NOVEL APPLICATIONS OF RANDOM FOREST FOR EXPLORING POPULATION STRUCTURE OF ATLANTIC SALMON (*SALMO SALAR*) IN LABRADOR, CANADA

by

Emma V. A. Sylvester

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2017

Table of Contents

List of Tables

List of Figures

Abstract

The detection of population-genetic structure is useful for understanding patterns of gene flow, population distribution, and wildlife management and conservation. In this work, we examine approaches for inferring the modern genetic structure of Atlantic salmon (*Salmo salar*). We explore the utility of machine-learning algorithms (random forest, regularized random forest, and guided regularized random forest) compared with F_{ST}-ranking for selection of single nucleotide polymorphisms (SNP) for fine-scale population assignment within a marine embayment, Lake Melville, Labrador. Using an unpublished SNP dataset for Atlantic salmon and validating our approaches with a published SNP data set for Alaskan Chinook salmon (*Oncorhynchus tshawytscha*), we demonstrate improved self-assignment accuracy and provide evidence of population structure consistent with F-statistics. We compare the level of population structure in greater Labrador that is resolved using a preliminary panel of SNPs selected with guided regularized random forest with an established panel of 101 microsatellites. We ask if salmon originating from rivers draining into Lake Melville show evidence of discrete genetic population structure relative to those outside of the embayment. Finally, we investigate environmental parameters associated with the observed genetic structure and seek to explain the mechanisms driving genetic differentiation in the area. We highlight the potential for applications of machine-learning approaches in population genetics and uncover fine-scale structure with potential impact on fisheries management techniques.

List of Abbreviations Used

COSEWIC Committee on the Status of Endangered Wildlife in Canada

DAPC Discriminant analysis of principal components

DU Designatable Unit

EAA Environmental association analysis

GRRF Guided regularized random forest

GSI Genetic stock identification

HWE Hardy-Weinberg Equilibrium

IBD Isolation by distance

LD Linear discriminants

LOO Leave-one-out

MAF Minor allele frequency

MDA Mean decrease in accuracy

NJ Neighbour-joining

OOB Out-of-bag

PC(A) Principal component analysis

RDA Redundancy analysis

RF Random forest

RRF Regularized random forest

RSE Residual squared error

SARA Species at Risk Act

SFA Salmon fishing area

SNP Single nucleotide polymorphism

Acknowledgements

Thank you to all the members of the Beiko, Bentzen, Bradbury, and Clément labs for continuous suggestions and support, particularly (and in no particular order) Alex Keddy, Michael Hall, John Horne, Ian Paterson, Beth Watson, Ryan Stanley, Nick Jeffery, Sarah Lehnert, and Judith Savoie. Thanks, as well to the Nunatsiavut Government, the Sivunivut Inuit Community Corporation, the Innu Nation, the Labrador Hunting and Fishing Association and local fishers for their support and active participation in the salmon project. This work was funded by an NSERC (Natural Sciences and Engineering Research Council of Canada) Strategic Grant to Paul Bentzen, Robert Beiko and Marie Clément, and the Canada Graduate Scholarship (CGS-M), as well as NSGS (Nova Scotia Graduate Scholarship), Atlantic Salmon Conservation Foundation, Labrador Institute (Atlantic Canada Opportunities Agency and Department of Business, Tourism, Culture and Rural Development) and Olin Fellowships (Atlantic Salmon Federation). Lastly, thank you to my family for their continued support, my love and co-fish-lover, Greg, for his patience and encouragement, Saturday-morning yard sales, and Lindy Hop for getting me through the last couple years.

Chapter 1

Introduction

1.1 Population Genetics in Fisheries Research

The study of population genetics (or genomics) for wildlife conservation and management is, like many sciences, a diverse and expanding field. Not only do new, interesting questions arise out of existing work, as technology develops we have increasing access to a variety of types and immense quantities of genetic data, and with that, the opportunity for novel approaches to uncover interesting patterns. Ultimately, population genetics asks “what diversity is present in this population?” or more accurately, “what is the genetic structure of individuals and populations in a given area?” For the purposes of conservation, particularly of exploited species, we aim to identify this structure to retain genetic diversity to reduce inbreeding depression in often declining wildlife populations, and to preserve adaptive genetic variation.

Applying these objectives to mixed-stock fisheries management involves identifying the stock (population) composition of harvests, and ensuring subpopulations are not overexploited (Bradbury et al., 2015). This approach, termed genetic stock identification or GSI can be done through mixed-stock analysis or assignment of individuals to a reference population (McKinney et al., 2017; Anderson 2010; Guinand et al., 2002). Although a reference is often unavailable, it is possible to infer a population of origin for highly structured populations (e.g. those that demonstrate natal homing behaviour) by sampling individuals at sites of putative population origin and establishing a model of population structure to then assign individuals of unknown origin. Though modelling attempts to detect patterns, it can also be implemented for data reduction, in which a reduced set of genetic markers, or loci, are selected to adequately explain variation in the data (Figure 1.1C). Often, various approaches for a given step in the GSI workflow (Figure 1.1) are implemented for a cross-method comparison.

Rapid advances in sequencing and genotyping technologies, such as next-generation sequencing or NGS, have enabled the development of large panels of informative single nucleotide polymorphisms (SNPs) from genome-wide scans. These substitution mutations are usually biallelic, as multiple substitution events occurring at the same nucleotide are highly unlikely. Rare SNPs that are polyallelic are often removed from the data due to downstream computational complexity. Relative allele frequency of SNPs across populations can be used to inform genetic structure. Markers selected particularly for maximum self-assignment accuracy are likely to be useful for assignment across both broad and small-scale studies (Larson et al., 2014a); however, the trade-off between panel size and self-assignment accuracy often results in panels that, at an adequate performance threshold, are too large to be of practical value for fisheries applications, due to the costs of analysis. For this reason, methods to select informative loci without reducing accuracy of the assignment model, known generally as feature-selection, have been developed (Helyar et al., 2011; Rosenberg et al. 2005). To date, methods to select informative loci have mostly been restricted to simple, univariate ranking. Machine-learning methods may provide a more robust approach, allowing for the establishment of smaller panels with greater accuracy (Topchy et al., 2004; Guinand et al., 2002), by using computational algorithms to recognize underlying patterns within data. An iterative process, these methods apply non-parametric methods to subsets of variables or features (here, genetic markers) to find trends between sets of data, or within the data itself (Guinand et al., 2002). This allows for greater consideration of feature importance by evaluating how variables interact and influence the model in a variety of combinations. In classification, states of features across samples (individuals) of a known class or group are assessed to place individuals into their assigned class with maximum accuracy. Feature-selection identifies key features that convey a high degree of information for the classification process.

To avoid overestimating the accuracy of a given method, often referred to as upward grading bias (Anderson, 2010) the feature-selection algorithm is applied to a subset of individuals with known origin, referred to as a training set (Figure 1.1C). The remaining individuals, or those in the test set, are assigned to a population using an

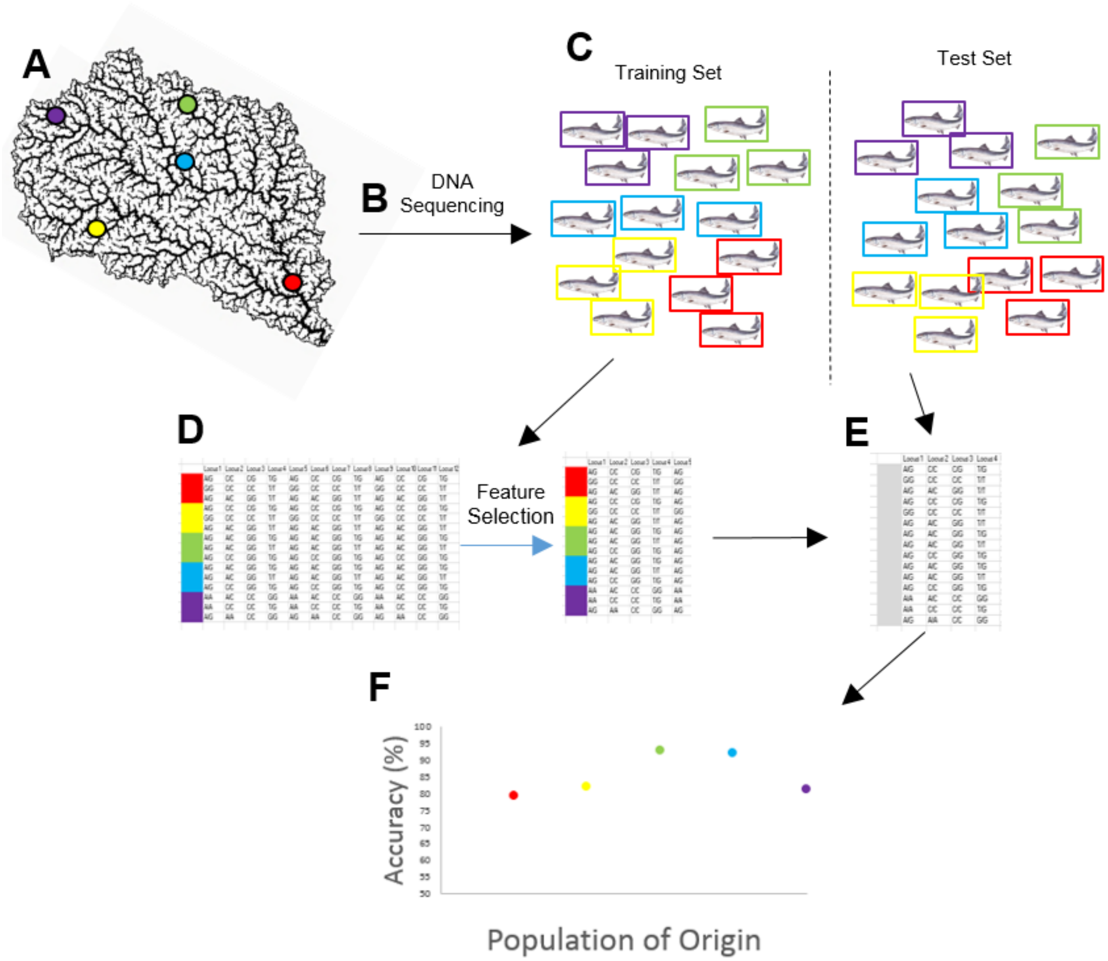

Figure 1.1: A generalized workflow to assess a method for feature-selection for population assignment of individuals. A) Tissue is sampled from individuals at several sampling/spawning sites (in this case, within a river network) indicative of population of origin and putative population structure. B) DNA is sequenced from tissues to create a panel of all loci within all individuals. C) Individuals are split into a training and test set. D) A feature/loci selection technique is implemented on genotype data from 'training' individuals. E) 'Test' individuals are assigned to a population with GSI methods using only selected loci from step D. F) Accuracy of the locus-selection method (step D) is assessed by determining the proportion of correctly assigned individuals.

algorithm that is given only the selected loci from the feature-selection step (Figure 1.1E). The success of the feature-selection technique can be assessed by calculating the proportion of individuals correctly assigned given their putative population of origin, that is, the proportion of individuals from a given sampling location whose assignment (Figure 1.1E) matches that of the sampling site.

In Chapter Two, we investigate novel approaches to select informative loci for population assignment of Atlantic salmon (*Salmo salar*) in a small geographic area of Labrador, Canada.

