimates for the total run size of the Kuskokwim River from previous work (Bue et al. 2007) for the annual census size (N_c). Although this is likely an underestimate of the census size, it provides an upwardly biased estimate of dispersal distance and is therefore a conservative over-estimate of the potential gene flow among populations. This estimate also includes fall-run fish, but those run sizes are small relative to the summer-run populations and the estimate of Bue et al. (2007) is conservative. The density ($D = \frac{N_c}{\text{Distance}}$) was estimated by dividing the harmonic mean of the annual census size estimates from Bue et al. (2007) by the total linear distance of the Kuskokwim River, which was calculated by summing the linear distances between the centroids all of the major drainages for summer-run chum salmon (Table 3.1).

The effective density (D_e) was determined from a sensitivity analysis of plausible ratios of $\frac{N_E}{N_C}$. In order to determine this ratio, we needed to account for two factors: (1) chum salmon have overlapping generations and an average age of return of four years in the northern hemisphere (Groot & Margolis 1991), which means that the yearly census size of a population is only a portion of the existing population. Therefore, we multiplied the census data by four to account for potential future breeders that remained at sea. (2) The effective population size (N_e) can be considered the efficiency of passing genes from

one generation to the next. For salmon, the determination of N_e can be challenging because of fluctuating population sizes and overlapping generations (Falconer & Mackay 1996; Waples 2002). Although these methods make some large assumptions, our purpose was to estimate the order of magnitude of the dispersal distance (i.e. is it tens, hundreds, or thousands of kilometers?). Therefore, we used a range of $\frac{N_e}{N_c}$ ratios from 0.0125 to 0.75, which brackets previous estimates of N_e for populations of chum salmon in Norton Sound (0.063 to 0.619; Burkhart & Dunmall 2005; Olsen et al. 2005) and reported the mean dispersal distances for the range of $\frac{N_e}{N_c}$ ratios.

Results

Measures of divergence

The log-likelihood ratio tests were all highly significant ($p > 0.001$) except for the populations from the Middle Kuskokwim River (Table 3.3). However, the total divergence among populations within regions was less than the total divergence among regions ($p < 10^{-4}$), which suggests further reduction of genetic divergence within regions than one would expect given the total genetic variation in the geographic region. We also tested for panmixia with data from the [21P50L] data set. The G -tests within and among these three systems differed (Table 3.4), which suggests these populations do not represent panmixia. Finally, the θ_{CSW} values over all loci for coastal southwestern populations were an order of magnitude lower compared to all of the populations in western Alaska (Table 3.2).

Principal components analysis

The PCA for the [25P70L] data set (Figure 3.2) indicated little divergence among southern Norton Sound, Lower Yukon, summer-run Kuskokwim, and Northern Bristol Bay populations. No additional resolution was provided by the third or fourth components. The Kotzebue Sound samples (A1-A3) are divergent from all other samples as are the Middle Yukon samples (D1, D2). The samples from southern Bristol Bay (H1, H2), the late-run Upper Kuskokwim populations (G2, G3), and northern Norton Sound populations (B1, B2) were also divergent, but the samples from the Nushagak River (H3) in northern Bristol Bay, the Takotna River (G1) in the upper reaches of the Kuskokwim system, and the Unalakleet River (B3) in southern Norton Sound clustered with the Lower Yukon (C4, C6), Lower Kuskokwim (E1-E3), and Middle Kuskokwim samples (F1, F3-F7).

Neighbor-joining trees

The consensus neighbor-joining tree supported the PCA (Figure 3.3a). The Kotzebue Sound, Norton Sound, Middle Yukon, and Upper Kuskokwim populations have large bootstrap estimates for the nodes connecting them; but the tree reveals weak genetic structure among coastal southwestern populations. The neighbor-joining tree drawn with the observed allele frequencies corroborates the weak genetic structure of the coastal southwestern populations seen with the consensus tree.

Outlier analysis

The outlier analysis identified four loci that differed significantly ($p < 0.05$) from the null distribution and indicated that those regions of the DNA or closely linked regions may have experienced divergent selection among individuals from the 14 chum salmon populations in the analysis (Figure 3.4). The same four SNPs were identified as outliers when only the SNP data were used and when the combined SNP and microsatellite data were used. None of the microsatellites appeared to be outliers. All four of the SNPs were discovered in previous work that used Eco-TILLING to identify informative genetic markers for MSA during the discovery process (Garvin & Gharrett 2007). Although the outlier test in Arlequin 3.5 can exhibit type I error (Narum & Hess 2011), we removed the data for these four loci prior to the IBD analysis to assume that the markers we used were selectively neutral.

Isolation by distance

The tests for IBD differed when the different models were used. The test for IBD with data from 66 loci (the 25P70L data minus the four outliers) and pairwise geographic distances as they exist today among the populations from the Lower Yukon and Kuskokwim rivers was not significant. However, the test for IBD was significant when the geographic distances were calculated by assuming an historic connection between those rivers (Figure 3.5). In addition, the slope of the line was approximately 5-fold higher than for the IBD analysis with the present-day geographical distances. The tests for IBD with data from the [21P50L] data were similar when compared to the values

when 70 loci were used (data not shown). This suggests that 50 loci provide sufficient information for the IBD analyses and data for those 50 loci were used for tests that included more populations.

The test for IBD between the Kuskokwim and Nushagak drainages was not significant if present-day connections were assumed among populations but the test was significant when the connection between the Stony and Mulchatna rivers was assumed (Figure 3.6). We also tested for IBD within the Yukon and Kuskokwim rivers to determine if the slopes of the lines were similar, which might suggest parallel historical demographic processes. The test for IBD for the seven populations within the Yukon was not significant ($p < 0.461$, slope = -4.39×10^{-4}), but the test for IBD within the Kuskokwim was significant ($p < 0.04$, slope = 3.15×10^{-6}). The IBD analysis within the Kuskokwim was similar when the 25P70L data were used (slope = 3.03×10^{-6} ; $p < 0.02$). There were insufficient samples from the Nushagak River for either data set to test for IBD within that drainage.

Dispersal distance on the Kuskokwim River

The data from all available putatively neutral markers (66 loci) for the ten summer-run chum salmon populations from the Kuskokwim River were used for an IBD analysis. We used the slope from the [25P70L] data and the mean density of chum salmon to estimate the straying distance for $\frac{N_e}{N_c}$ ratios that ranged from 0.0125 to 0.75 (Table 3.5). The straying distances spanned from 28.7 to 222.3 km.

Discussion

In this work we tested the hypothesis that recent historical connections between major drainages in western Alaska allowed gene flow among now-distant populations and resulted in the present-day reduced genetic divergence among coastal southwestern chum salmon populations that inhabit a substantial geographic area. The IBD analysis and the dispersal distances that we calculated among summer run chum salmon populations on the Kuskokwim River (between 28.7 and 222.3 km) support the hypothesis. Although the dispersal estimates exceeded those observed for pink salmon (*O. gorbuscha*) between adjacent streams (Gharrett et al. 2001) (maximum likelihood estimates from several years of data), they do not support long distance migration among populations for the cause of present-day weak genetic structure among these populations.

Chum salmon populations between Japan and Kamchatka and between the Gulf of Alaska and the Pacific Northwest have genetic diversity estimates that are several-fold higher than for populations in the geographic area between Kotzebue Sound and Bristol Bay in western Alaska (Seeb & Crane 1999; Beacham et al. 2009b; Seeb et al. 2011). The exception is fall-run fish that spawn in the upper reaches of both the Yukon and Kuskokwim rivers, which are highly divergent from the summer-run populations but are not the focus of this study and do not confound MSA applications. This reduced genetic divergence among summer-run populations in western Alaska is likely a result of the geological history of the freshwater and marine habitat since the LGM.

The additional weak divergence of coastal southwestern populations nested within western Alaska could be explained by two processes that occurred recently (geologically speaking): (1) population reduction (extirpation) and (2) recolonization into a dynamic ecosystem in which major river systems were repeatedly altered. Historical evidence indicates at least two significant population reductions and several instances of major modifications to river drainages that occurred during the period from the Pleistocene/Holocene transition to the present.

Archaeological evidence from Broken Mammoth in central Alaska near the Tanana River revealed that one or more species of salmonid was present in central Alaska approximately 14 000 years ago (Hoffecker & Elias 2003), and certainly freshwater species survived the LGM (McPhail & Lindsey 1986). But paleo-climate data suggest that the environmental conditions would not have been as favorable as they are presently for populations of anadromous salmon. Rivers that were present at the LGM would have drained at the present-day Bering Sea shelf, and as a result, the estuarine habitat that is necessary for the early life-history stages of chum salmon may have been reduced. In addition, the climate likely caused the Bering Sea to be ice-covered for 9 months of the year, and food abundances based on diatom microfossil assemblages from sediment cores suggest that the Bering Sea during the LGM was much less productive than it is today (Sancetta 1983; Mann & Hamilton 1995). For these reasons, chum salmon populations that existed in the warm period prior to the LGM were likely reduced or in some instances extirpated. Any chum salmon populations that may have survived the LGM would ecologically resemble those that exist in similar environmental conditions today

(e.g. those in the Beaufort and northern Chukchi seas); the populations would have been small and many of them ephemeral.

However, sea levels rose rapidly enough to cause shorelines to retreat up to 100 meters per year (Fairbanks 1989; Manely 2002) and would have created potential estuaries on the relatively flat newly formed coastline. By about 10 000 years ago, the connection between the Bering and Chukchi seas had been re-established and the planktonic productivity in the Bering Sea began to increase (Sancetta 1983; Mann & Hamilton 1995). Beringian chum salmon spawning habitat flooded by rising seas after the LGM would have forced populations to colonize new locations or become extirpated.

It is likely that the coastal southwestern chum salmon are either the descendants of a paleo-Beringian invasion from the rising seas, expansions from small populations that survived the LGM, or both. By approximately 5000 years ago, the Bering Sea had reached its current level, and the present shorelines had been established (Manely 2002). Subsequent to this (Nelson & Creager 1977) and perhaps as recently as 1200 years ago (Dupré 1988; Shaw 1998), the Yukon River mouth moved north to flow into Norton Sound. Populations in the lower reaches of the Yukon and Kuskokwim rivers would have formed a large meta-population that was repeatedly connected, disconnected, and reconnected as the Yukon River meandered across the broad flat plain of Beringia and the present-day Yukon-Kuskokwim Delta. Many populations would have winked in and out and some may have merged in this fluctuating habitat, which was a result of variable discharges from glaciers that melted with the warming climate. Even today about one-third of the present-day flow of the Yukon River is due to glacier melt (Brabets et al.

2000); any substantial retreat of glaciers would have caused changes in the flows of the Yukon River and its drainages.

There were several periods during the Holocene in which temperatures rapidly returned to those of the LGM and were followed by rapid warming that may have caused increased flows to establish connections among river drainages (Alley 2000). The first began after the Younger Dryas between 9000 and 5000 years ago when the northern hemisphere experienced the maximum temperatures of the Holocene. The discharge rates of drainages in western Alaska were likely highly variable as they were in the Atlantic Ocean (Fairbanks 1989), and is still the case in many Alaskan drainages presently. These increased flows could have established connections among the headwaters of the Nushagak and the Stony River as well.