1.2 Landscape Genetics in Fisheries Research

While it is helpful to understand the genetic population structure of a fishery species, it is also important to identify possible mechanisms influencing this structure. Landscape genetics not only seeks to determine the genetic diversity present in a population, but also how the landscape affects the distribution of this diversity. Genetic divergence, the accumulation of genetic differences (e.g. allele frequencies) across groups of individuals resulting in defined subpopulation structure and potentially speciation, can result from many phenomena. Divergence may be influenced by selection, or by random (neutral) processes. Neutral factors include genetic drift, in which small populations undergo random changes in relative allele frequency, causing them to diverge, as well as dispersal or gene flow, in which migrants may introduce new genetic variants or alter the genetic composition of a population. This often occurs as localized gene flow, in which individuals are exchanged between neighbouring populations, or from historical scenarios such as secondary contact between formerly isolated populations. Alternatively, non-neutral, or selective forces influence genetic divergence through local adaptation (Manel and Holderegger, 2013; Manel et al., 2003). By investigating how physical or environmental variables at a given spawning site vary with genetic structure, landscape genetics aims to identify possible underlying mechanisms to explain the observed genetic diversity.

In addition to asking interesting questions from an evolutionary perspective, landscape genetics research plays a role in policy making to conserve exploited species (Manel and Holdregger, 2013). To be considered an evolutionarily significant unit, or a designatable unit (DU), for conservation purposes by the Committee on the Status of Endangered Wildlife in Canada (COSEWIC, 2015), a population must not only be genetically discrete, but also evolutionarily significant. This significance is based on four main criteria, although to be a considered a DU a population does not need to exhibit all four. There may exist (1) evidence of local adaption, or genetic divergence due to selection. A population may also demonstrate (2) endemism, in which a genetically distinct group is isolated to a particular area, or (3) deep phylogenetic divergence relative to other populations of the species. The population may also be important for (4) range connectivity, in which its removal or extinction would result in isolation of surrounding populations. As such, developing approaches to provide evidence of these phenomena is important to influence protection of species and resources. By identifying environmental variables influencing local adaptation of a population, conservation efforts may be directed accordingly to focus on key habitat features to promote the maintenance of genetic diversity in managed populations (Hilborn et al., 2003; Manel and Holdregger, 2013).

Environmental association analysis (EAA) refers to the general statistical approaches involved in identifying correlations between genetic and environmental data, based on the principle that local adaptation evolves due to environmental or ecological heterogeneity across a population range (Lotterhos and Whitlock, 2015). Disentangling neutral from adaptive influences on genetic variation is a difficult computational task as these factors often vary similarly with geography and are therefore often highly correlated, resulting in similar signals reflected within the data (Frichot et al., 2013). This necessitates determining an underlying ‘neutral model’ to describe the genetic differentiation expected with only neutral influence on the process of divergence. Loci that exhibit patterns of population structure, and correlations with environmental parameters outside of this model can then be considered for evidence of local adaptation. This can be accomplished by identifying outlier loci after accounting for neutral structure and removing putative neutral loci from the dataset (Ferchaud and Hansen, 2016), or by including neutral genetic structure as an explanatory variable within EAA, such that correlations outside of this variable can be identified (Bradbury et al., 2015; Bourret et al., 2014). This inference, however, is dependent on the type of genetic marker used for analysis.

1.2.1 Genetic Markers: SNPs and Microsatellites

Different types of molecular markers carry varying amounts of information, depending on the number of alleles per locus. They can also differ in their impact on trait or phenotypic expression, and undergo varying mutation rates (DeFaveri et al., 2013). Landscape associations with panels consisting of a particular marker type are likely to identify different factors influencing population structure compared to panels with alternative genetic information.

SNPs are widely present within the genome, can be genotyped with low error and do not require access to a species' fully sequenced genome. Because of these advantages, the use of SNPs for GSI have been increasing, relative to microsatellites (Hess et al., 2011; Hauser et al., 2011). Microsatellites are regions of repeating k-mers, often 2-4 nucleotides in length, in which allelic variation is defined by variation in the number of repeats. As repeat regions are more prone to mutations, microsatellites experience a higher mutation rate relative to SNPs (Nishant et al., 2009). Microsatellites contain more information per locus due to a high length variability across individuals, (Hauser et al., 2011); consequently, fewer (up to a tenth of the number of biallelic SNPs) are required to obtain equivalent detection of population structure (Hess et al., 2011). However, the need for more SNPs can be readily met with NGS techniques with relatively low cost. Both types of marker have been used for identification of population structure and landscape associations, though the use of large SNP panels is becoming more frequent throughout the literature (e.g. Bradbury et al., 2015a; Larson et al., 2014a; Bourret et al., 2013; Dionne et al., 2008). Although most studies implement a single marker type, harmonizing observed patterns in population structure using multiple molecular markers could be beneficial for understanding mechanisms of divergence and moving forward in population genetics research (Groot et al. 2015; DeFaveri et al., 2013). This integration, however, is not always straight-forward as comparisons of microsatellites and SNPs for population structure are often equivocal. Fine-scale structure has been found best resolved for threespine stickleback (*Gasterosteus aculeatus*) and Chinook salmon (*Oncorhynchus tshawytscha*) using microsatellites (DeFaveri et al., 2013; Hess et al., 2011), while assignment success was highest in sockeye salmon (*Oncorhynchus nerka*) and Atlantic salmon using SNPs (Hauser et al., 2011; Glover et al., 2010). Evidence of introgression in Newfoundland populations of Atlantic salmon was apparent only when analyzing SNP arrays (Bradbury et al., 2015b). Regardless of the molecular markers used, even when filtering SNPs for outlier detection, Moore et al. (2014) found similar resolution in North American Atlantic salmon population structure.

In the context of EAA for Atlantic salmon, microsatellite variation has most often been found to correlate with neutral factors that influence population size or carrying capacity such as river size or flow volume (Bradbury et al., 2014; Ozerov et al., 2012; Dillane et al., 2008). Due to their multi-allelism, signals of drift may be more easily detected when using microsatellites than SNPs. However, associations have also been made with putatively adaptive influences, such as temperature (Dionne et al., 2008) and water chemistry (Bradbury et al., 2014). Similarly, both neutral effects and adaptively associated parameters have been found to influence genetic differentiation of SNPs (Rougemont and Bernatchez, 2017; Zueva et al., 2014; Bourret et al., 2013; Palstra et al., 2007). Non-microsatellite-associated SNPs are more likely to be located within coding regions of the genome, microsatellites may be proximal to genes, resulting in genetic hitchhiking, in which the selection of a gene results in a similar change in allele frequency at a near-by locus due to physical proximity on the genome, without direct selection of the locus itself. SNPs and microsatellites alike are influenced by selective and neutral forces, necessitating approaches to identify true signals of selection.

1.3 Our Study System: Atlantic Salmon in Labrador

Historically a contentious location for French, British, and later American proprietary interests, Labrador fishery catch was monitored and extensively exploited across the region by the early 19th century (Taylor, 1985; Dunfield, 1985). Although fishery production regularly fluctuated, the progression of industry and human encroachment, as well as over-exploitation in the 20th century have recently led to significant declines in stocks, necessitating management strategies (Mills et al., 2013; Dunfield, 1985). Atlantic salmon in Newfoundland and Labrador are harvested according to Salmon Fishing Areas (SFAs), and assessed and monitored according to DUs (DFO, 2016). In fisheries, management units are often not aligned with biologically or genetically distinct populations (Reiss et al., 2009). Established with the initiation of a management plan in 1984, Labrador is composed of three SFAs, located North of Lake Melville, South of Lake Melville, and adjacent to the Strait of Belle Isle (O'Connell et al., 1992), while the whole of Labrador consists of a single DU. Despite the regular moratoria in the area, continued reductions in Atlantic salmon populations are largely attributed to reduced marine survival likely resulting from environmental and ecosystem changes, including those associated with climate change (DFO, 2016; Mills et al., 2013). Widely distributed and extensively exploited, Atlantic salmon exemplify opportunities and challenges within population and landscape genetics research (COSEWIC, 2011, Larson et al. 2014a; Bradbury et al., 2015a; 2015b; 2016). They are of particular conservation concern, protected under the Species at Risk Act (SARA) in parts of their ranges, due to substantial population declines. Anadromous populations home to their natal streams to spawn (natal philopatry) with low rates of straying (Hendry et al., 2004) and exhibit hierarchical population structure (Bradbury et al., 2015a; Bourret et al., 2013). As there is a greater differentiation between European and North American populations than within either continent, North American populations were thought to have derived from a single ancestral population already diverged from the European lineage (Ståhl, 1987). More recent evidence suggests at least one secondary contact event from a European lineage may have occurred in Newfoundland and southern Labrador due to the presence of shared alleles associated with European ancestry (Rougemont and Bernatchez, 2017; Bradbury et al., 2015b; Nilsson et al., 2001; King et al., 2000). How historical processes, such as recolonization, may have influenced current population structure and distribution may be influence modern findings in population genomics approaches. In Chapter Three we will further discuss the population structure of Labrador Atlantic salmon.

Within Labrador approximately 13,200 salmon are harvested each year (Bradbury et al., 2015a). FSC (Food, Social, and Ceremonial) fishery practices are conducted by Innu First Nations, Inuit (Nunasiavut) and Metis (NunatuKavut) groups and constitute important traditional and recreational harvests (ICES, 2013) necessitating a better understanding of stock assessment for management of these populations. Most of the salmon harvest occurs within Lake Melville, a 3,069 km^2 marine embayment (Figure

1.2). As a somewhat isolated area, physiographic characteristics such as temperature and water chemistry differ between Lake Melville and other areas of Labrador. As smolt spend some time in the embayment area before exiting into the broader ocean, these characteristics may play a selective role, influencing juvenile survivorship (Ozerov et al., 2012). If there is evidence of local adaptation affecting differentiation between Lake Melville and elsewhere in coastal Labrador, Lake Melville may require distinct management strategies, and may meet the criteria to be considered as a separate SFA.

1.4 Our Contribution

Fine-scale assignment of Atlantic salmon can be difficult due to relatively low divergence (Bradbury et al., 2015a), necessitating novel approaches to detect subtle genetic differences across subpopulations. In Chapter 2, we will discuss widespread methods for SNP selection in population genetics, and compare accuracy in individual assignment using a well-established parameter, the fixation index (F_{ST}) and novel applications of random forest-classification, a machine-learning approach that implements a series of decision trees for ranking features. We use parr (salmon juveniles) sampled from 11 rivers running into Lake Melville to evaluate the potential for individual assignment within this small geographic area. We test various sizes of subsets of genetic markers to determine the minimum panel size required to reach an adequate self-assignment accuracy, ≥ 90%. We then investigate patterns of incorrect assignment across methods to infer genetic structure across these rivers. To assess the broader applicability of our proposed methods, we also apply our feature-selection techniques to a published SNP dataset of Chinook salmon from coastal Alaska and the Yukon River (Larson et al., 2014b).

In Chapter 3 we investigate the overall structure within coastal Labrador, and apply a robust landscape genetics approach, redundancy analysis (RDA), to uncover the mechanism for divergence in this area. We use parr sampled from these same 11 rivers in Chapter 2, with some overlap of individuals, as well as parr sampled from an additional 24 sites in coastal Labrador (Figure 1.2). We apply these methods to two established panels: (1) 101 microsatellites, and (2) 376 single-SNP genotypes. We conclude with a discussion of our contribution and indicate the direction of future work in this area.