The weak genetic structure that is present today may indicate that these large meta-populations have not yet reached migration-drift equilibrium. If chum salmon have been present in these drainages since the Holocene Maximum, which is roughly 1200 generations, there should have been time to establish equilibrium (Waples et al. 2008). However, this state is reached at a rate that is proportional to the effective population size and inversely proportional to the migration rate in a population ($t_{1/2} \approx \frac{\ln 2}{2m + \frac{1}{2N_E}}$; Crow and Aoki 1984). Low gene flow or large effective population size would retard progress toward an equilibrium.

Many of the coastal southwestern chum salmon populations are very large, and physical connections between the Yukon/Kuskokwim and the Kuskokwim/Nushagak

rivers would have resulted in large numbers of migrants between nearby populations. It is important to note here that we detected an IBD pattern within the Kuskokwim River drainage but not within the Yukon River drainage (too few populations were available to test for IBD within the Nushagak River). This is consistent with the idea that as the mouth of the Yukon River shifted to the north, it may have disrupted population structure in that habitat and erased IBD patterns that had been previously established. Migrants that formed populations in these areas may have originated from the Kuskokwim River or upstream on the lower Yukon River.

Subsequent to the Holocene Maximum, two neoglacial periods were followed by rapid warming events that may have also altered discharge rates from western Alaskan drainages and established migration corridors. Data from sediment samples taken from Ongoke Lake in southwestern Alaska indicate that approximately 1600 years ago this part of the world experienced the coldest, driest climate in the past 2000 years (Chipman et al. 2008) (Figure 3.1a), known as the First Millennial Cold (FMC) period. Other work has established that glaciers advanced during this same period, which was followed by a warmer climate and glacial retreat (Wiles et al. 2008; Barclay et al. 2009). Approximately 300 years ago the Little Ice Age (Mann 2002), which was colder but wetter than the preceding cold period, caused glaciers to advance far enough to displace native Tlingit people from their village in what is now Glacier Bay National Park (Appleton et al. 2010). The cooler climate caused the Kaskawulsh Glacier at the southern end of Kluane Lake in the Yukon Territory to advance (Figure 3.1a) and block the outlet of the lake (Clague et al. 2006). The rise of the lake level caused the outlet to form at the lake's

southern end and empty into the White River, which now contributes 10% to the flow of the Yukon River (Brabets et al. 2000). Finally, glaciers in the Ahklun mountains in southwestern Alaska that formed during the Little Ice Age were reduced by 50% in volume during the subsequent warm period (Levy et al. 2004). Any or perhaps all of these events may have disrupted IBD patterns that had formed in coastal southwestern populations after they had expanded from the LGM.

Alternatively a more recent environmental perturbation may have caused the extirpation or severe population reduction of chum salmon in western Alaska that was followed by population expansion and colonization. During the FMC period, the abundance of sockeye salmon from Kodiak Island declined substantially (Finney et al. 2002), and archaeological data indicate that native communities near Cape Nome in Norton Sound (so-called Norton Phase) ceased the use of salmon as a food resource for nearly 300 years even though they had done so for the 1500 years prior to that period. This coincided with a migration of Norton-Phase people south into the Alaska Peninsula near the Ugashik and Naknek drainages (Dumond 1998). Approximately 1200 years ago, simultaneously cultural and technological changes occurred in native communities located near the Ugashik River, the Naknek River, the Pacific coast of Shelikof Strait, Cook Inlet, and Kodiak Island, and the use of salmon for subsistence food became common (Bockstoce 1973, 1979; Yesner 1998). This was considered to be highly significant ($p < 0.0005$) in the context of 9000 years of human occupation in Alaska (Mills 1994).

The cold and dry climate may have created unsuitable habitat in the topographically flat areas of coastal southwestern Alaska and salmon populations may have declined substantially between 1600 and 1200 years ago. Interestingly, populations of lake trout (*Salvelinus namaycush*) have survived several glacial periods throughout Alaska, but have never colonized the same flat areas inhabited by coastal southwestern chum salmon populations, presumably because of reduced flows and shallow water habitat (McPhail & Lindsey 1986).

Flows in major rivers during the FMC would have been reduced by the lack of precipitation and the reduction of glacial melt. The decrease in flows, combined with a reduction in water temperatures may have eliminated the viable habitat that is necessary for chum salmon. The development of salmon eggs in freshwater habitat is tightly coupled to water temperature so that fry emerge when food resources are available for growth (Beacham & Murray 1990), and chum salmon depend on oxygenated water from either turbulent areas of the river or upwelling from groundwater (Groot & Margolis 1991). As environmental conditions improved, recolonization could have occurred from distant sources and likely would have begun in the middle and upper reaches of the drainages where habitat would have improved first because it is the least flat (flows would have provided hospitable habitat). Over time, those upper river populations provided colonists for the lower reaches, which produced the IBD signal we observed among the Kuskokwim River populations, but that signal was erased on the Yukon River because the Yukon Delta habitat is the least favorable and was only very recently reinhabited.

This more recent possible extirpation event does not preclude the possibility of expansion through connections among western Alaskan drainages. In addition, it could also explain the genetic similarity of the populations from the Lower Yukon, the Kuskokwim, and the Nushagak rivers to southern Norton Sound drainages because the latter geographic area could have been recolonized at the same time as the former. Future geological or hydrological work focused on the timing of the most recent connection between the Yukon and Kuskokwim, and Kuskokwim and Nushagak rivers may provide support for one of these hypotheses as could analysis similar to that of Finney et al. (2000) in western Alaskan salmon habitat.

A more recent timing for one of these hypotheses is suggested by the fact that the southwestern coastal Alaskan chum salmon populations have divergence estimates that are an order of magnitude lower than for chum salmon in western Alaska overall ($\theta_{\text{CSW}} = 0.001$ versus $\theta = 0.016$). In addition, observed divergence estimates among Chinook salmon that were introduced into New Zealand rivers at the end of the 19th century (Kinnison et al. 2002) and pink salmon that colonized new habitat as glaciers receded in Glacier Bay National Park about 125 years ago (Kondzela 2010) are also an order of magnitude higher than the coastal southwestern chum salmon populations here. Chinook salmon would be expected to show greater divergence and the anthropogenic introduction produced a colonization mechanism that was likely very different than western Alaskan chum salmon, but pink salmon typically demonstrate IBD patterns similar to those of chum salmon (Quinn 2005) and would have followed a natural progression of colonization. Regardless, the difficulties associated with applied genetic work to chum

salmon populations from coastal western Alaska may be a result of recent events, and therefore large numbers of neutral markers may be inadequate.

Acknowledgements

We would like to thank several funding agencies for providing stipend and laboratory support: the Rasmuson Foundation, the Arctic–Yukon–Kuskokwim Sustainable Salmon Initiative (www.aykssi.org) awarded through the Bering Sea Fishermen’s Association, the University of Alaska Experimental Program to Stimulate Competitive Research (EPSCoR), and the U.S. National Oceanic and Atmospheric Administration (NOAA) Alaska Fisheries Science Center (AFSC). This work was also supported by a grant of the HPC resources from the Arctic Region Supercomputing Center. We would like to thank Edward Neal of USGS for his advice on hydrology and effects on salmon, Hans Thompson of ADF&G for his assistance with GIS data, and Diane Whited from the Flathead Lake Biological Station at the University of Montana. The findings and conclusions presented by the authors, however, are their own and do not necessarily reflect the views or positions of the funding agencies, or the University of Alaska Fairbanks, School of Fisheries and Ocean Sciences

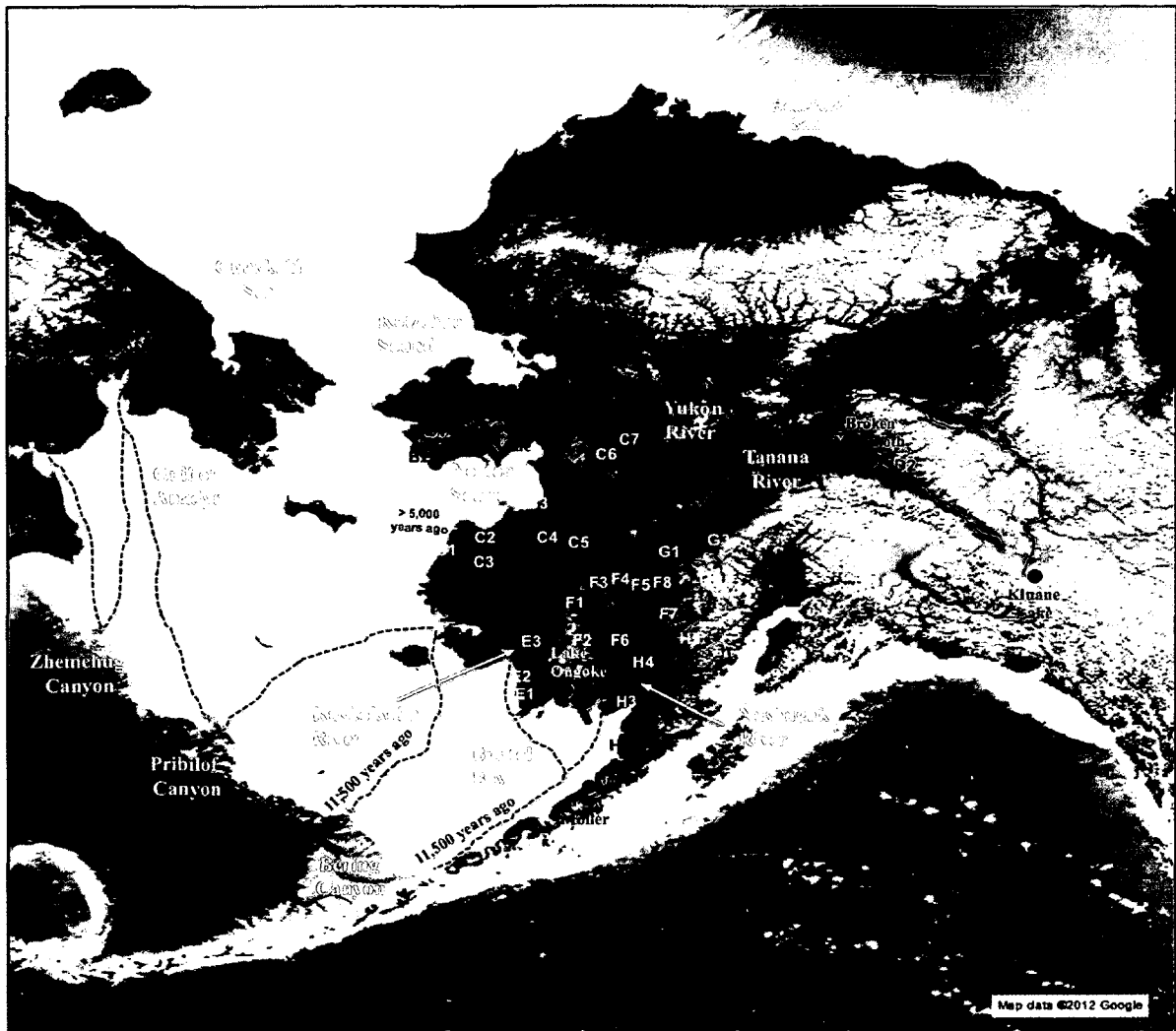


Figure 3.1a. Area of study. Populations of chum salmon that were sampled are identified with alpha-numeric codes (Table 3.1). The codes in yellow represent the coastal southwestern populations that show the weakest genetic structure and are difficult for MSA. Current river systems are indicated with blue lines and black broken lines indicate known and likely historical river systems. Other place names mentioned in the text are provided.

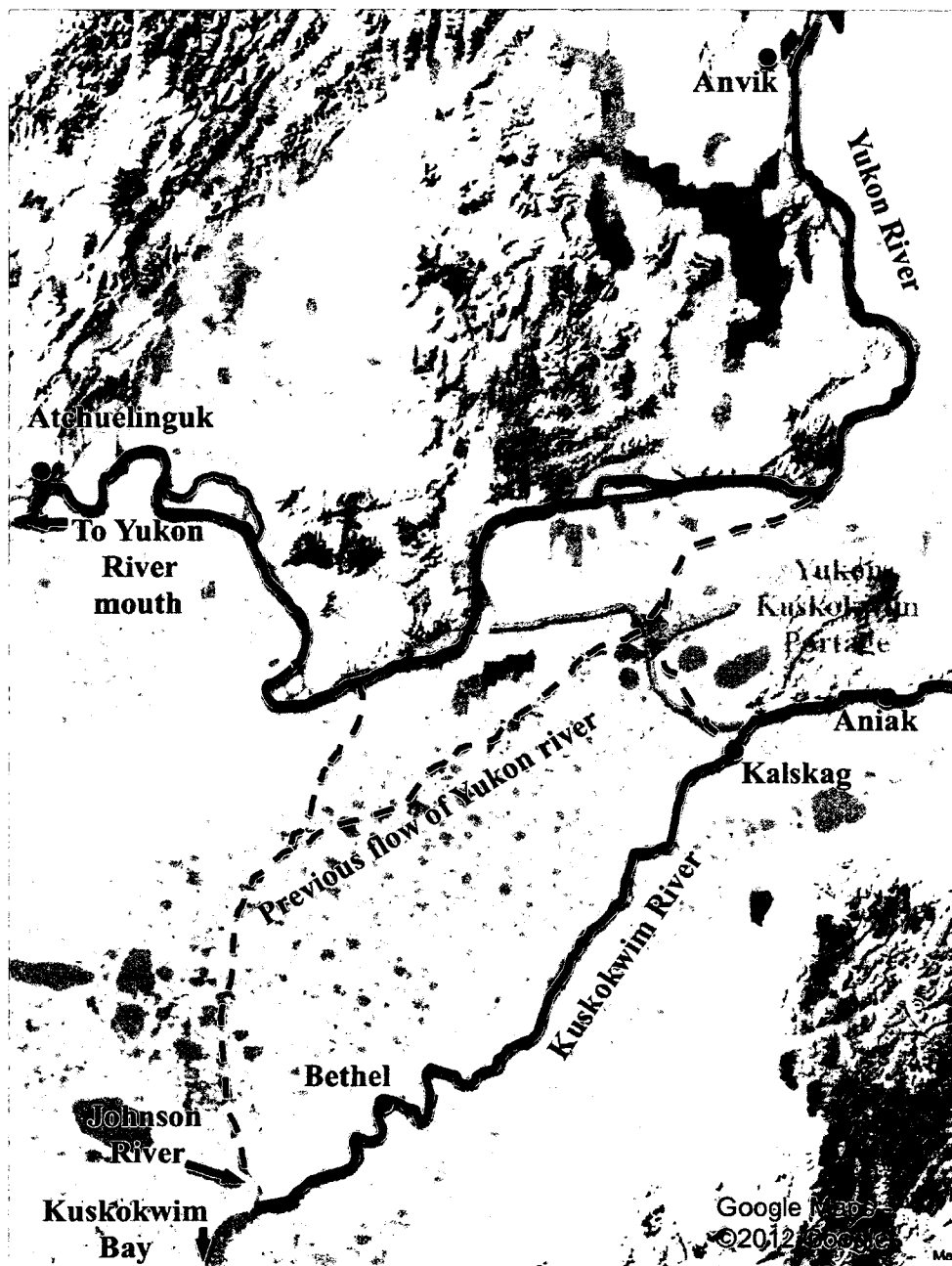


Figure 3.1b. Historical Yukon-Kuskokwim connections. A reproduction from Shepard and Wanless (1971) that shows the connections between the Lower Yukon and Kuskokwim rivers at two locations as well as the current Yukon-Kuskokwim Portage. The villages of Bethel, Kalskag, Atchuelinguk, and Aniak provide reference points.

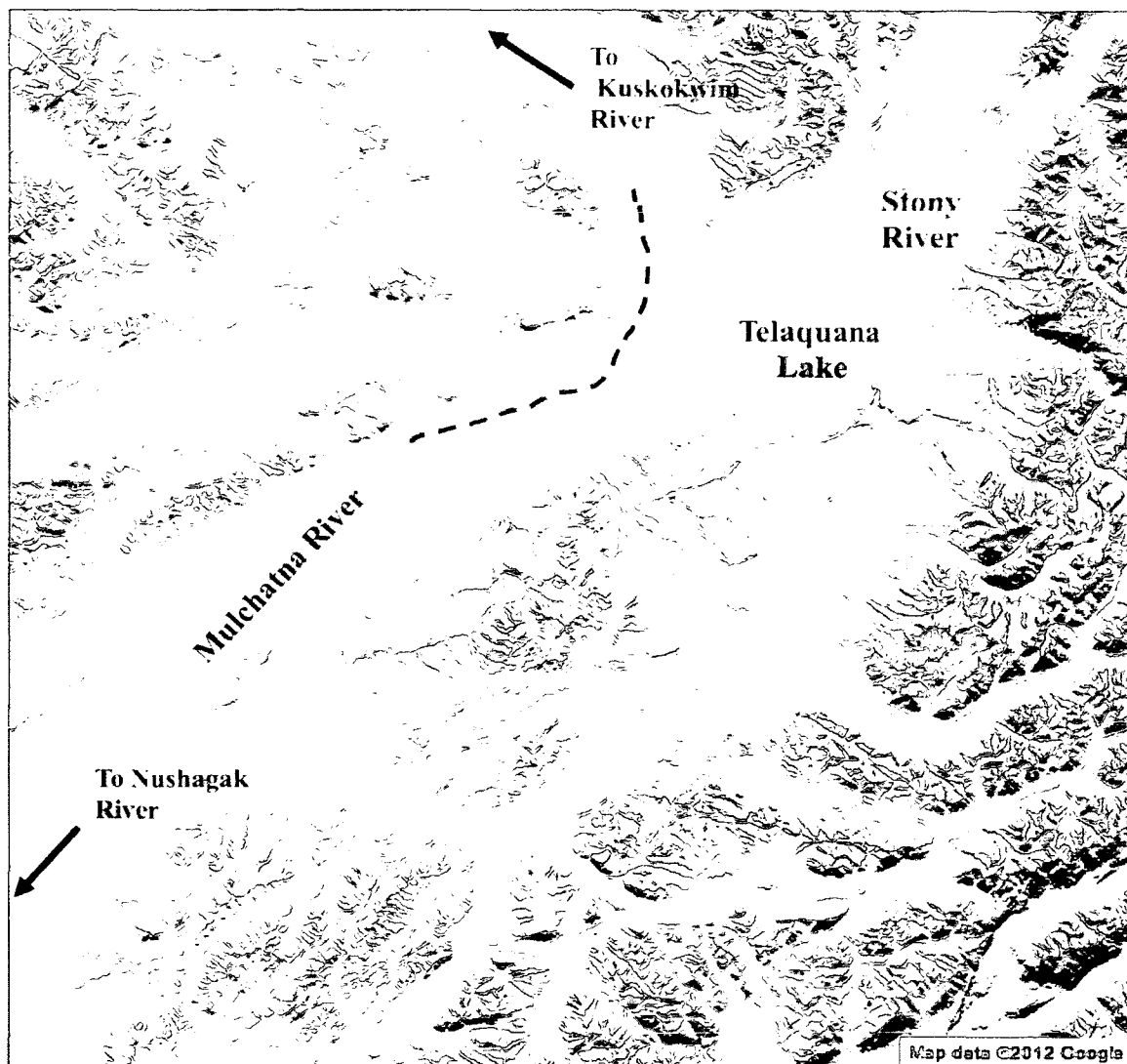


Figure 3.1c. Historical Kuskokwim-Nushagak connections. A topographical map that shows the likely connection between the upper Nushagak (Mulchatna) River and the Middle Kuskokwim (Stony) River. Current river systems are indicated with blue lines and black broken lines indicate likely historical river systems.

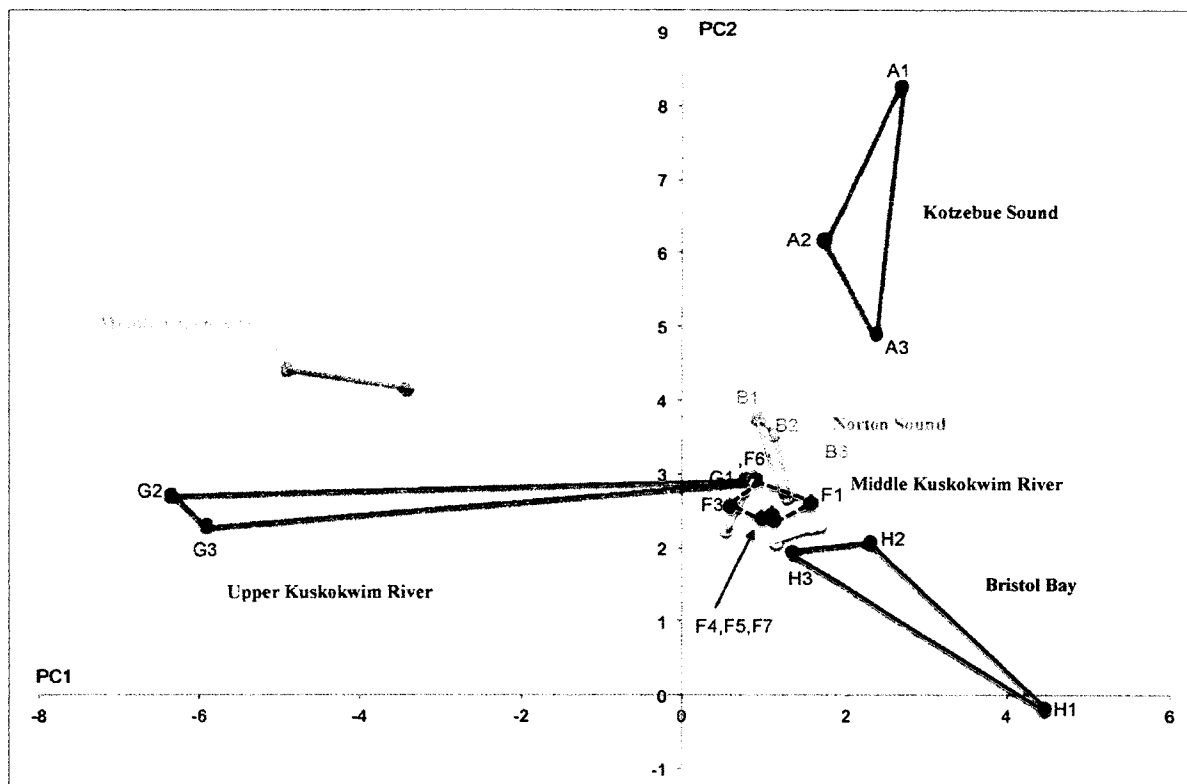


Figure 3.2. PCA of western Alaskan chum salmon populations. Populations are indicated with an alpha-numeric symbol, which corresponds to geographic regions listed in 3.1. A graph of first and second components show distinct clusters for Kotzebue Sound, Middle Yukon River, northern Norton Sound, southern Bristol Bay, and late-run Upper Kuskokwim River populations.