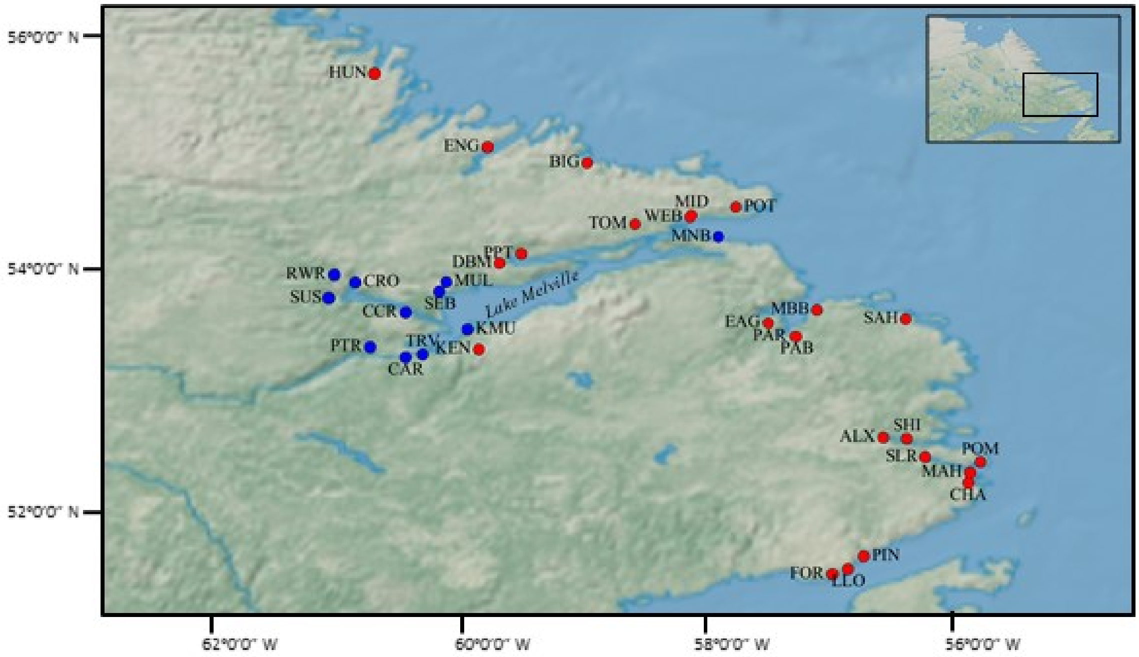

Figure 1.2: Sampling rivers across coastal Labrador. Individuals sampled at site locations in blue were used to study feature-selection techniques for GSI in Chapter Two. Data from site locations in both red and blue were used to investigate overall population structure and EAA in Chapter Three. For full names of rivers, see Table 2.1.

Chapter 2

Random Forest Feature-Selection for Individual Assignment

2.1 Individual Assignment

Genetic assignment of individuals to their source populations is useful for uncovering spatial distribution of populations and migration patterns (e.g. André et al., 2016) relevant to wildlife management and conservation (Manel et al., 2005). For exploited species, assignment tests may be used to monitor population-specific exploitation, ensuring the maintenance of genetic diversity and improving management practices through the identification of over-exploited stocks. Assignment tests have been implemented in commercial fishery species such as herring, *Clupea harengus* L., (Bekkevold et al., 2015), Atlantic cod, *Gadus morhua* L., (André et al., 2016), Chinook salmon, *Oncorhynchus tshawytscha*, (Smith et al., 2005; Templin et al., 2011, Larson et al. 2014a) and Atlantic salmon, *Salmo salar* (Karlsson et al., 2011, Bradbury et al., 2015a). These studies rely on genetic differences among populations to assign individuals to their source populations across large spatial scales (e.g. Bekkevold et al., 2015). Resolution of spatially distinct biological units across fine spatial scales can be difficult as weak genetic divergence may limit the accuracy of assignment tests (Larson et al., 2014a). Developing methods to detect this divergence and improve assignment accuracy may benefit management practices across both large and small geographic scales.

In this chapter, we identify and evaluate various sizes of SNP panels using global F_{ST} and three variations of RF: standard RF, Regularized Random Forest (RRF), and Guided Regularized Random Forest (GRRF) (Deng and Runger 2013). We aim to identify one or more methods for selection of an optimal panel, while comparing the trade-off between panel size and self-assignment accuracy across methods and identifying the minimum panel size required to achieve a minimum overall self-assignment accuracy of 90%. We provide evidence of successful implementation of machine-learning approaches on a metapopulation scale for site-by-site (river) classification to establish a relevant, non-redundant, reduced panel of genetic markers. By testing these novel approaches on an unpublished set of Atlantic salmon SNPs, and a published Chinook salmon data set, we explore methods for capitalizing on large genomic datasets for genetic population assignment, with potential for application across a range of systems.

2.1.1 Feature-Selection Techniques in Population Genetics

Currently, the most widely used methods for SNP selection in ecological research rely on measures of population differentiation (see Helyar et al., 2011; Rosenberg et al. 2005 for review). Most commonly, SNPs are ranked by fixation index, F_{ST} (Karlsson et al., 2011; Larson et al., 2014a; Larson et al., 2014c; Lemay and Russello, 2015; André et al., 2016). As a measure of differentiation of populations, F_{ST} for SNP selection can be calculated at each locus between subpopulations (pairwise F_{ST}) or for a metapopulation relative to the overall population (global $F_{ST;}$ Foll and Gaggiotti, 2006). Though widely used, it is difficult to gauge the applicability of F_{ST}-based methods across different study systems because published studies are often biased towards research demonstrating successful self-assignment. As F_{ST}-based methods only consider loci through a single, univariate rank for importance (Brieuc et al., 2015), the overall performance of the selected panel may be limited. Commonly used Bayesian algorithms rank locus importance based on inference of population structure through linkage disequilibrium (for example, Carlson et al., 2004), or deviation from Hardy-Weinburg Equilibrium (HWE). The latter has been implemented in user-friendly formats such as STRUCTURE (Pritchard et al., 2000) and BAPS (Corander et al., 2008).

As an alternative, iterative algorithms implemented in the software BELS (Bromaghin, 2008) and genetic algorithms (Topchy et al., 2004) have been proposed for informative SNP selection (Rosenberg, 2005). Though potentially an improvement for assignment-focused marker selection, like those previously described, both methods are computationally intensive, potentially limiting the number of markers considered. BELS also lacks consideration of various possible subsets of SNPs (Helyar et al., 2011). There is the potential for further work on machine-learning approaches such as k-nearest neighbour and genetic algorithms (Topchy et al., 2004).

In contrast to simple ranking, random forest (RF) is a machine-learning approach that considers a subset of features or predictive variables (e.g. SNPs) at each node to grow a series of decision trees (Breiman, 2001). In the classification implementation, an individual is assigned to a class (e.g. population), using a bootstrapped sample of these features or loci. Features can be ranked by importance based on the change in classification error affected by the presence or absence of a feature in a subset. The RF algorithm also considers loci in various combinations of subsets, improving the power of the algorithm to rank these features or loci for importance. The increasing popularity of RF in biological research has provided ample evidence to indicate its potential for successful use in population genetics. The regression implementation has been used to select SNPs to predict phenotypes (Bureau et al., 2005; Brieuc et al., 2015; Pavey et al., 2015) and to identify environmental parameters that may have an influence on population structures in landscape genetics (Zhan, 2016). RF classification has been applied as a method of feature-selection to predict microbial community structure, using phylogenetic and functional trait data (Ning and Beiko, 2015) and to select genes for functionality using microarray data (Díaz-Uriarte and De Andres, 2006; Deng and Runger, 2013; André et al., 2016); however, to our knowledge it has yet to be applied to SNP selection for population assignment.

2.1.2 Random Forest Feature-Selection

For RF classification, measures of importance of each feature can be calculated based on the reduction in accuracy of the model when the feature in question (i.e. SNP) is not included in a subset of features within a tree (Breiman, 2001). Decision trees based on subsets lacking highly informative features will have a higher error or reduced classification accuracy to a known class (i.e. river) when an important feature is removed, compared to an irrelevant marker, the removal of which will result in no reduction in model accuracy. This difference in model accuracy, averaged across decision trees with and without the locus in question is termed the mean decrease in accuracy (MDA). We used this measurement to rank loci based on importance in assignment (classification). Features or SNPs with a relatively high MDA will be deemed highly important for accurate classification. As the actual MDA value indicates relative importance, a strict cut-off threshold will vary for each data set. The overall error of a forest of trees is assessed as the out-of-bag (OOB) error. Similar to the calculation of MDA across subsets of features, OOB error is an average of model error across the bootstrapped (or bagged) samples or individuals.

Regularized random forest (RRF) and guided regularized random forest (GRRF) are variations on the RF algorithm designed to address issues with RF, and to optimize features for selection (Deng and Runger 2013). RRF uses a customizable parameter, the penalty coefficient (λ), which penalizes features at a node when making a classification decision. To be selected for importance and included in the selected panel, a feature must be more informative than the other features in the subset considered at a node as well as those already selected for importance, despite this penalty. As such, RRF is a more stringent application of RF and influences the selected feature set (panel) size. A larger λ (approaching 1) leads to a smaller penalty, resulting in a larger selected panel. Using the minimum regularization (λ=1) a feature must still be more informative than the already selected features to be included in the subset. Though this additional component to the RF algorithm provides a more stringent approach, the efficacy of RF and RRF may be limited by the number of nodes within the forest that consider a feature for importance to the model. That is, as a locus may not be present in many nodes, it may not be considered for importance often enough to truly inform the selection process, a problem referred to as node sparsity (Deng and Runger 2013).

GRRF addresses node sparsity by using an input of importance measures (from a previous RF run, for instance) to weigh each feature. This customizes the algorithm such that the penalty coefficient applied to features of presumably greater importance is less than that applied to features of less importance. GRRF uses an alternative parameter, gamma (γ), to control the weight of the importance score applied to each feature. A larger value of γ (approaching 1) leads to a smaller overall λ and will therefore result in a smaller feature set (Deng and Runger, 2013). The ability to fine-tune parameters to target

Table 2.1 Site locations and sample size for all study collections of juvenile salmon, sampled in 2013 and 2014.

River Name	Sample Size	Site ID	Latitude (N)	Longitude (W)
Cape Caribou River	21	CB	53°32'48,8"	60°36'27,0"
Caroline Brook	20	CL	53°15,232'	60°31,899'
Peters River	21	PR1	53°20'10,4"	60°47'15,3"
		PR2	53°20,345'	60°37,293'
Red Wine River	22	RW1	53°52,764'	61°27,976'
		RW2	53°52,928'	61°28,730'
Susan River	22	SR1	53°44,365'	61°3,275'
		SR2	53°44,184'	61°02,216'
Crooked River	21	CR	53°50,991'	60°48,863'
Kenamu River	22	KE	52°50,952'	60°08,279'
Main Brook River	21	MB	54°04,355'	57°52,374'
Mulligan River	17	MU	53°52,138'	60°05,392'
Sebaskachu River	22	SK1	53°47,397'	60°08,523'
		SK2	53°46,10'	60°10,575'
Traverspine River	22	TR	53°08,853'	60°27,769'

sizes of SNP panels that achieve maximum classification accuracy in combination offers a unique property to the RRF and GRRF algorithms and demonstrates their suitability for comparing assignment accuracy across panel sizes.

2.2 Sampling and Genotyping

A total of 231 juvenile (parr) Atlantic salmon were sampled from 11 rivers (1-2 sites per river) within Lake Melville, Labrador (Table 2.1, Fig. 2.1) in 2013 and 2014 by electrofishing and angling. Heart samples were collected and placed in 95% ethanol. DNA was isolated using the DNeasy Blood and Tissue kit or DNeasy 96 Blood and

Tissue kit (Qiagen, Toronto, ON, Canada) following the manufacturer's protocol, including the optional RNase A treatment. DNA samples were quantified using the Qubit dsDNA HS Assay Kit (Life Technologies, Burlington, ON, Canada) with assays read on a Qubit v2.0 (Life Technologies) or using the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies) with assays read on a FLUOStar OPTIMA fluorescence plate reader (BMG Labtech, Ortenberg, Germany). The DNA quality for all samples was verified by agarose gel electrophoresis of 100 ng of extracted DNA, visualized using SYBR Safe (Life Technologies), and documented using a Gel Logic 200 (Kodak, Rochester, New York, United States). Individuals were genotyped using a 220K target, bi-allelic SNP Affymetrix Axiom array developed by the Centre for Integrative Genetics (CiGene, Ås, Norway). These SNPs were a subset of those in the 930K XHD *Ssa*l array (dbSNP accession numbers ss1867919552–ss1868858426).