0.0010

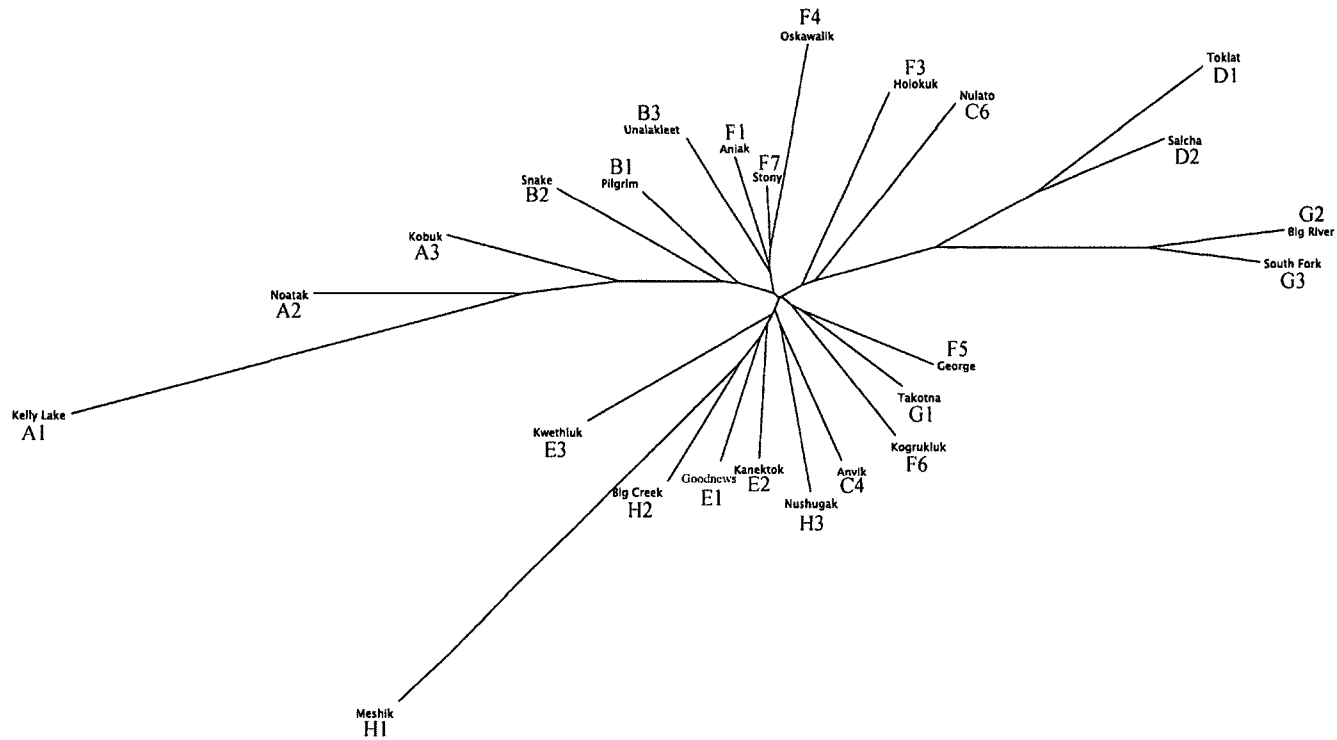


Figure 3.3a. The neighbor-joining tree drawn with genetic chord distances. The edge lengths reflect genetic distances from observed data allele frequencies and alpha-numeric codes correspond to Table 3.1 and Figure 3.1.

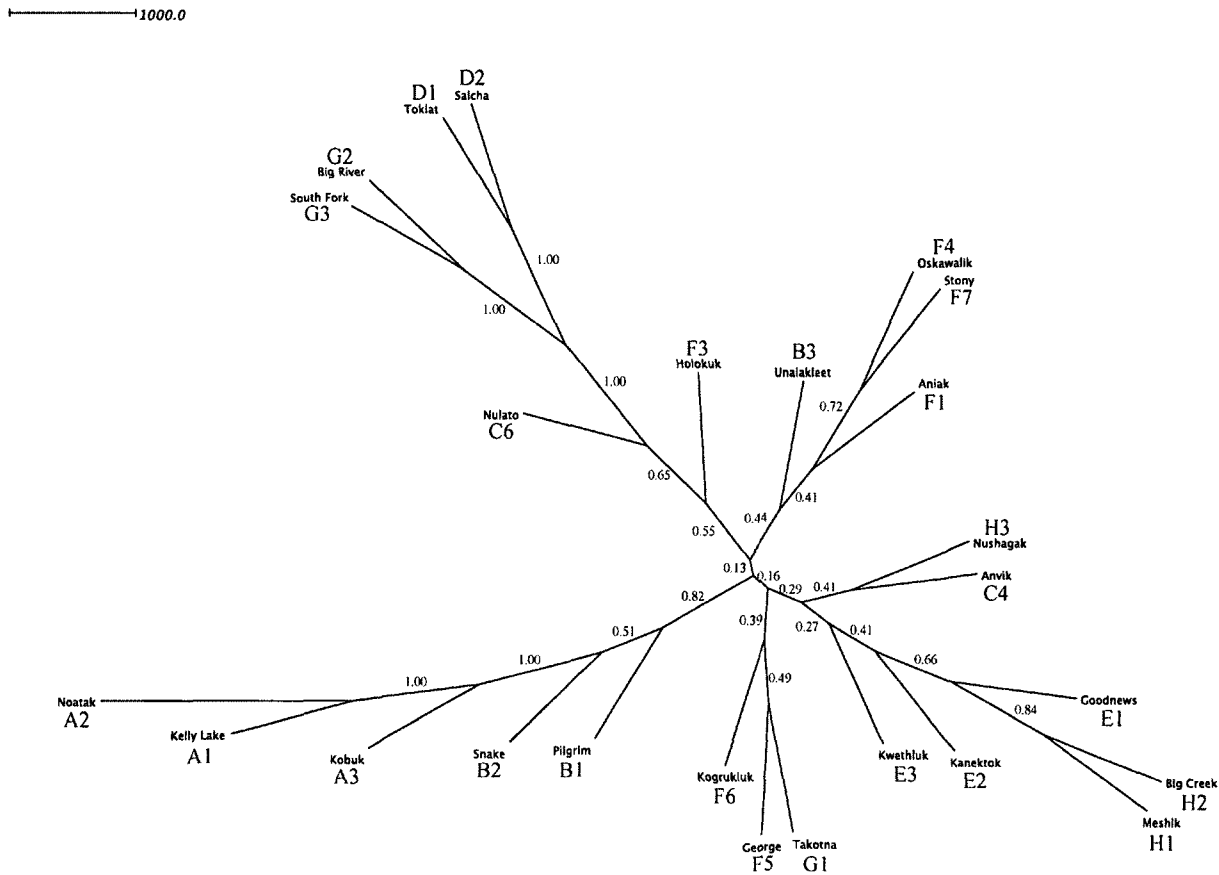


Figure 3.3b. Consensus neighbor-joining tree. Bootstrap estimates are given for the proportion of trees out of 1000 that gave that edge after sampling loci with replacement. The edge lengths reflect consensus numbers and alpha-numeric codes correspond to Table 3.1 and Figure 3.1.

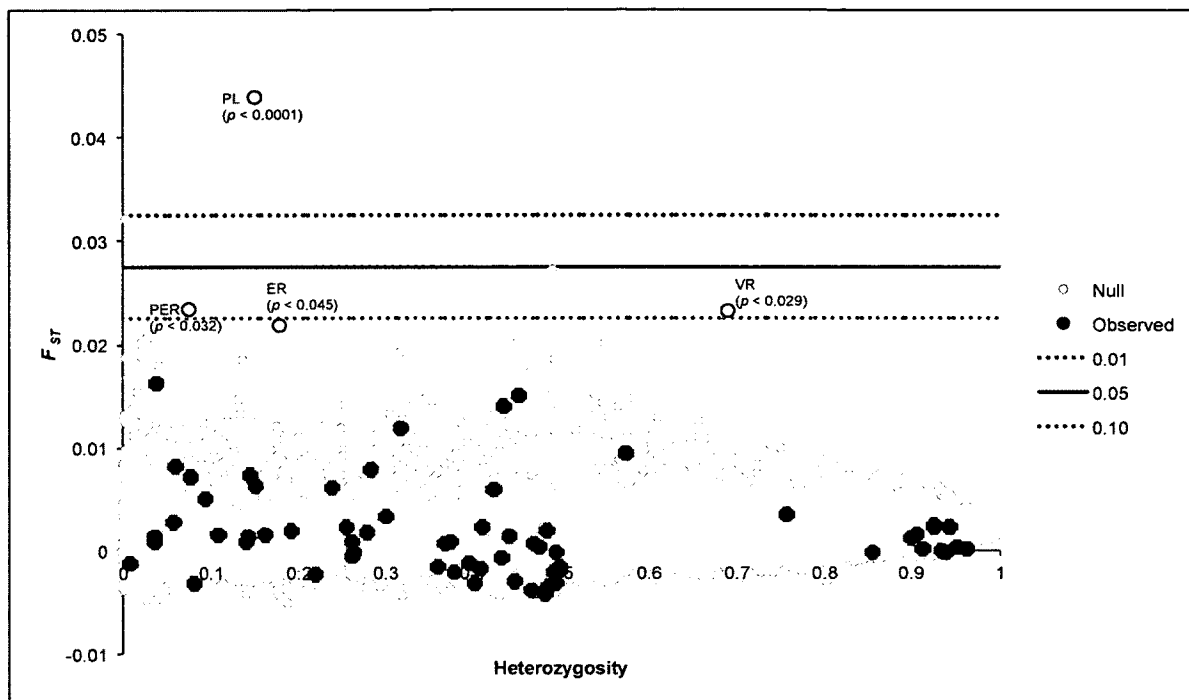


Figure 3.4. Outlier analysis. Data from 50 nuclear SNPs and 12 microsatellites used in this study were analyzed. The gray symbols denote the simulated null distribution based on a hierarchical population structure. The black symbols are the values for the empirical data. Outlier loci are indicated with empty circles and the name and associated probabilities (p) are shown next to each circle.

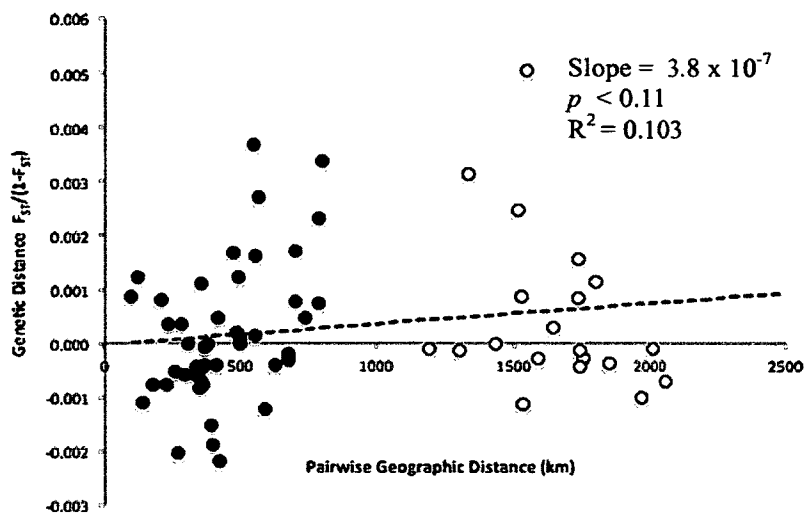


Figure 3.5a. IBD between the Yukon and Kuskokwim rivers with present geographical distances. The analysis was performed among the summer-run chum salmon populations from the Lower Yukon and the Kuskokwim rivers. The slope of the line is given along with the probabilities for the Mantel test in GENEPOP.