Ten fish were genotyped twice to assess genotyping error rate. Loci with inconsistent calls among replicates were removed from the data set. Loci were then filtered in PLINK v. 1.07 (Purcell et al., 2007) for global minor allele frequency (MAF) below 5%. One locus was also removed for having more than 5% missing data across all sites. Pairwise population F_{ST} (Weir and Cockerham, 1984) was calculated using Arlequin (Excoffier et al., 2005), Table 2.2. Additional missing genotype data, consisting of 0.08% of the data, were imputed using the function rfImpute in the RandomForest package, using 5000 trees with all other parameters set to default.

We further reduced our panel for downstream feature-selection by removing redundant SNPs and SNPs in linkage disequilibrium using the genepop_toploci function in the R package 'genepopedit' (Stanley et al., 2016) at an R^2 threshold of 0.2 and a minimum global F_{ST} of 0.05. Though this is a highly stringent approach, reductions in the dataset are helpful both to reduce computational load and to increase consistency of markers across subsets (and therefore confidence in the importance of selected SNPs). As evidence suggests that under linkage disequilibrium RF performance may be reduced, redundancy in the dataset should be considered prior to or during the feature-selection process (Toloşi and Lengauer, 2011; Meng et al. 2009).

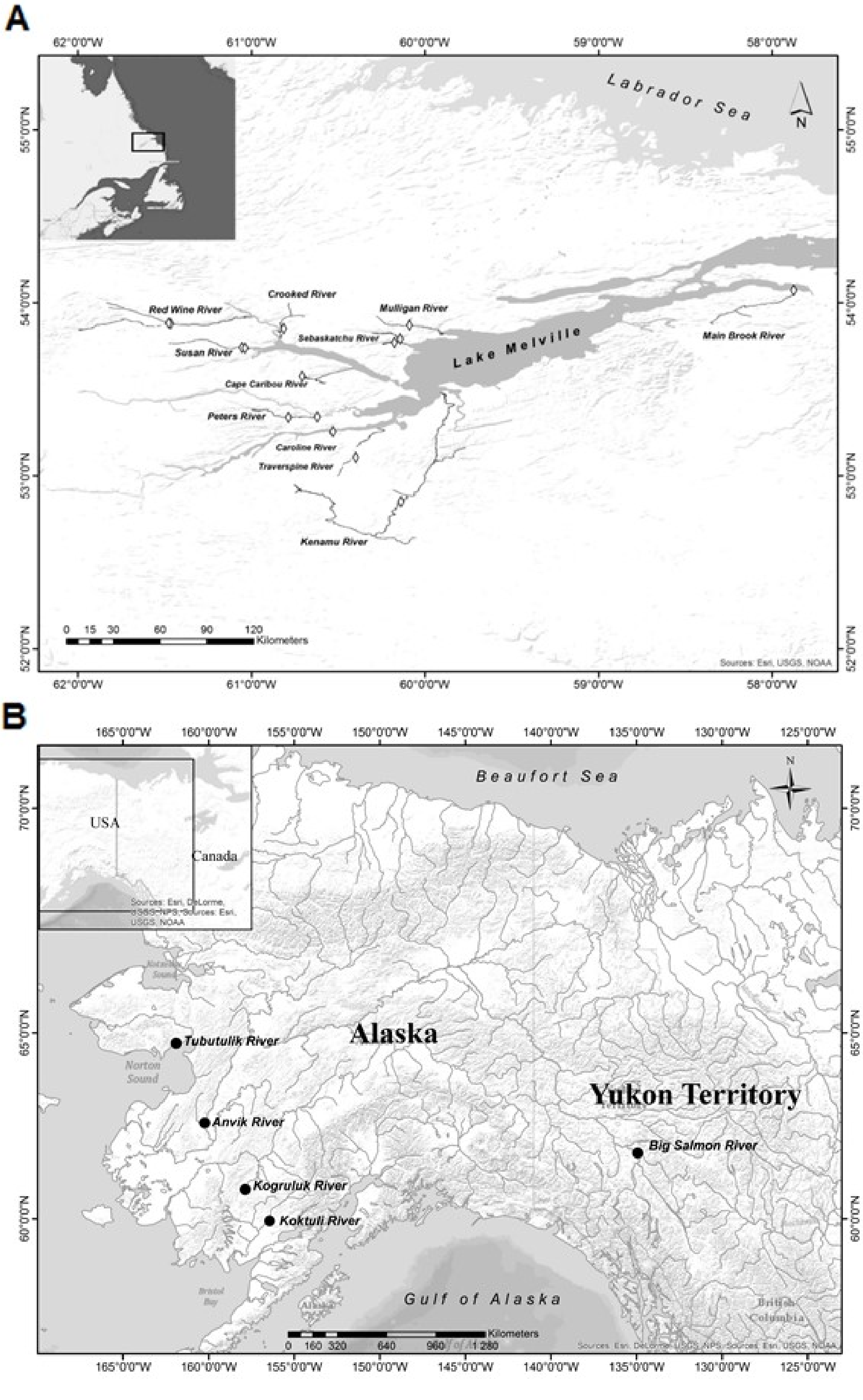

Figure 2.1: Sampling locations of (A) Atlantic salmon (*Salmo salar)* from Lake Melville, Labrador, Canada and (B) Chinook salmon (*Oncorhynchus tshawytscha*) from Western Alaska and the Yukon River. See Table 2.1 for site coordinates, site ID, and sample size for Atlantic salmon sampling. Coordinates for Chinook salmon sampling sites were obtained from Larson et al. (2014a). Maps were created using ArcGIS (ESRI, 2011).

Chinook salmon data contained 10,944 SNPs identified through *Sbf1* restriction-site-association DNA (RAD) sequencing for 265 adult individuals from five locations: four populations in coastal western Alaska and one in Yukon River (Fig. 2.1B). SNPs were removed from an original pool of 42,351 putative loci, if genotyped in <80% of individuals, and were reduced to one SNP per RAD tag (Larson et al., 2014a). Further, SNPs were filtered for linkage disequilibrium, evidence of paralogous sequences, deviation from Hardy-Weinberg equilibrium and MAFs of <0.05 (Larson et al. 2014a). Data were imputed and filtered for F_{ST} and redundancy as described above.

2.2.1 Overall Panel Characteristics

Of the original 220K SNPs genotyped for Atlantic salmon, 276 were called inconsistently across samples. Overall genotyping accuracy was > 99.8%. After removing these loci and initial filtering for MAF, 93,058 SNPs remained in the Atlantic salmon dataset for further selection. Average global, locus-specific F_{ST} (mean: 0.059, range: 0 - 0.58) and pairwise population F_{ST} ranking across the whole panel (Table 2.2, Fig. 2.2) indicated relatively low genetic differentiation. After controlling for linkage disequilibrium and co-variance in the panel across all chromosomes, and filtering at a global F_{ST} of 0.05, 8,434 non-redundant loci remained in the panel, with F_{ST} frequency distribution similar to that observed in the unfiltered dataset (Fig. 2.2). The 10,944 SNP panel accessed for this study (Larson et al. 2014b) was reduced to 2,178 SNPs after filtering at a global F_{ST} of 0.05 and linkage threshold of 0.2. For pairwise population F_{ST}, see Larson et al. (2014a). The size of the panel ranged from 51 to 697 SNPs and 41 to 528 SNPs for the Atlantic salmon and Chinook salmon datasets, respectively (Table 2.3). Though SNPs were most often selected by only a single selection method, some SNPs were identified by more than one method (Fig. 2.3). A total of 17 and 32 SNPs were selected by all four SNP selection methods for Atlantic and Chinook salmon, respectively. Overlap in SNPs occurred more often with Chinook salmon data, likely a result of the smaller panel size (2,178 SNPs) relative to the 8,434 SNPs in the Atlantic salmon panel.

Table 2.2: Pairwise population F_{ST} (bottom diagonal) and p-values (top diagonal) calculated using 1000 iterations in Arlequin 3.5.2.2 (Excoffier et al. 2005). P-values of zero indicate values < $e10^{-6}$.

	CB	CL	PR	RW	SR	CR	KE	MB	MU	SK	TR
CB	-	0	0	0	0	0	0	0	0	0	0
CL	0.034	-	0	0	0	0	0	0	0	0	0.0009
PR	0.131	0.136	-	0	0	0	0	0	0	0	0
RW	0.024	0.033	0.122	-	0	0.0048	0	0	0	0	0
SR	0.037	0.045	0.133	0.025	-	0	0	0	0	0	0
CR	0.019	0.027	0.120	0.003	0.023	-	0	0	0	0	0
KE	0.025	0.022	0.125	0.026	0.037	0.023	-	0	0	0	0
MB	0.053	0.061	0.151	0.046	0.057	0.042	0.058	-	0	0	0
MU	0.061	0.068	0.154	0.055	0.065	0.050	0.067	0.036	-	0	0
SK	0.076	0.084	0.165	0.070	0.079	0.064	0.080	0.056	0.031	-	0
TR	0.036	0.010	0.138	0.036	0.049	0.033	0.017	0.069	0.075	0.090	-

2.3 SNP Selection for Identifying Panel Subsets

Ideally, assignment analysis with loci selected for population assignment would implement a training/holdout approach, such that the individuals used for marker selection would be different from those used for assignment analysis (Anderson, 2010). Though upward grading bias (over-estimations of assignment accuracy) is effectively diminished by this approach, a completely independent training and holdout set is often unfeasible due to limitations in sample size. To overcome this, Anderson (2010) proposes a combined training/holdout/*leave-one-out* strategy where a subset of individuals (training set) are used for locus selection, and all individuals are used to establish a baseline for assignment. However, self-assignment accuracy is calculated based solely upon the assignment of the individuals in the holdout set. As such, all loci were selected using a subset of individuals. For both datasets one-third of the individuals from each site (approximately 7 for Atlantic salmon data and 19 for Chinook salmon data) were randomly selected for all methods of locus selection.

2.3.1 RF-based SNP Selection

Data was formatted using a custom R script such that individuals at a given locus were assigned 0, 0.5, or 1, for an individual that is homozygous for the minor allele, heterozygous, or homozygous for the major allele, respectively. We ran RF using the R package ‘randomForest’ (Liaw and Wiener, 2002) on our filtered datasets. To determine our appropriate *ntree* parameter (number of trees), we ran RF using 125, 250, 500, 1000, 2000, 4000, and 8000 trees, 10 times each. As overall out-of-bag (OOB) error stabilized at approximately 2000 trees for both Atlantic and Chinook data, we accepted this as suitable for our analysis (Boulesteix et al., 2012). The m_{try} parameter (the number of features considered at a node) was tested at default (the square root of the number of features), half default, and twice default, as suggested by Liaw and Weiner (2002). Error was lowest at twice default for

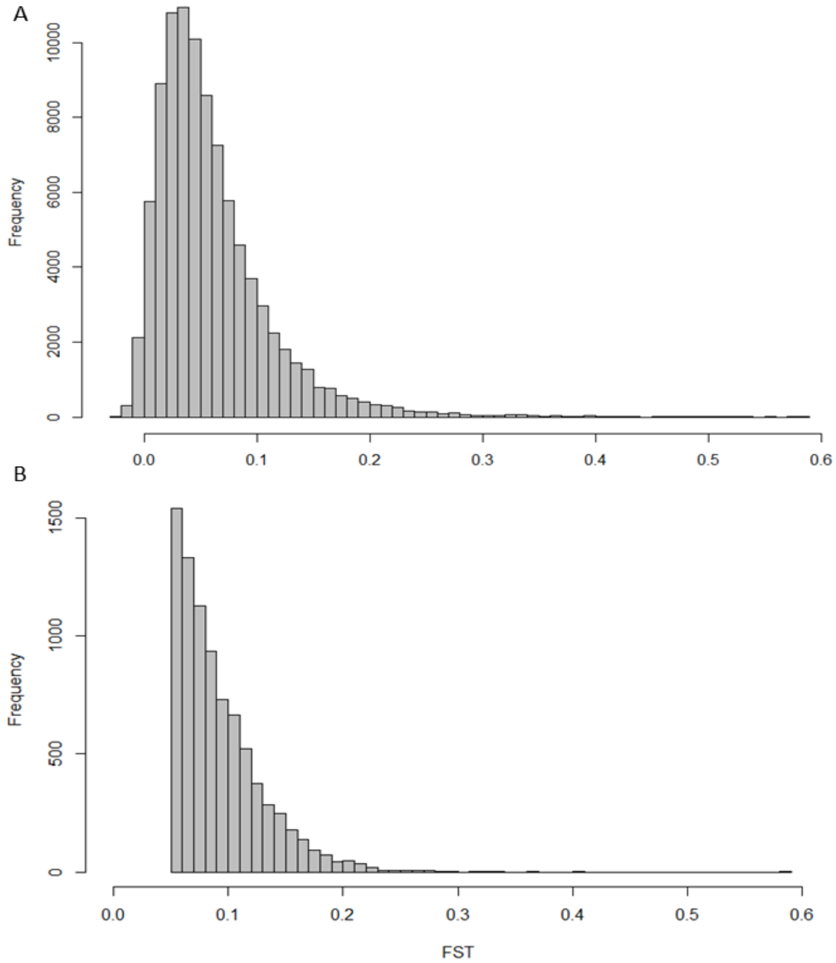

Fig. 2.2: Frequency distribution of global F_{ST} (A) across all loci after initial filtering (93,058 SNPs) and (B) after filtering for redundancy (R^2 linkage threshold of 0.2) and F_{ST} (threshold of 0.05) in genepopedit (Stanley et al. 2016).