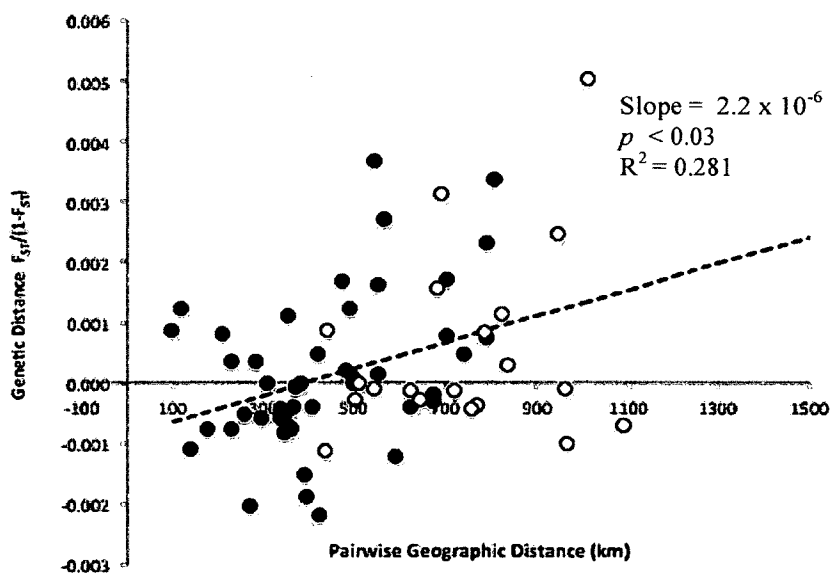


Figure 3.5b. IBD between the Yukon and Kuskokwim rivers with past geographical distances. The analysis was performed among the summer-run chum salmon populations from the Lower Yukon and the Kuskokwim rivers. The slope of the line is given along with probabilities for the Mantel test in GENEPOP.

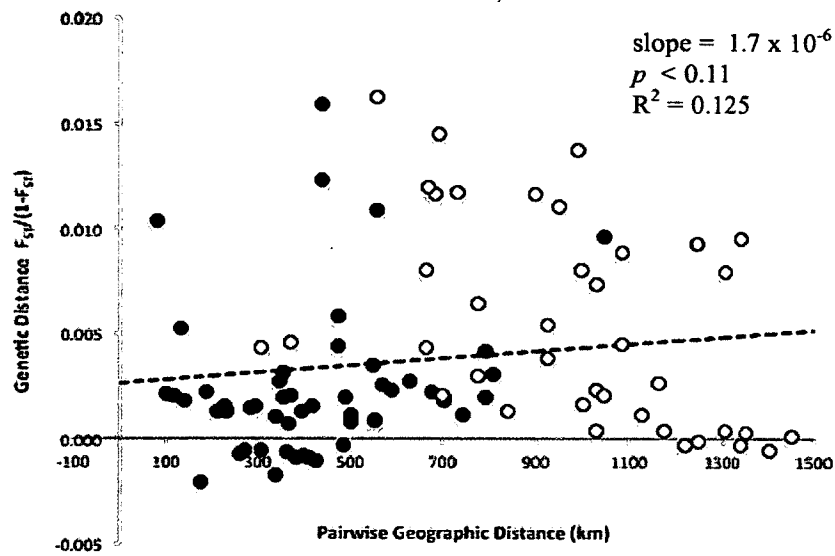


Figure 3.6a. IBD between the Kuskokwim and Nushagak rivers with present geographical distances. The analysis was performed among the summer-run chum salmon populations from the Lower Yukon and the Kuskokwim rivers. The slope of the line is given along with the probabilities for the Mantel test in GENEPOP

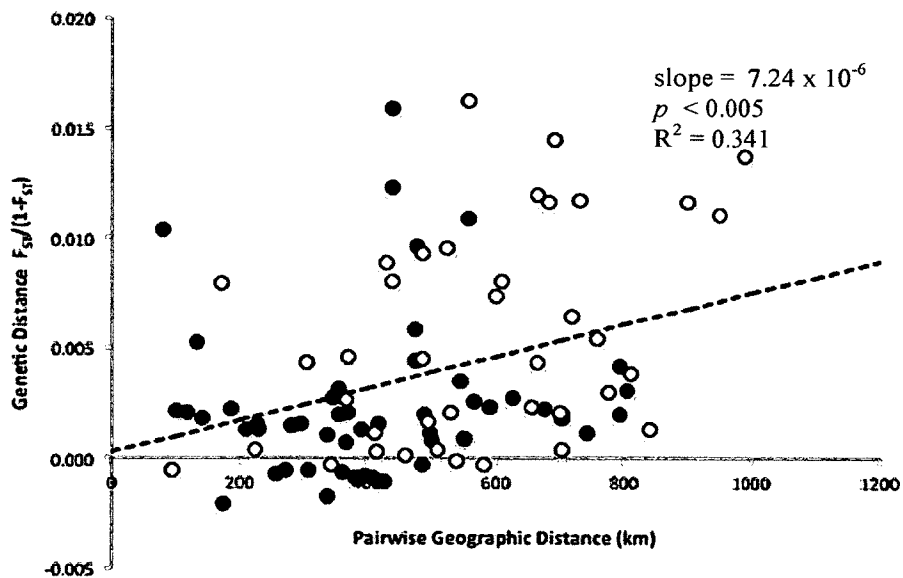


Figure 3.6b. IBD between the Kuskokwim and Nushagak rivers with present geographical distances. The analysis was performed among the summer-run chum salmon populations from the Lower Yukon and the Kuskokwim rivers. The slope of the line is given along with the probabilities for the Mantel test in GENEPOP.

Table 3.1. Geographical and run timing information of the samples. The Lat and Lon values are the latitude and longitude of the location where the samples were taken. The cLat and cLon values are the latitude and longitude of the centroid of each drainage. Fifty of the 58 SNPs were available for a detailed analysis of the coastal southwestern populations. The final two columns indicate whether a population was included in the 25P70L or the 21P50L data set.

| Sample Location | Code | Regional Group | Year | N | Lat | Lon | cLat | cLon | Source | Run Timing | 25P70L | 21P50L |
|-----------------|------|------------------------|-----------|-----|-------|---------|-------|---------|--------|------------|--------|--------|
| Kelly Lake | A1 | Kotzebue Sound | 1991 | 96 | 67.92 | -162.35 | 67.27 | -162.48 | ADF&G | Summer | Yes | No |
| Noatak | A2 | Kotzebue Sound | 1991 | 96 | 67.98 | -162.51 | 67.49 | -161.99 | ADF&G | Summer | Yes | No |
| Kobuk | A3 | Kotzebue Sound | 2000 | 96 | 66.92 | -160.81 | 67.06 | -156.51 | USFWS | Summer | Yes | No |
| Pilgrim | B1 | Norton Sound | 2004 | 96 | 65.16 | -165.22 | 64.92 | -165.10 | KWRK | Summer | Yes | No |
| Snake | B2 | Norton Sound | 2004 | 96 | 64.50 | -165.41 | 64.52 | -165.41 | KWRK | Summer | Yes | No |
| Unalakleet | B3 | Norton Sound | 2005 | 96 | 63.87 | -160.79 | 63.97 | -159.94 | KWRK | Summer | Yes | No |
| Black River | C1 | Lower Yukon River | 2006 | 95 | 62.35 | -165.35 | 62.09 | -164.81 | ADF&G | Summer | No | Yes |
| Andreasfky | C2 | Lower Yukon River | 1993 | 93 | 62.12 | -162.81 | 63.20 | -162.58 | USFWS | Summer | No | Yes |
| Achuelinguk | C3 | Lower Yukon River | 1989 | 93 | 61.96 | -162.83 | 62.01 | -162.73 | USFWS | Summer | No | Yes |
| Anvik | C4 | Lower Yukon River | 1989 | 75 | 62.68 | -160.20 | 63.12 | -160.59 | USFWS | Summer | Yes | Yes |
| Innoko | C5 | Lower Yukon River | 1993 | 86 | 62.25 | -159.56 | 62.68 | -159.56 | ADF&G | Summer | No | Yes |
| Nulato | C6 | Lower Yukon River | 2003 | 48 | 64.71 | -158.14 | 64.65 | -158.79 | USFWS | Summer | Yes | Yes |
| Gisasa | C7 | Lower Yukon River | 1994 | 95 | 65.25 | -157.71 | 64.81 | -159.02 | ADF&G | Summer | No | Yes |
| Toklat | D1 | Middle Yukon River | 1994 | 96 | 64.45 | -150.31 | 63.82 | -150.05 | USFWS | Fall | Yes | Yes |
| Salcha | D2 | Middle Yukon River | 1994 | 96 | 64.47 | -146.98 | 64.83 | -144.71 | USFWS | Summer | Yes | No |
| Goodnews | E1 | Lower Kuskokwim Bay | 1989 | 96 | 59.10 | -161.56 | 59.32 | -160.68 | USFWS | Summer | Yes | Yes |
| Kanektok | E2 | Lower Kuskokwim Bay | 1989 | 75 | 59.75 | -161.93 | 59.84 | -160.35 | USFWS | Summer | Yes | No |
| Kwethluk | E3 | Lower Kuskokwim River | 1989 | 77 | 60.81 | -161.45 | 60.20 | -159.90 | USFWS | Summer | Yes | Yes |
| Aniak | F1 | Middle Kuskokwim River | 1992 | 95 | 61.57 | -159.49 | 60.80 | -159.51 | ADF&G | Summer | Yes | Yes |
| Salmon | F2 | Middle Kuskokwim River | 2007 | 96 | 61.06 | -159.20 | 60.75 | -159.69 | USFWS | Summer | No | No |
| Holokuk | F3 | Middle Kuskokwim River | 2007 | 63 | 61.54 | -158.59 | 61.30 | -158.31 | USFWS | Summer | Yes | Yes |
| Oskawalik | F4 | Middle Kuskokwim River | 1994 | 58 | 61.75 | -158.18 | 61.56 | -157.79 | ADF&G | Summer | Yes | Yes |
| George | F5 | Middle Kuskokwim River | 2007 | 96 | 61.90 | -157.71 | 62.17 | -157.26 | USFWS | Summer | Yes | Yes |
| Kogrukuk | F6 | Middle Kuskokwim River | 2007 | 96 | 60.85 | -157.85 | 60.55 | -158.29 | USFWS | Summer | Yes | Yes |
| Stony | F7 | Middle Kuskokwim River | 1994 | 151 | 61.77 | -156.59 | 61.16 | -154.10 | ADF&G | Summer | Yes | Yes |
| Tatlawiksuk | F8 | Middle Kuskokwim River | 2007 | 96 | 61.92 | -156.24 | 62.16 | -155.53 | USFWS | Summer | No | No |
| Takotna | G1 | Upper Kuskokwim River | 2007 | 96 | 62.96 | -155.60 | 62.66 | -156.64 | USFWS | Summer | Yes | Yes |
| Big River | G2 | Upper Kuskokwim River | 2008 | 96 | 62.61 | -155.01 | 62.69 | -154.38 | ADF&G | Fall | Yes | No |
| South Fork | G3 | Upper Kuskokwim River | 2008 | 96 | 63.09 | -154.64 | 62.06 | -153.48 | ADF&G | Fall | Yes | No |
| Meshik | H1 | Bristol Bay | 1989 | 75 | 56.81 | -158.66 | 56.60 | -158.42 | USFWS | Summer | Yes | No |
| Big Creek | H2 | Bristol Bay | 1988/2000 | 96 | 58.29 | -157.53 | 58.18 | -155.76 | USFWS | Summer | Yes | No |
| Nushagak | H3 | Bristol Bay | 1988 | 75 | 58.80 | -158.63 | 60.11 | -156.99 | USFWS | Summer | Yes | Yes |
| Stuyahok | H4 | Bristol Bay | 1992/1993 | 87 | 60.19 | -156.29 | 60.18 | -156.15 | ADF&G | Summer | No | Yes |
| Mulchatna | H5 | Bristol Bay | 1994 | 95 | 59.95 | -156.41 | 60.61 | -154.42 | ADF&G | Summer | No | Yes |
| Togiak | H6 | Bristol Bay | 1993 | 95 | 59.08 | -160.34 | 59.19 | -160.38 | ADF&G | Summer | No | Yes |

Table 3.2. Summary statistics of the loci. ‘A’ refers to the number of alleles, ‘ H_e ’ and ‘ H_o ’ are the expected and observed heterozygosity, ‘ θ ’ is Weir and Cockerham’s (1984) F_{ST} , θ_{CSW} is the F_{ST} value when only data from the coastal southwestern populations are used. The θ value for the mitochondrial locus is ϕ_{ST} . Grey boxes indicate values for outlier loci. Some SNP loci are composed of multiple linked SNPs.

| Locus | Locus Type | A | H_e | H_o | θ | θ_{CSW} | D_{EST} | Source |
|------------|------------|----|-------|-------|----------|----------------|-----------|---------|
| Oki100 | mSAT | 23 | 0.902 | 0.871 | 0.012 | 0.001 | 0.095 | UAF/ABL |
| Omm1070 | mSAT | 40 | 0.961 | 0.955 | 0.004 | 0.000 | 0.099 | UAF/ABL |
| Omy1011 | mSAT | 30 | 0.924 | 0.922 | 0.009 | 0.002 | 0.092 | UAF/ABL |
| One101 | mSAT | 34 | 0.896 | 0.893 | 0.014 | 0.001 | 0.107 | UAF/ABL |
| One102 | mSAT | 21 | 0.910 | 0.886 | 0.004 | 0.000 | 0.041 | UAF/ABL |
| One104 | mSAT | 30 | 0.929 | 0.919 | 0.018 | 0.000 | 0.182 | UAF/ABL |
| One111std | mSAT | 93 | 0.924 | 0.912 | 0.016 | 0.000 | 0.161 | UAF/ABL |
| One114 | mSAT | 47 | 0.922 | 0.913 | 0.008 | 0.002 | 0.078 | UAF/ABL |
| Ots103 | mSAT | 43 | 0.947 | 0.934 | 0.008 | 0.000 | 0.125 | UAF/ABL |
| Ots3std | mSAT | 20 | 0.767 | 0.740 | 0.031 | 0.003 | 0.085 | UAF/ABL |
| Ots68 | mSAT | 40 | 0.944 | 0.925 | 0.010 | 0.002 | 0.138 | UAF/ABL |
| Ssa419 | mSAT | 21 | 0.862 | 0.855 | 0.012 | 0.000 | 0.066 | UAF/ABL |
| AHR178 | SNP | 2 | 0.491 | 0.496 | 0.011 | -0.002 | 0.011 | ADF&G |
| ARF | SNP | 2 | 0.308 | 0.289 | 0.022 | 0.009 | 0.010 | ADF&G |
| CCT3220 | SNP | 2 | 0.355 | 0.354 | 0.009 | 0.001 | 0.005 | ADF&G |
| CKS389 | SNP | 2 | 0.377 | 0.363 | 0.012 | 0.000 | 0.006 | ADF&G |
| CL | SNP | 2 | 0.402 | 0.405 | 0.005 | 0.002 | 0.003 | UAF/ABL |
| COPA | SNP | 2 | 0.057 | 0.058 | 0.014 | 0.005 | 0.001 | ADF&G |
| ctgf105 | SNP | 2 | 0.263 | 0.266 | 0.014 | 0.000 | 0.004 | ADF&G |
| CTS1627 | SNP | 2 | 0.484 | 0.490 | 0.003 | -0.003 | 0.003 | ADF&G |
| DM20 | SNP | 2 | 0.492 | 0.508 | 0.010 | -0.002 | 0.008 | ADF&G |
| EIF4EB | SNP | 2 | 0.082 | 0.082 | 0.019 | 0.006 | 0.002 | ADF&G |
| ER | SNP | 2 | 0.174 | 0.166 | 0.024 | 0.016 | 0.005 | UAF/ABL |
| FARSLA242 | SNP | 2 | 0.061 | 0.062 | 0.030 | 0.000 | 0.002 | ADF&G |
| GAPDH | SNP | 2 | 0.488 | 0.475 | 0.022 | 0.000 | 0.019 | ADF&G |
| GHII | SNP | 2 | 0.406 | 0.378 | 0.021 | -0.001 | 0.013 | ADF&G |
| GnRH527 | SNP | 2 | 0.358 | 0.361 | 0.012 | -0.001 | 0.006 | ADF&G |
| GPH105 | SNP | 2 | 0.461 | 0.425 | 0.031 | -0.002 | 0.024 | ADF&G |
| hnRNPL239 | SNP | 2 | 0.108 | 0.104 | 0.017 | -0.003 | 0.002 | ADF&G |
| HP182 | SNP | 2 | 0.365 | 0.350 | 0.010 | -0.002 | 0.005 | ADF&G |
| HSP90BA299 | SNP | 2 | 0.007 | 0.007 | 0.000 | -0.002 | 0.000 | ADF&G |
| IGF11 | SNP | 2 | 0.049 | 0.050 | 0.018 | 0.001 | 0.001 | ADF&G |
| IL8r272 | SNP | 2 | 0.183 | 0.179 | 0.013 | 0.000 | 0.003 | ADF&G |
| IN | SNP | 3 | 0.317 | 0.300 | 0.017 | 0.004 | 0.008 | UAF/ABL |
| IN1 | | | | | | | | |
| IN2 | | | | | | | | |
| IS | SNP | 4 | 0.564 | 0.555 | 0.017 | 0.006 | 0.020 | UAF/ABL |
| ISOII | | | | | | | | |
| ISOP | | | | | | | | |
| KPNA287 | SNP | 2 | 0.092 | 0.089 | 0.021 | 0.003 | 0.002 | ADF&G |
| MAPK1135 | SNP | 2 | 0.242 | 0.245 | 0.012 | 0.002 | 0.004 | ADF&G |

Table 3.2 Continued

| Locus | Locus Type | A | H _c | H _o | θ | θ_{CSW} | D _{EST} | Source |
|---------|------------|------|----------------|----------------|----------|-----------------------|------------------|-----------------|
| MARKS | SNP | 2 | 0.426 | 0.417 | 0.008 | 0.010 | 0.006 | ADF&G |
| MOESIN | SNP | 2 | 0.125 | 0.124 | 0.018 | 0.008 | 0.003 | ADF&G |
| PER | SNP | 2 | 0.080 | 0.078 | 0.029 | 0.024 | 0.003 | UAF/ABL |
| PL | SNP | 2 | 0.138 | 0.138 | 0.030 | 0.039 | 0.005 | UAF/ABL |
| RACP | SNP | 2 | 0.399 | 0.401 | 0.012 | -0.003 | 0.007 | ADF&G |
| RAS1 | SNP | 2 | 0.422 | 0.403 | 0.027 | -0.004 | 0.021 | ADF&G |
| RF | SNP | 2 | 0.450 | 0.437 | 0.036 | 0.004 | 0.028 | ADF&G |
| RH | SNP | 2 | 0.058 | 0.057 | 0.004 | 0.003 | 0.000 | UAF/ABL |
| SP | SNP | 2 | 0.499 | 0.517 | 0.003 | -0.003 | 0.003 | UAF/ABL |
| TCP178 | SNP | 2 | 0.127 | 0.122 | 0.020 | 0.005 | 0.003 | ADF&G |
| TF278 | SNP | 2 | 0.3511 | 0.337 | 0.054 | -0.002 | 0.027 | ADF&G |
| TSHA1 | SNP | 2 | 0.259 | 0.237 | 0.012 | 0.004 | 0.005 | ADF&G |
| u1519 | SNP | 2 | 0.265 | 0.257 | 0.020 | 0.002 | 0.007 | ADF&G |
| U200 | SNP | 2 | 0.486 | 0.474 | 0.014 | 0.000 | 0.013 | ADF&G |
| U202 | SNP | 2 | 0.124 | 0.122 | 0.042 | 0.000 | 0.005 | ADF&G |
| U212 | SNP | 2 | 0.059 | 0.055 | 0.028 | 0.010 | 0.002 | ADF&G |
| U216 | SNP | 2 | 0.245 | 0.239 | 0.006 | -0.002 | 0.002 | ADF&G |
| U217 | SNP | 2 | 0.488 | 0.501 | 0.017 | 0.002 | 0.015 | ADF&G |
| U302195 | SNP | 2 | 0.412 | 0.424 | 0.040 | 0.012 | 0.026 | ADF&G |
| U502241 | SNP | 2 | 0.212 | 0.216 | 0.014 | -0.002 | 0.003 | ADF&G |
| U503272 | SNP | 2 | 0.144 | 0.142 | 0.003 | 0.001 | 0.000 | ADF&G |
| U504 | SNP | 2 | 0.498 | 0.499 | 0.003 | -0.001 | 0.004 | ADF&G |
| U505112 | SNP | 2 | 0.438 | 0.441 | 0.007 | 0.001 | 0.005 | ADF&G |
| U506110 | SNP | 2 | 0.205 | 0.195 | 0.028 | 0.001 | 0.007 | ADF&G |
| U507286 | SNP | 2 | 0.497 | 0.503 | 0.006 | -0.002 | 0.006 | ADF&G |
| U509219 | SNP | 2 | 0.500 | 0.510 | 0.003 | -0.001 | 0.004 | ADF&G |
| U510204 | SNP | 2 | 0.304 | 0.301 | 0.019 | 0.006 | 0.009 | ADF&G |
| U511271 | SNP | 2 | 0.156 | 0.145 | 0.007 | 0.002 | 0.001 | ADF&G |
| U514150 | SNP | 2 | 0.251 | 0.240 | 0.018 | 0.001 | 0.005 | ADF&G |
| VR | SNP | 4 | 0.708 | 0.683 | 0.022 | 0.023 | 0.050 | UAF/ABL |
| VR1 | | | | | | | | |
| VR2 | | | | | | | | |
| VR3 | | | | | | | | |
| VT | SNP | 2 | 0.472 | 0.461 | 0.014 | -0.004 | 0.011 | UAF/ABL |
| ZAN132 | SNP | 2 | 0.449 | 0.451 | 0.010 | -0.003 | 0.008 | ADF&G |
| MT | SNP | 10 | 0.250 | N/A | 0.033 | 0.020 | N/A | UAF/ABL |
| MT5 | SNP | | | | | | | UAF/ABL |
| MT12 | SNP | | | | | | | UAF/ABL |
| MT18 | SNP | | | | | | | UAF/ABL |
| MT21 | SNP | | | | | | | UAF/ABL |
| MT27 | SNP | | | | | | | UAF/ABL |
| CR30 | SNP | | | | | | | Sato et al 2004 |
| CR231 | SNP | | | | | | | Sato et al 2004 |
| CR386 | SNP | | | | | | | Sato et al 2004 |
| Overall | mSAT | 36.8 | 0.907 | 0.894 | 0.012 | 0.001 | 0.090 | |
| Overall | SNP | 2.2 | 0.305 | 0.320 | 0.016 | 0.001 | 0.003 | |
| Overall | SNP & mSat | 8.2 | 0.408 | 0.481 | 0.015 | 0.001 | 0.009 | |