Table 2.3**:** Properties of panels selected for assignment analysis by F_{ST}-rank, Random Forest (RF), Regularized Random Forest (RRF), and Guided Regularized Random Forest (GRRF). Panel size column indicates '(Rank) Panel Size' for RF-selected panels. See Fig. 2.3 for intersections across methods.

Method	**Parameter for selection**	**Parameter Value**	**Panel Size Atlantic Salmon**	**Panel Size Chinook Salmon**
F_{ST}	Top ranked	-	60	47
		-	85	65
		-	104	88
		-	130	112
		-	184	134
		-	266	182
		-	344	240
		-	508	384
		-	519	454
		-	670	509
RF	Within (x) rank across all 5 runs	-	(800) 66	(400) 41
		-	(825) 90	(600) 74
		-	(850) 110	(700) 91
		-	(875) 135	(850) 125
		-	(900) 157	(950) 167
		-	(950) 201	(1000) 216
		-	(1050) 298	(1100) 277
		-	(1200) 435	(1250) 341
		-	(1400) 605	(1400) 437
		-	(1500) 697	(1500) 519
RRF	Penalty coefficient (λ)	0.75	51	47
		0.8	83	71
		0.825	114	94
		0.85	140	110
		0.875	180	150
		0.9	275	191
		0.925	336	260
		0.95	515	364
		0.975	604	470
		0.99	710	528
GRRF	Weight of penalty (γ)	0.25	60	47
		0.2	85	65
		0.175	104	88
		0.15	130	112
		0.125	184	134
		0.1	266	182
		0.075	344	240
		0.05	508	384
		0.025	519	454
		0.01	670	509

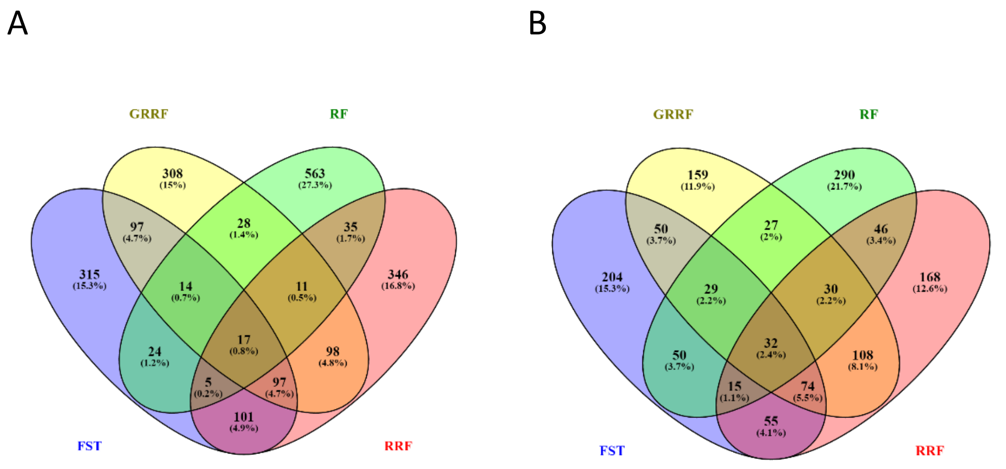

Figure 2.3: Overlap of SNPs from largest panels created using $F_{\mathrm{ST},}$ Random Forest (RF), Regularized Random Forest (RRF) and Guided Regularized Random Forest (GRRF) for (A) Atlantic salmon data and (B) Chinook salmon data (Larson et al. 2014a). See Table 2.3 for panel information.

both Atlantic and Chinook data and was therefore used as such for our analyses. We used a minimum node size (minimum size of terminal nodes or leaves) of five, allowing larger trees to be grown while decreasing run time (see 'randomForest' R documentation), with all other parameters set to default (Liaw and Wiener, 2002).

For feature-selection, we used five runs of RF, resulting in five separate lists of SNPs ranked by MDA. Panels of various sizes were created by identifying SNPs present in all five lists at 10 ranking levels. These levels were selected to create panels of 40-700 SNPs, after ensuring that each list contained only features with a positive MDA. For example, SNPs consistently ranked within the top 800 loci in all five lists were aggregated to form a consensus panel of 67 SNPs (Table 2.2).

RRFs and GRRFs were run using the R package 'RRF' (Deng and Runger, 2013). Both methods were run using the same parameters as those used for RF (described above). We tested ten parameter values for the penalty coefficient (λ) running RRF and 10 parameter values for gamma (γ) when running GRRF (Table 2.2). Parameters were selected to encompass a range of regularization penalties and to ensure a diversity of panel sizes for individual assignment. A vector of importance measures (MDA scores) determined by a single RF run for feature (SNP) rank was applied for feature weight in GRRF, as described above.

2.3.2 F_{ST}-based SNP Selection

We tested F_{ST} as a method of SNP selection using panels of loci ranked by global F_{ST} calculated using the R package ‘genepopedit’ (Stanley et al., 2016). To assess the assignment power of various panel sizes of SNPs ranked by F_{ST}, we created panels of size equal to those established using GRRF for cross-method comparison (Table 2.3). To visualize the overlap of SNPs selected across all methods, Venn diagrams were created for the largest panels across all SNP-selection methods using Venny v.2.1 (Oliveros, 2015).

2.4 Approaches for Individual Assignment

Assignment tests for GSI when the population of origin is already known (or suspected, based on sampling location) or for a simulated baseline, can apply a variety of classification algorithms, such as those discussed above for use in the feature-selection step of the pipeline. In general, most assignment methods have implemented maximum likelihood, for instance, the user-friendly software, ONCOR (Kalinowski et al., 2008) and GENECLASS2 (Piry et al., 2004), or Bayesian approaches, such as STRUCTURE (Pritchard et al., 2000) and BAPS (Corander et al., 2008). These two general approaches, however are susceptible to over-estimation of assignment accuracy as population allele frequencies often deviate from baseline (sample) allele frequencies, particularly over resampling of individuals to create a simulated baseline (Anderson et al., 2008). As an alternative, Anderson et al. (2008) proposes gsi_sim (genetic stock identification simulation) to estimate individual assignment and mixture proportions. By limiting the training set used for marker selection to a subset of individuals as described above, and implementing a LOO cross-validation method, gsi_sim controls for high grading bias within power analysis without reducing the sample size of the dataset. Gsi_sim creates simulations of individual genotypes through bootstrap sampling and assigns these individuals to a population based on the true baseline calculated across all individuals. This is particularly useful for studies with relatively low sample sizes as there is likely a larger difference between baseline allele frequencies and actual parametric population frequencies. The improvement of assignment accuracy is also expected to be largest for fine-scale studies, where genetic differences in populations are expected to be small. ‘Assigner’ is an R package developed to run filtering procedures and conduct assignment and mixture analysis with NGS data and has the option to implement gsi_sim. ‘Assigner’ (Gosselin et al., 2015) was used here to implement gsi_sim (Anderson et al., 2008), to conduct assignment analysis. Whitelists, or lists of loci to be considered for assignment, were created from each SNP selection method using custom R scripts for input into ‘assigner’. Though all individuals were used to create the baseline for gsi_sim, only the assignment of the holdout individuals was used to assess self-assignment accuracy.

Significance of SNP selection method was determined by an ANOVA comparing second degree polynomial models with and without the SNP-selection method term. We investigated consistent patterns of incorrect assignment across putative populations (rivers) by observing assignment matrix heatmaps of the smallest panels across all SNP selection methods. We also compared pairwise population F_{ST} values to discrepancies in pairwise mismatches (the number of individuals incorrectly assigned between paired populations) between F_{ST}-rank and GRRF selection methods, to further assess the optimal application of each method. That is, for a given pair of putative populations, the proportion of individuals that were incorrectly assigned from one study site to the other when using GRRF for SNP selection was subtracted from the proportion of individuals incorrectly assigned (within that pair of sites) using F_{ST} -rank. This allowed us to visualize a preferred method for sites at a given pairwise $F_{ST.}$

2.5 Panel Performance

2.5.1 Atlantic salmon in Labrador, Canada

Across panel sizes, we found that panels selected by F_{ST} ranking had the lowest self-assignment accuracy on average (mean=79.4%, SE=1.8) (Fig. 2.4A). Self-assignment accuracy for panels selected using RF, RRF and GRRF perform better overall (RF: mean=81.8%, SE=1.8; RRF: mean=81.5, SE=2.6; GRRF: mean=82.1, SE=2.5). An ANOVA comparing the fit of polynomial models with and without considering SNP selection method indicated marginal significance ($F_{28,37}$=2.54, p=0.048). The difference between methods varied with panel size. In the smallest panel sizes (50-100 SNPs), F_{ST}-ranked panels had better or comparable self-assignment accuracy with RF-based panels (Fig. 2.4A). In small- to medium-sized panels (101-200 SNPs), RF-selected panels performed best (up to 7.8 percentage points for panels of comparable size), while GRRF-selected panels most often have the highest self-assignment accuracy in larger panels (>200 SNPs). In all cases but the three smallest panel sizes (60, 85 and 104 SNPs), GRRF-selected panels outperformed F_{ST} -selected panels of the same size by a margin of 3.2 to 4.9 percentage points. For smaller panels, RF-selected panels outperformed F_{ST} -selected panels by up to 5 percentage points, although the highest accuracy of the smallest panel was 70.64%, observed in the F_{ST} -selected panel. A threshold of 90% accuracy overall was achieved only with the largest panels created using GRRF and RRF, composed of 670 and 710 SNPs, respectively.

We also investigated how self-assignment varied across sites (Fig. 2.5). Many sites showed consistently high (above 90%) self-assignment regardless of SNP selection method, whereas others have a higher frequency of mis-assignment. In these latter sites (Caroline River and Traverspine River; Red Wine River and Crooked River), the margin in performance between F_{ST} and RF-selected panels widened, in some cases by up to 40 percentage points, as seen in Caroline River (Fig. 2.5A). Some study sites showed a higher self-assignment accuracy with F_{ST} -based methods and some with RF-based methods (Fig. 2.5A). To understand these patterns, we compared pairwise population F_{ST} values with the difference in the proportion of mismatches across paired sites between F_{ST} and the best performing RF-based method overall, GRRF (Fig. 2.6). While we expected that populations with a low pairwise F_{ST} value may tend to be more successful with one SNP selection method over another, we did not find consistency across panels. As pairwise F_{ST} values increased, these differences shifted toward zero, but at low pairwise F_{ST} values there was no tendency for more mismatches to occur in one method over another (Fig. 2.6).

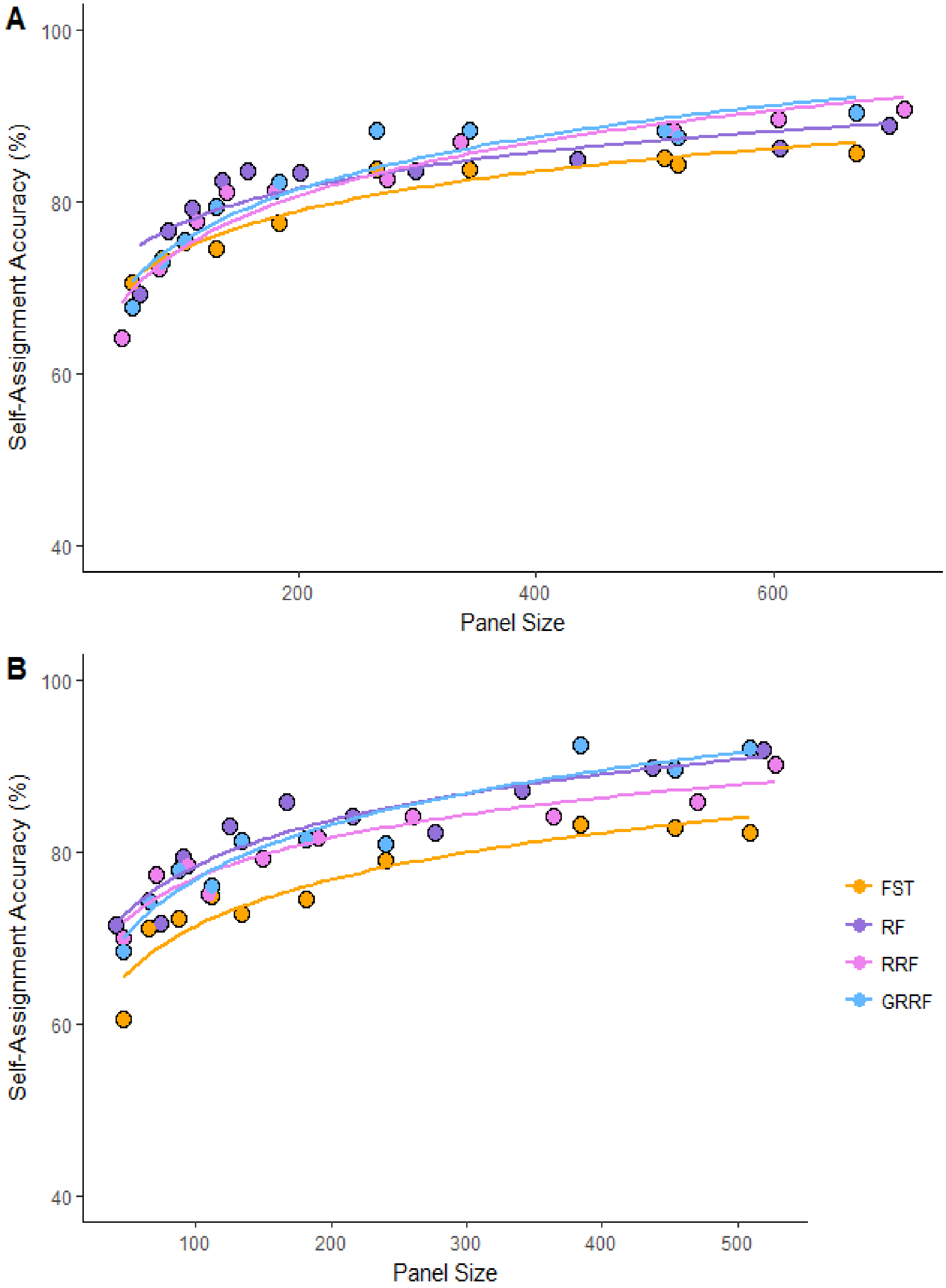

Figure 2.4: Average, overall self-assignment accuracy of identified SNP panels (50-700 SNPs) for (A) Atlantic salmon and (B) Chinook salmon (Larson et al. 2014a) calculated across sampling sites. SNP selection method (F_{ST} rank, RF, RRF and GRRF) is indicated by colour.

To identify patterns of mis-assignment, we created heatmaps demonstrating mis-assignment from 'assigner' outputs for the smallest panel sizes from all methods to ensure consistency in observed patterns (Fig. 2.7). From this we observed a high rate of mis-assignment between Red Wine River and Crooked River, and between Caroline River, Traverspine River, and to a lesser degree, Kenamu River. Regardless of the method of SNP selection, we observed that incorrectly assigned individuals from Red Wine River frequently assigned to Crooked River (30.0% of all individuals), and vice versa (35.7% of all individuals). Incorrectly assigned individuals from Caroline River were often assigned to Traverspine River (30.7% of individuals). Although individuals from Traverspine River generally self-assigned well, incorrectly assigned individuals often assigned to Caroline River (13.3% of all individuals) (Fig. 2.7). Up to 10% of individuals from Traverspine River and Caroline River incorrectly assigned to Kenamu River, while incorrectly assigned individuals from Kenamu River most often assigned to Traverspine River or Caroline River (up to 13.3%). We also observed consistent self-assignment of 81% of individuals in Peter's River (Fig. 2.5a). Regardless of panel-selection method, the same four individuals mis-assigned to Crooked River, Red Wine River or Kenamu River (Fig 2.7). It is also worth noting that these incorrectly assigning individuals were sampled from a site upstream of the river mouth, while all other samples were collected near the river mouth (Table 2.1), possibly indicating within-river population structure. These consistent patterns in mis-assignment between and within geographically proximate sites (Fig. 2.1) illustrate the difficulty with population assignment at the finest spatial scales. Although there appears to be some level of genetic divergence between individuals at each of these sites, either computational methods are limited in their ability to detect and fully discern these populations, or they are in fact genetically and behaviourally the same population with higher genetic diversity than nearby populations.

2.5.2 Chinook salmon in Alaska, USA

Similar to our findings with the Atlantic salmon data, we found consistently higher self-assignment accuracy with RF-based selection methods (RF: mean=82.7%,

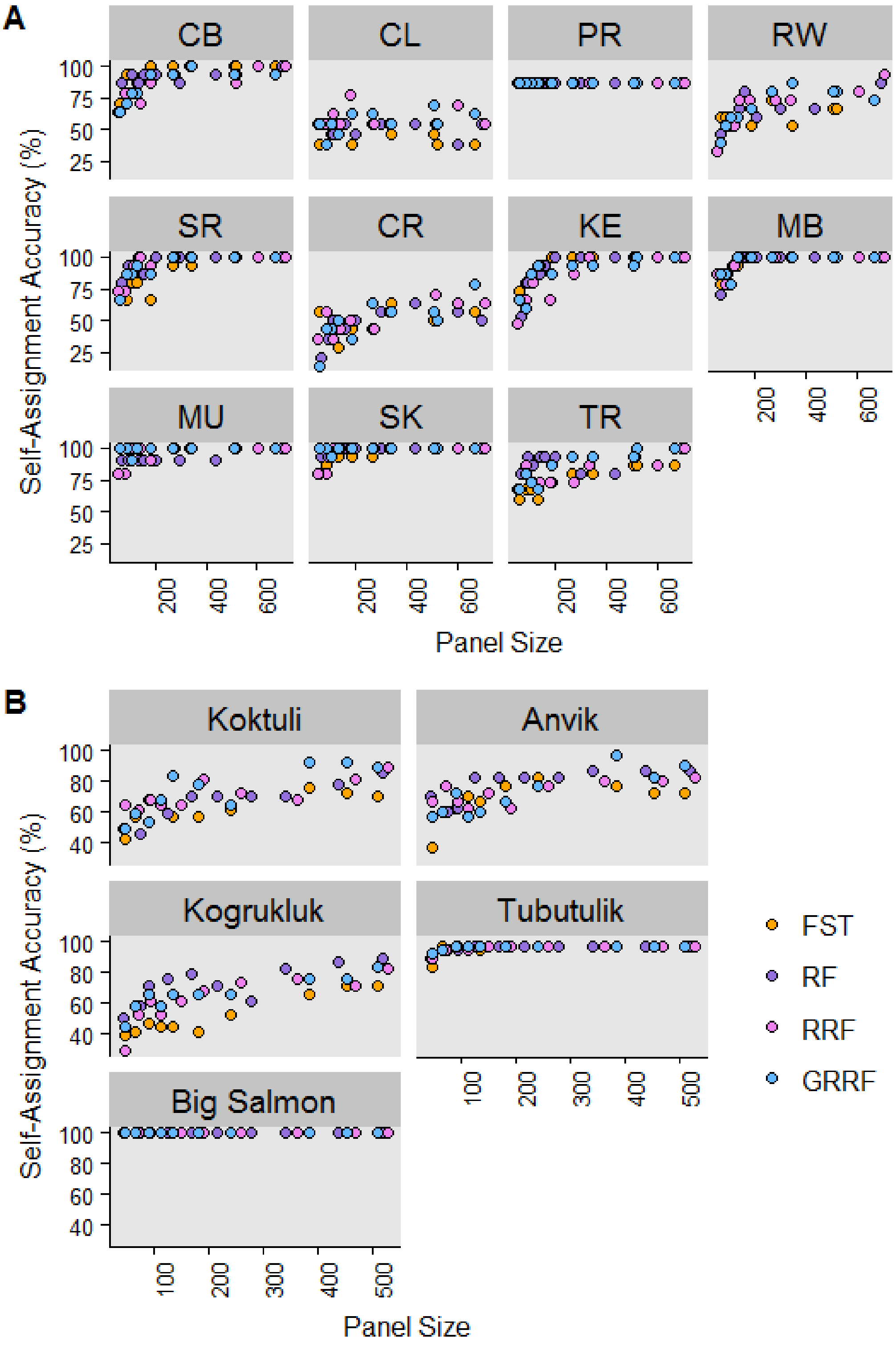

Figure 2.5: Self-assignment accuracy of identified SNP panels (50-700 SNPs) across all sampling sites as indicated by site ID (see Table 1) for (A) Atlantic salmon and (B) Chinook salmon (Larson et al. 2014a). SNP selection method (*F_{ST}* rank, RF, RRF and GRRF) is indicated by colour. Note differences in y-axis range between A and B.