Table 3.3. Log-likelihood ratio (G) tests for the [25P70L data set].

| Regional Group | G | df | p-value |
|-----------------------|-----------------------|-----------|-----------------------------|
| Kotzebue | 1358.5 | 358 | $< 10^{-6}$ |
| Norton Sound | 657.6 | 358 | $< 10^{-6}$ |
| Lower Kuskokwim | 239.9 | 179 | 1.3×10^{-3} |
| Middle Yukon | 437.2 | 179 | $< 10^{-6}$ |
| Upper Kuskokwim | 1349.0 | 358 | $< 10^{-6}$ |
| Middle Kuskokwim | 391.8 | 358 | 1.1×10^{-1} |
| Lower Kuskokwim | 1030.5 | 895 | 1.1×10^{-3} |
| Bristol Bay | 969.6 | 358 | $< 10^{-6}$ |
| Total within | 6434.1 | 3043 | $< 10^{-6}$ |
| Total among | 6694.3 | 1253 | $< 10^{-6}$ |
| Total | 13128.4 | 4296 | 0.00E+00 |

$$F_{df_{among} df_{within}} = \frac{\frac{G_{among}}{df_{among}}}{\frac{G_{within}}{df_{within}}}$$

$$F = \frac{\frac{6694.3}{1253}}{\frac{6434.1}{3043}} \ll 10^{-6}$$

Table 3.4. Log-likelihood ratio (G) tests for the [21P50L data set].

| Regional Group | G | df | p-value |
|-----------------------|-----------------------|-----------|-----------------------------|
| Yukon | 414.35 | 300 | 1.3×10^{-5} |
| Kuskokwim | 660.89 | 450 | $< 10^{-5}$ |
| Bristol Bay | 527.27 | 150 | $< 10^{-5}$ |
| Overall | 1602.51 | 900 | $< 10^{-5}$ |

Table 3.5. Dispersal distances of chum salmon. Values were estimated with the slope of the IBD analysis from the Kuskokwim River.

| N_E/N_C Ratio | 0.0125 | 0.025 | 0.050 | 0.100 | 0.250 | 0.500 | 0.75 |
|-----------------|--------|--------|--------|--------|---------|---------|---------|
| N_E | 57000 | 115000 | 230000 | 460000 | 1150000 | 2300000 | 3450000 |
| D_E (fish/km) | 29.2 | 58.4 | 116.9 | 233.7 | 584.3 | 1168.6 | 1752.9 |
| Dispersal (km) | 222.3 | 157.2 | 111.2 | 78.6 | 49.7 | 35.2 | 28.7 |

References

- Alley R.B. 2000. The Younger Dryas cold interval as viewed from central Greenland. *Quaternary Science Reviews* 19: 213-226.
- Appleton S., Wiles G., Howell W., Jarvis S., and Lawson D. 2010. Tree records of environmental change in Glacier Bay National Park and Preserver, Southeast Alaska and their relation to 18th century Tglingit migration. The Geological Society of America, Denver.
- Barclay D.J., Wiles G.C., and Calkin P.E. 2009. Holocene glacier fluctuations in Alaska. *Quaternary Science Reviews* 28(21-22): 2034-2048.
- Beacham T.D., Candy J.R., Le K.D., and Wetklo M. 2009a. Population structure of chum salmon (*Oncorhynchus keta*) across the Pacific Rim, determined from microsatellite analysis. *Fisheries Bulletin* 107: 244-260.
- Beacham T.D., Candy J.R., and Wallace C. 2009b. Microsatellite stock identification of chum salmon on a Pacific rim basis. *North American Journal of Fisheries Management* 29: 1757-1776.
- Beacham T.D., and Murray C.B. 1990. Temperature, egg size, and development of embryos and alevins of five species of Pacific salmon: a comparative analysis. *Transactions of the American Fisheries Society* 119: 927-945.
- Beamish R.J., and Bouillon D.R. 1993. Pacific salmon production trends in relation to climate. *Canadian Journal of Fisheries and Aquatic Sciences* 50: 1002-1016.

- Beaumont M., and Nichols R. 1996. Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Academy of Sciences of London. Series B, Biological Sciences* 263: 1619-1626.
- Bockstoce J. 1973. A prehistoric population change in the Bering Strait region. *Polar Record* 16(105): 793-803.
- Bockstoce J. 1979. *The Archaeology of Cape Nome, Alaska*, p. 135. University of Pennsylvania, Philadelphia.
- Brabets T.P., Wang B., and Meade R.H. 2000. Environmental and hydrologic overview of the Yukon River basin, Alaska and Canada. U.S. Geological Survey, Anchorage. Water-Resources Investigations Report 99-4204
<http://pubs.usgs.gov/wri/wri994204/pdf/wri994204.pdf>
- Bue B.G., Molyneaux D.B., Schaberg K.L. 2007. Kuskokwim River chum salmon run reconstruction Fishery Data Series No. 08-64, pp. 1-38. Alaska Department of Fish and Game, Anchorage.
- Burkhart G., Dunmall K. 2005. The Snake River, Eldorado River and Pilgrim River salmon escapement enumeration and sampling project summary report, 2005 Kawerak Inc, pp. 1-55. Alaska Department of Fish and Game, Anchorage.
<http://www.kawerak.org/servicedivisions/nrd/fish/forms/2004/2004%20Snake%20Eld%20Pil%20Project%20Summary.pdf>
- Cavalli-Sforza L.L., Edwards A.W.F. 1967. Phylogenetic analysis: models and estimation procedures. *American Journal of Human Genetics* 19, 233-257.

- Chipman M.L., Clarke G.H., Clegg B.F., Gregory-Eaves I., and Hu F.S. 2008. A 2000 year record of climatic change at Ongoke Lake, southwest Alaska. *Journal of Paleolimnology* 41(1): 57-75.
- Clague J.J., Luckman B.H., Van Dorp R.D., *et al.* 2006. Rapid changes in the level of Kluane Lake in Yukon Territory over the last millennium. *Quaternary Research* 66(2): 342-355.
- Creager J.S., and McManus D.A. 1967. Geology of the Floor of the Bering and Chuckchi Seas - American Studies. In *The Bering Land Bridge*. Edited by Hopkins D.M. Stanford University Press, Stanford.
- Crow J.F., Aoki K. 1984. Group selection for a polygenic behavioral trait: Estimating the degree of population subdivision. *Proceedings of the National Academy of Sciences* 81: 6073-6077.
- Dougan J. 2010 Waiver for Yukon-Kuskokwim Portage. United States Department of the Interior, Anchorage.
http://www.blm.gov/pgdata/etc/medialib/blm/ak/aktest/rdi.Par.38638.File.dat/Y-K_Portage_Decision_09-02-2010.pdf
- Dumond D.E. 1998. Maritime adaptation on the northern Alaska peninsula. *Arctic Anthropology* 35(1): 187-203.
- Dupré W.R., 1988. Yukon Delta coastal processes study. National Oceanic and Atmospheric Administration, Outer Continental Shelf Environmental Assessment Program 1988, Final Report 58. pp. 393-447.

- Excoffier L., and Lischer J.E.L. 2010. Arlequin suite ver 3.5: A new series of programs to perform genetic analyses under Linux and Windows. *Molecular Ecology Resources* 10: 564-567.
- Excoffier L., Smouse P.E., and Quattro J.M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131: 479-491.
- Fairbanks R.G. 1989. A 17,000-year glacio-eustatic sea level record: influence of glacial melting rates on the Younger Dryas event and deep-ocean circulation. *Nature* 342(6250): 637-642.
- Falconer D.S., and Mackay T.F.C. 1996. *Introduction to Quantitative Genetics*. Longman Group Ltd Essex.
- Felsenstein J. 2004. PHYLIP (Phylogeny Inference Package) version 3.9. Distributed by the author. Department of Genetics, University of Washington, Seattle..
- Finney B.P., Gregory-Eaves I., Douglas M.S.V., and Smol J.P. 2002. Fisheries productivity in the northeastern Pacific Ocean over the past 2,200 years. *Nature* 416: 729-733.
- Foll M., and Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180: 977-993.
- Fournier D.A., Beacham T.D., Riddell B.E., and Busack C.A. 1984. Estimating stock composition in mixed stock fisheries using morphometric, meristic, and

- electrophoretic characteristics. Canadian Journal of Fisheries and Aquatic Sciences 41(3): 400-408.
- Garvin M.R., and Gharrett A.J. 2007. DEco-TILLING: An inexpensive method for SNP discovery that reduces ascertainment bias. Molecular Ecology Notes 7: 735-746.
- Garvin M.R., and Gharrett A.J. 2010. Application of SNP markers to chum salmon (*Oncorhynchus keta*): Discovery, genotyping, and linkage phase resolution. Journal of Fish Biology 77(9): 2137-2162.
- Garvin M.R., Saitoh K., Brykov V., Churikov D., and Gharrett A.J. 2010. Single nucleotide polymorphisms in chum salmon (*Oncorhynchus keta*) mitochondrial DNA derived from restriction site haplotype information. Genome 53: 501-507.
- Gharrett A.J., Lane S., McGregor A.J., and Taylor S.G. 2001. Use of a genetic marker to examine genetic interaction among subpopulations of pink salmon (*Oncorhynchus gorbuscha*). Genetica 111: 259-267.
- Gharrett A.J., Riley R.J., and Spencer P.D. 2012. Genetic analysis reveals restricted dispersal of Northern rockfish along the continental margin of the Bering Sea and Aleutian Islands. Transactions of the American Fisheries Society 141(2): 370-382.
- Gilk S.E., Templin W.D., Molyneaux D.B., Hamazaki T., Pawluk J.A. 2009. Biological and genetic characteristics of fall and summer chum salmon (*Oncorhynchus keta*) in the Kuskokwim River, Alaska. American Fisheries Society Symposium 70: 161-179

- Gisclair B.R. 2009 Salmon bycatch management in the Bering Sea walleye pollock fishery: threats and opportunities for Western Alaska. American Fisheries Society Symposium 70:799–816.
- Groot C., and Margolis L. 1991. Pacific Salmon: Life Histories. University of Washington Press.
- Hawkins S.L., Varnavskaya N.V., Matzak E.A., et al. 2002. Population structure of odd-broodline Asian pink salmon and its contrast to the even-broodline structure. Journal of Fish Biology 60, 370-388.
- Hoffecker J.F., and Elias S.A. 2003. Environment and archeology in Beringia. Evolutionary Anthropology: Issues, News, and Reviews 12(1): 34-49.
- Hopkins D.M. 1967. The Cenozoic History of Beringia - A Synthesis. In The Bering Land Bridge. Edited by Hopkins DM. Stanford University Press, Stanford.
- Huson D.H., Richter D.C., Rausch C., et al. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8(1): 460.
- Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24: 1403-1405.
- Kinnison M.T., Bentzen P., Unwin M.J., and Quinn T.P. 2002. Reconstructing recent divergence: evaluating nonequilibrium population structure in New Zealand Chinook salmon. Molecular Ecology 11: 739-754.
- Kondzela C.M. 2010. The colonization mechanism of pink salmon populations in Glacier Bay, Alaska, based on genetic data. PhD thesis. University of Alaska Fairbanks.