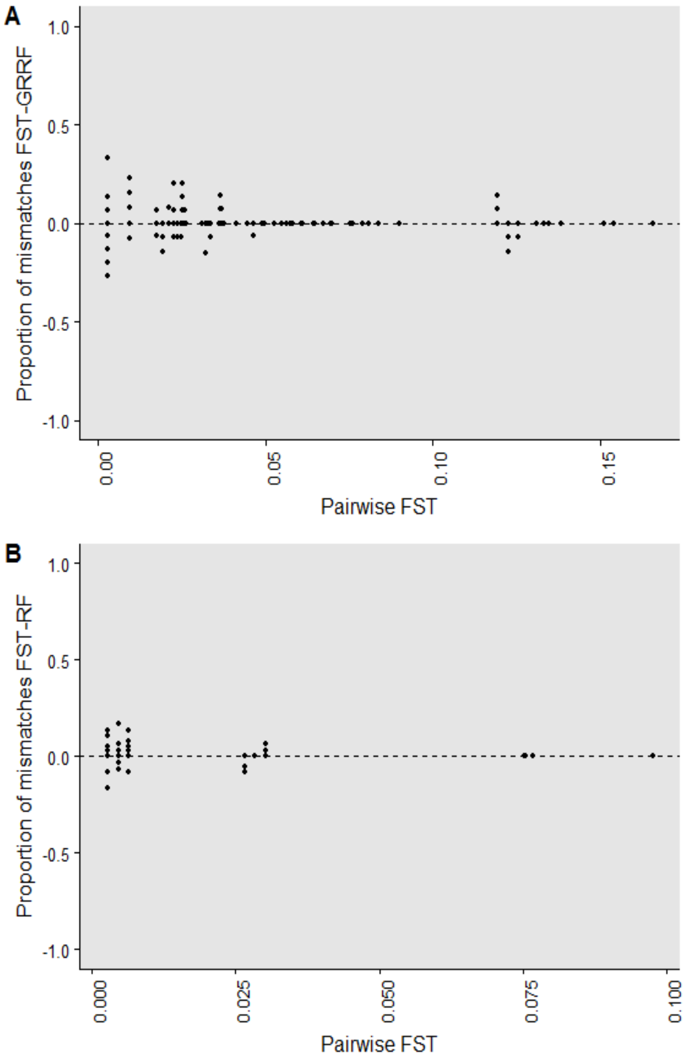

Figure 2.6: Difference between F_{ST} and best overall RF-based SNP-selection method showing proportion of individuals from one study location incorrectly assigned to an alternative location, sorted by pairwise population F_{ST} for (A) Atlantic salmon, comparing F_{ST} and GRRF and (B) Chinook salmon (Larson et al. 2014a), comparing F_{ST} and RF.

SE=2.16; RRF: mean=80.7%, SE=1.84; GRRF: mean=81.5%, SE=2.5) compared to F_{ST}-selected panels (mean=75.4%, SE=2.18) (Fig. 2.4B) for the Chinook salmon dataset. SNP selection method was found to have a significant effect on the polynomial model ($F_{28,37}$=4.08, p=0.001). As observed with the Atlantic salmon data, smaller to medium-sized panels (up to 200 SNPs) performed best with RF SNP selection (up to 11.2 percentage points for panels of comparable size), while GRRF often had the highest self-assignment accuracy of the larger panels. However, unlike the Atlantic salmon data, F_{ST}-selected panels showed reduced self-assignment accuracy at both small and large panel sizes. GRRF-selected panels outperform F_{ST} -selected panels of the same size by a margin of 1 to 9.8 percentage points. A 90% self-assignment accuracy threshold was reached with the largest panels of all RF-based selection methods, and with a panel of 384 SNPs selected by GRRF at 92.4% overall accuracy.

Self-assignment accuracy decreased (Fig. 2.5B) and mis-assignment increased among closely associated sites (Anvik River, Koktuli River, and Kogrukluk River) with reduced pairwise F_{ST} values (Larson et al., 2014a). Larson et al. (2014a) found the lowest genetic divergence between these three rivers, particularly between Kogrukluk River and Koktuli River, with these rivers showing the lowest pairwise F_{ST} (0.003) and highest occurrence of overlap in a principal component analysis (PCA). Accordingly, we found the highest rate of incorrect assignment occur between these two rivers (Fig. 2.8). Though F_{ST} -selected panels most often had the lowest accuracy, this was not consistent across all sites. As with the Atlantic data, we investigated the relationship between pairwise population F_{ST} values and the difference in the number of mismatches occurring between a given pair of populations when using F_{ST} values versus the best performing method overall, RF. Though higher pairwise F_{ST} is associated with reduced differences between these approaches, there is no indication that outperformance of a particular method is associated with F_{ST} (Fig. 2.6B).

2.6 Discussion of Results

Overall, in both Atlantic salmon and Chinook salmon, we achieved self-assignment accuracy above 90% for most populations using targeted panels of loci, comparable to or higher than that of broad-scale (Ozerov et al., 2013; Moore et al., 2014; Bradbury et al., 2015; Bradbury et al., 2016) and fine-scale (Vähä et al. 2016) mixed-stock analyses. Machine-learning algorithms in contrast to F_{ST}-rank, allow SNPs to be selected based on their relevance directly to the study question, be it correlation with a phenotype (for example, Brieuc et al., 2015) or classification to a reference population. Machine-learning techniques also consider the importance of loci in combinations with other loci, in contrast to loci selected solely on individual importance. If combinations of markers perform better than expected given the individual characteristics of each marker, then machine-learning methods might select relevant markers that would otherwise go undetected. For phenotype-genotype studies, this approach is more likely to consider and identify important loci involved in polygenic traits, which may otherwise be discarded. In a SNP selection study targeting disease indicators (Shah and Kusiak, 2004) a set of 172 SNPs was reduced by 85% with little cost to the performance of the assignment model. It is not surprising then, that machine-learning algorithms may increase the accuracy of population assignment.

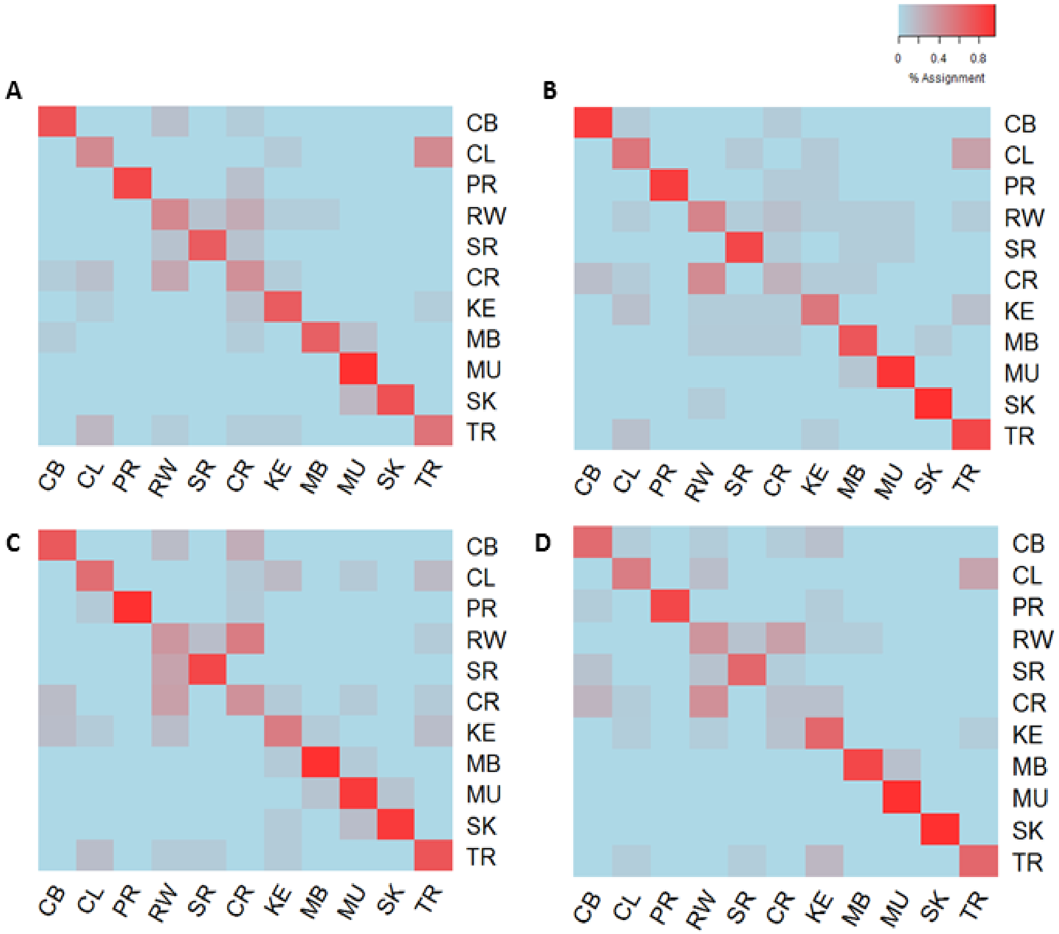

Figure 2.7: Assignment matrix heat maps indicating average percent assignment of Atlantic salmon data calculated across the smallest panels (51-66 SNPs) established using (A) F_{ST} rank, (B) RF, (C) RRF and (D) GRRF. Colour intensity indicates probability of an individual from a reference population (rows) being assigned to a given population (columns).

In the Atlantic salmon dataset, we observed an improvement of up to 40 percentage points within a single site and up to 7.8 percentage points in overall assignment accuracy, compared to F_{ST}-selected panels of similar size. This improvement in self-assignment accuracy was most apparent in larger panel sizes. In the three smallest panel sizes, F_{ST}-selected panels had comparable accuracy to those selected using RF methods. We observed frequent and consistent mis-assignment in particular sites across SNP selection methods (Fig. 2.5A, Fig. 2.7A). Caroline River and Traverspine River, as well as Red Wine River and Crooked River showed higher levels of mis-assignment with each other than most other rivers, though self-assignment was still higher than would be expected if individuals were randomly assigned to one of these two paired sites (i.e.

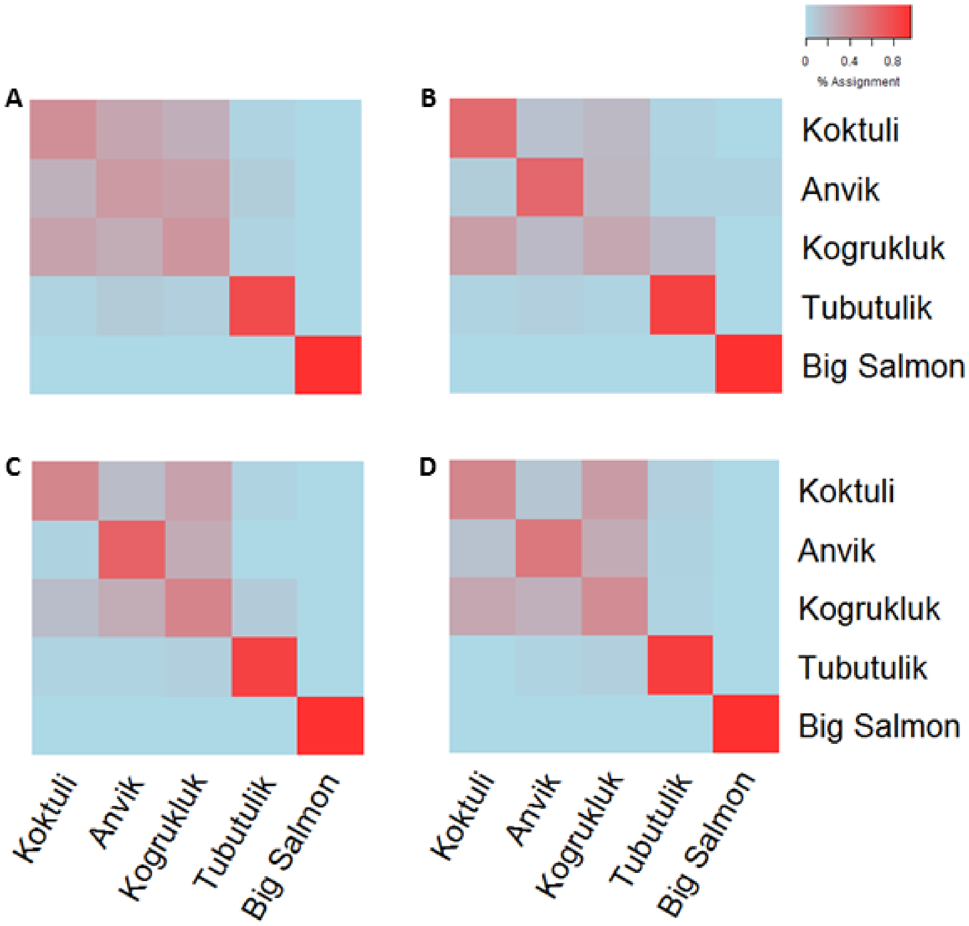

Figure 2.8: Assignment matrix heat maps indicating average percent assignment of Chinook salmon data calculated across the smallest panels (41-47 SNPs) established using (A) F_{ST} rank, (B) RF, (C) RRF and (D) GRRF. Colour intensity indicates probability of an individual from a reference population (rows) being assigned to a given population (columns).