- Kondzela C.M., Garvin M.R., Wilmot R.L., *et al.* In Preparation. Population genetic structure of chum salmon (*Oncorhynchus keta*) across the Pacific Rim based on single-nucleotide-polymorphisms and microsatellites.
- Kruse G.H. 1998. Salmon run failures in 1997-1998: A link to anomalous ocean conditions? *Alaska Fisheries Research Bulletin* 5(1): 55-63.
- Levy L.B., Kaufman D.S., and Werner A. 2004. Holocene glacier fluctuations, Waskey Lake, northeastern Ahklun Mountains, southwestern Alaska. *The Holocene* 14(2): 185-193.
- Lewis P.O., and Zaykin D. 2001. Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c).
<http://lewis.eeb.uconn.edu/lewishome/software.html>.
- Linderman J.C., Bergstrom D.J. 2009. Kuskokwim management area: salmon escapement, harvest, and management. *American Fisheries Society Symposium* 70: 541-599.
- Maddren A.G. 1910. The Innoko Gold-Placer district Alaska with accounts of the central Kuskokwim valley and the Ruby Creek and Gold Hill placers (U.S. Geological Survey). Government Printing Office, Washington D. C.
- Manely W.F. 2002. Postglacial flooding of the Bering Land Bridge: A geospatial animation. INSTAAR, University of Colorado.
http://instaar.colorado.edu/QGISL/bering_land_bridge.
- Mann D.H., and Hamilton T.D. 1995. Late Pleistocene and Holocene paleoenvironments of the north Pacific coast. *Quaternary Science Reviews* 14: 449-471.

- Mann M.E. 2002. Little Ice Age. In Encyclopedia of Global Environmental Change. Edited by Munn T. John Wiley & Sons. pp. 504-509.
- McDonald J.H. 2009. Handbook of Biological Statistics. Sparky House Publishing.
- McPhail J., and Lindsey C. 1986. Zoogeography of the freshwater fishes of Cascadia (the Columbia system and rivers north to the Stikine). In The Zoogeography of North American Freshwater Fishes. Edited by Hocutt C, Wiley E. John Wiley and Sons, New York. pp. 615-637.
- Mills R.O. 1994. Radiocarbon calibration of archaeological dates from the central Gulf of Alaska. Arctic Anthropology 31(1): 126-149.
- Narum S.R., and Hess J.E. 2011. Comparison of F_{ST} outlier tests for SNP loci under selection. Molecular Ecology Resources 11: 184-194.
- Nelson H., and Creager J.S. 1977. Displacement of Yukon-derived sediment from Bering Sea to Chukchi Sea during Holocene time. Geology 5(3): 141-146.
- Olsen J.B., Crane P.A., Flannery B.G., *et al.* 2010. Comparative landscape genetic analysis of three Pacific salmon species from subarctic North America. Conservation Genetics 12(1): 223-241.
- Olsen J.B., Beacham T.D., Le K.D., *et al.* 2006. 2005 Arctic Yukon Kuskokwim Sustainable Salmon Initiative Final Report. Genetic variation in Norton Sound chum salmon populations. United States Fish and Wildlife Service, Anchorage.
- Quinn T.P. 2005. The Behavior and Ecology of Pacific Salmon and Trout. University of Washington Press, Seattle.

- Rousset F. 2008. GENEPOP'007: a complete reimplementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* 8(1): 103-106.
- Ryman N., Palm S., Andre C., et al. 2006. Power for detecting genetic divergence: differences between statistical methods and marker loci. *Molecular Ecology* 15, 2031-2045.
- Sancetta C. 1983. Effect of Pleistocene glaciation upon oceanographic characteristics of the north Pacific Ocean and Bering Sea. *Deep-Sea Research* 30(8A): 851-869.
- Seeb L.W., and Crane P.A. 1999. Genetic heterogeneity in chum salmon in western Alaska, the contact zone between northern and southern lineages. *Transactions of the American Fisheries Society* 128: 58-87.
- Seeb L.W., Crane P.A., Kondzela C.M., et al. 2004. Migration of Pacific Rim chum salmon on the high seas: insights from genetic data *Environmental Biology of Fishes* 69: 21-36.
- Seeb L.W., Templin W.D., Sato S., et al. 2011. Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. *Molecular Ecology Resources* 11: 195-217.
- Shaw R.D. 1998. An archaeology of the central yupik: a regional overview for the Yukon-Kuskokwim Delta, northern Bristol Bay and Nunivak Island. *Arctic Anthropology* 35(1): 234-246.
- Shepard F.P., and Wanless H.R. 1971. *Our Changing Coastlines*. McGraw-Hill, New York.

- Sokal R.R., and Rohlf F.J. 1994. *Biometry: The principles and practices of statistics in biological research*. W. H. Freeman.
- Templin W.D., Seeb J.E., Jasper J.R., Barclay A.W., and Seeb L.W. 2011. Genetic differentiation of Alaska Chinook salmon: the missing link for migratory studies. *Molecular Ecology Resources* 11: 226-246.
- Utter F.M., McPhee M.V., and Allendorf F.W. 2009. Populations genetics and management of Arctic-Yukon-Kuskokwim salmon populations. *American Fisheries Society Symposium* 70: 97-123.
- Waples R.S. 2002. Effective Size of fluctuating salmon populations. *Genetics* 161: 783-791.
- Waples R.S., Pess G.R., and Beechie T. 2008. Evolutionary history of Pacific salmon in dynamic environments. *Evolutionary Applications* 1: 189-206.
- Weir B., and Cockerham C. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution* 38: 1358-1360.
- Whited D.C., Kimball J.S., Lucotch J.A., *et al.* 2012. A riverscape analysis tool developed to assist wild salmon conservation across the north Pacific rim. *Fisheries* 37(7): 305-314.
- Wiles G.C., Barclay D.J., Calkin P.E., and Lowell T.V. 2008. Century to millennial-scale temperature variations for the last two thousand years indicated from glacial geologic records of Southern Alaska. *Global Planetary Change* 60(1-2): 115-125.

- Wilmot R.L., Everett R.J., Spearman W.J., *et al.* 1994. Genetic stock structure of western Alaskan chum salmon and a comparison with Russian far east stocks. *Canadian Journal of Fisheries and Aquatic Sciences* 51(Suppl. 1): 84-94.
- Wolfe R.J., Spaeder J. 2009. People and salmon of the Yukon and Kuskokwim drainages and Norton Sound Alaska: Fishery harvests, culture change, and local knowledge systems. *American Fisheries Society Symposium* 70: 349-379.
- Yesner D.R. 1998. Origin and development of maritime adaptations in the northwest Pacific region of North America: a zooarchaeological perspective. *Arctic Anthropology* 35(1): 204-222.

General Conclusions

Most of this thesis involves some of the latest technology to address questions regarding the conservation and management of chum salmon. The molecular genetic and statistical computer algorithms that I developed and/or used here will almost certainly be obsolete before I've settled into my next job. But I think that the questions that I answered and the problems that I addressed will provide firm footholds to explore hypotheses that were generated here.

Molecular genetics has been a rapidly evolving field that is now maturing. Adopting technology developed for human genetic studies to solve problems in other species has oftentimes been analogous to putting a square peg in a round hole because human-based studies were largely focused on one or a few individuals, but that is changing. The focus is now to study genomes of many individuals, which is ideal for population genetics, conservation, and management studies of populations. The first two chapters of this work are primarily a statement that we need to understand the limitations of the tools that we use or develop. The topic may seem mundane, but the conclusions that we draw are only as good as the data we produce.

The last chapter may on the surface appear to address very different questions, but in actuality, draws on what was just stated above: what are the limitations of the tools we develop? This chapter sought to understand the basis for the weak genetic structure of populations in western Alaska. Although many published works and anecdotal accounts have highlighted this fact, none have explored it in detail. My co-authors and I have provided a reasonable explanation for the divergence, but I believe that a more complete

answer will develop if the next studies follow what we have done, and bring together other broad disciplines. A combination of geological, archaeological, oceanographic, and molecular genetic methods will likely provide a clearer picture and results from the fields individually will probably help to explain the combined distribution of many other species in this broad geographic area. Such knowledge will be of paramount importance as management and conservation efforts focus on monitoring species' response to a rapidly changing climate that we can no longer control.

General References

- Anderson E (2010) Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology* 10, 701-710.
- Anderson E, Waples RS, Kalinowski ST (2008) An improved method for estimating the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences* 65, 1475-1486.
- Augerot X (2005) *Atlas of Pacific Salmon*. University California Press, Berkeley.
- Beacham TD, Candy JR, and Wallace C (2009). Microsatellite stock identification of chum salmon on a Pacific rim basis. *North American Journal of Fisheries Management* 29, 1757-1776.
- Beamish RJ, Bouillon DR (1993) Pacific salmon production trends in relation to climate. *Canadian Journal of Fisheries and Aquatic Sciences* 50, 1002-1016.
- Eggers DM (2009) Historical biomass of pink, chum, and sockeye salmon in the North Pacific Ocean. *American Fisheries Society Symposium* 70, 267-305.
- Garvin MR, Bielwaski JP, Gharrett AJ (2011) Positive Darwinian selection in the piston that powers proton pumps in complex I of the mitochondria of Pacific Salmon. *PLoS One* 6, e24127.
- Garvin MR, Saitoh K, Brykov V, Churikov D, Gharrett AJ (2010) Single nucleotide polymorphisms in chum salmon (*Oncorhynchus keta*) mitochondrial DNA derived from restriction site haplotype information. *Genome* 53, 501-507.

- Gisclair BR (2009) Salmon bycatch management in the Bering Sea walleye pollock fishery: threats and opportunities for Western Alaska. *American Fisheries Society Symposium* 70, 799–816.
- Hancock JM (1999) Microsatellites and other simple sequences: genomic context and mutational mechanisms. In: *Microsatellites Evolution and Applications* (eds. Goldstein DB, Schlotterer C) Oxford University Press, New York.
- Knapp G (2007) Implications of Aquaculture for Wild Fisheries: The Case of Alaska Wild Salmon. In Richard Arthur and Jochen Nierentz, *Global Trade Conference on Aquaculture, 29–31 May 2007, Qingdao, China*. *FAO Fisheries Proceedings* 9. Rome, FAO. 2007
- Kruse GH (1998) Salmon run failures in 1997-1998: A link to anomalous ocean conditions? *Alaska Fisheries Research Bulletin* 5, 55-63.
- Ricker WE (1954) Stock and recruitment. *Journal of the Fisheries Research Board of Canada* 11, 559-623.
- Seeb LW, Crane PA, Kondzela CM, et al. (2004) Migration of Pacific Rim chum salmon on the high seas: insights from genetic data *Environmental Biology of Fishes* 69, 21-36.
- Wolfe RJ, Spaeder J (2009) People and salmon of the Yukon and Kuskokwim drainages and Norton Sound Alaska: Fishery harvests, culture change, and local knowledge systems. *American Fisheries Society Symposium* 70, 349-379.
- Zimmerman C, See M, Volk E, et al. (2006) Arctic-Yukon-Kuskokwim Salmon Research & Restoration Plan. *American Fisheries Society Symposium* 70, 3-10