50%). This reduction in self-assignment accuracy likely reflects close genetic relationships or admixing between these neighbouring populations within the same river tributary. Alternatively, this may indicate multiple spawning sites (rivers) for the same population. Pairwise F_{ST} values were considerably lower for these pairs of rivers, indicating relatively low genetic divergence (Table 2.2). We also observed that assignment accuracy within Peter's River rarely deviated from 81%. Across all runs, individuals from Peter's River sampled from the site closest to the river mouth (Fig. 2.1A) were incorrectly assigned to Red Wine River, Crooked River or Susan River. We suspect that there may be genetic structuring occurring within Peter's River or that these individuals are progeny of recent migrants from one or more of these populations. More samples to detect population structure within these rivers may indicate the presence of distinct upstream and downstream populations within Peter's River, or other rivers with natural barriers influencing within-stream population structure. Though our study revealed clear patterns of mis-assignment in pairs, it is likely that patterns of incorrect assignment in other natural systems may be more complex (Vähä et al., 2016), particularly when assigning to a greater number of sites (Moore et al., 2015) or if the subpopulations in question are less genetically divergent. For such studies, GRRF or other modified machine-learning approaches may be well suited to SNP selection for accurate overall assignment accuracy, as shown by the successful application in the present study.

In Chinook salmon, our application of RF-based methods to a large (10,944 SNPs), published data set (Larson et al. 2014a) provided further evidence of the usefulness of RF feature-selection. RF-selected panels had consistently higher self-assignment accuracy compared to those selected by F_{ST}-ranking. Using a panel of 39 SNPs developed from expressed sequence tags, Larson et al. (2014a) obtained an overall accuracy of 54.4% using a LOO approach, comparable to our smallest F_{ST}-ranked panel of 47 SNPs, with an overall accuracy of 60.6% (Fig. 2.4B). However, the smallest RF-based panels resulted in overall self-assignment accuracy of 71.6%, 70.0% and 68.6% for RF, RRF and GRRF, respectively (Fig. 2.4B). Self-assignment accuracy of the largest panel (509 SNPs) using GRRF was comparable to that achieved using all 10,944 SNPs (Larson et al. 2014a) (92.0% and 96.4%, for the 509 SNP panel and 10,944 SNP panel respectively). Comparable self-assignment accuracy (above 90%) was reached using a panel of 500 multi-SNP (haplotype) loci (McKinney et al., 2017) selected based on F_{ST} rank with individuals assigned using gsi_sim. In this study McKinney et al. (2017) combined Koktuli River and Kogrukluk River into a single group for mixture analysis and individual assignment. That we achieved a similar level of self-assignment accuracy with single-SNP panels of equal or lesser size without combining sampling locations speaks to the predictive power of RF-based methods for marker selection. Populations with the lowest self-assignment accuracy (Anvik River, Kogrukluk River and Koktuli River) (Fig. 2.5B, Fig. 2.7B) were consistent with those found to be the least divergent, with the lowest pairwise F_{ST} (0.003 - 0.006) and high degree of overlap in a PCA analysis (Larson et al., 2014a). While FST-selected panels had the lowest accuracy for Kogrukluk River and Koktuli River, this disparity was reduced in Anvik River. The increased self-assignment accuracy obtained here is comparable with that achieved by McKinney et al. (2017) using haplotype genotypes of the same dataset. However, this improvement was achieved under a simpler assumed population structure as rivers with the lowest pairwise F_{ST} (Kogrukluk River and Koktuli River) were combined into a single class.

Overall, RF methods often outperformed the F_{ST} based method, however, the Atlantic and Chinook salmon data showed discrepancies in the optimal method of SNP selection for each site. By comparing pairwise F_{ST} with the difference in the number of mismatches between paired populations when using the best RF-based method and F_{ST} for SNP selection, we hoped to elucidate these findings. However, we did not find strong evidence that either of these methods perform better under certain conditions of population divergence (Fig. 2.6).

Across all analyses, we often observed fluctuations in self-assignment accuracy. There are many instances of accuracy decreasing with increasing panel size, even when markers were selected using the same method (Fig. 2.5, Fig. 2.7). Using a simulated baseline based on a subset of SNPs for individual assignment leaves room for noise and minor fluctuations depending on the SNPs used for assignment. Increasing panel size would not always increase accuracy if less-informative SNPs are included in the panel. Though our methods aim to select the most informative SNPs, those selected for classification based on the training set of individuals may not be informative for assignment when applied to the holdout individuals.

Although there was little difference observed between the three RF-based methods, in both datasets RF-selected panels had higher assignment accuracy in small- to medium-sized panels, while GRRF often outperformed other SNP-selection methods in the largest panels. This reduction in RF accuracy may be due to our applications of the RF approach. As we aggregated SNPs across five lists ranked by MDA, loci common across all lists at a lower rank may not be any more informative than those already included in the smaller panels, and will therefore contribute little to assignment accuracy. Conversely, GRRF continues to apply a penalty to SNPs regardless of panel size and thus selects SNPs that continue to contribute to the overall informativeness of the panel. We tested RRF and GRRF in addition to the basic RF approach to address the possible risk of node sparsity and to demonstrate the potential benefits of more stringent approaches. The easy implementation and customizable parameters for panel-size selection speak to the usability of these algorithms for subset selection. One additional benefit of GRRF is the customizable weighting of loci. We applied importance scores from a previous RF run to apply a non-uniform weight to the error penalty for each SNP. However, these scores could be manipulated to reflect additional information, such as location within known genes or importance to a phenotypic trait to allow for functional importance of loci to be considered in the SNP selection process. As such, we believe the comparison of all three approaches informs future use across genetic-based disciplines.

Sampling juveniles at spawning sites of anadromous fish increases the possibility of including siblings within the sample. Though this might inflate our estimates of self-assignment accuracy for Atlantic salmon, purging the dataset of siblings may actually reduce population estimates, depending on the severity of sibling removal (Waples and Anderson, 2017). The ideal threshold to remove individuals can be difficult to determine and varies for different systems and datasets (Waples and Anderson 2017). Further, this bias would be consistent across SNP selection methods, and does not detract from the benefits of machine-learning methods for SNP selection. The improved self-assignment accuracy obtained with RF methods for a larger sample of adult Chinook salmon (Larson et al., 2014a) demonstrate a wider range of the applicability of this approach.

We applied RF feature-selection to populations under a hierarchical genetic structure. Further tests of these methods may reveal that the applicability of RF is limited to highly structured populations under this type of hierarchical model. However, we demonstrate that within these populations of low differentiation (low pairwise F_{ST}), there is potential to develop these methods for further research. The resolution achieved using a single, small panel of SNPs for river-scale assignment offers new opportunities to improve fisheries management techniques. Ozerov et al. (2013) found that to distinguish populations of Atlantic salmon to a comparable (90%) accuracy, different sets of up to 150 SNPs were required to classify mixtures of individuals, depending on the populations in question. Although it is possible that there is some upward grading bias in our study, we applied the combined training-holdout and LOO method proposed by Anderson (2010) to reduce overestimation of self-assignment accuracy that might otherwise occur with relatively low sample sizes.

As we investigated overall assignment using a single panel at a time, we cannot be sure how each SNP in the subset distinguishes among individuals within a river. The low degree of overlap across RF runs (Table 2) indicates high variation in the RF ranking process. This is expected due to the randomness associated with considering subsets of features within each tree, but may be indicative of noise that must be filtered by the RF algorithm. Although the proportion of SNPs present in all 5 runs increases with increasing rank (Table 2), an adapted algorithm to increase consistency may also improve results. Though outside of the scope of the present study, investigating the potential for a deterministic approach could provide insight to the underlying genetic differentiation between certain populations and the process of feature ranking in RF. Our findings support the use of stringent applications of RF for feature-selection in a wildlife management context, such that a reduced panel may be established to allow for individual assignment to natal rivers. With this improvement in accuracy, these methods could be used to inform management policies to reduce exploitation of particular subpopulations. This study highlights the need for further investigation of machine-learning techniques, such as RF, that may be valuable for a range of ecological studies.

Chapter 3

Population and Landscape Genetics in Greater Labrador

The discovery of informative SNPs for accurate assignment of individuals within Lake Melville indicates that these populations are highly structured. Given this level of detectable structure for a small geographic area, populations may also be distinguishable across a broader scale. For a larger geographic area, assignment to spawning sites may not be achievable, but structure that is resolvable to general region may help to inform fishing practices and influence the monitoring and establishment of DUs. Identifying and preserving this genetic diversity can be crucial to managing fisheries harvests and ensuring the maintenance of the species. Due to the highly-structured nature of Atlantic salmon populations, we may expect that Lake Melville populations are differentiated from those in greater Labrador, particularly given differences in environmental conditions within the embayment.

In the previous chapter, we applied selected SNPs for individual assignment in Lake Melville. Here, we investigate the applicability of these SNPs for detecting overall structure across coastal Labrador. By targeting fewer molecular markers in individuals from an additional 24 rivers, we use amplicon-based sequencing to investigate the resolution that is achievable through a cost-effective approach for selecting SNPs.

As discussed in Chapter 1, the use of molecular tools for phylogeographic studies may identify environmental conditions that influence population structure. Understanding the contribution of neutral or adaptive factors that influence past and present population distribution allows the implementation of habitat-based conservation approaches, and may help predict future distribution in response to changing environmental conditions (Manel et al., 2003). We assess trends between population structure, climate variables, habitat conditions and geographic distance to uncover patterns that may indicate local adaptation to particular conditions, influencing the process of divergence.

3.1 Population Structure in Coastal Labrador

Atlantic salmon from Labrador often form a discrete genetic group relative to other regions of North America, though even this distinction is not consistent (Bradbury et al., 2014; Moore et al., 2014; King et al., 2001). Labrador populations have been shown to constitute a polyphyletic group with populations in Northern Newfoundland rivers, with no apparent pattern relating to geographic structure (Palstra et al., 2007; Verspoor et al., 2005), while more recent microsatellite analyses group Labrador populations with southeast Quebec (Moore et al., 2014; Dionne et al., 2008). Assessing these same populations using a SNP panel, Moore et al. (2014) found Labrador clustered into a discrete group, with some sites constituting their own cluster, depending on the subset of SNPs analysed. As research focused particularly on Labrador populations is limited, structure within the region has remained somewhat elusive but may be inferred through range-wide studies. Though only two Labrador sites were represented, a neighbour-joining (NJ) tree based on microsatellite data (King et al., 2001) aligns with that of other research separating Labrador into a distinct lineage, with further evidence showing greater divergence between north and south Labrador populations than any other within-region divergence, congruent with current SFA designations. A similar divergence was detected, separating Labrador into two clusters along a single principal component (PC) axis, though no geographic explanation for this split was made apparent (Bourret et al., 2013). In addition to genetic evidence of a north-south split within Labrador, differences in run-time, and possibly other life-history traits support the presence of distinct groups (Dempson et al., 2017). At a broader scope, the less likely fine-scale structure is to be detected, as between-region diversity is almost always greater than within-region diversity. Though this is to be expected, it does not indicate that within-region structure isn't present. On the contrary, that structure within Labrador can be even faintly detected when analysing across a broad scale, suggests that there exists differentiation that when inspected closely, may have important consequences for conservation.

3.1.1 Techniques to Identify Population Structure

Population structure may be detected through a variety of computational methods. While any clustering algorithm could theoretically be used to detect the number of discrete populations and the fidelity of individuals to those clusters, few software packages are used regularly in the field of population genetics. The most commonly used methods implement a Bayesian approach with Markov chain Monte Carlo algorithms, as applied in STRUCTURE (Pritchard et al., 2000) and GENELAND (Guillot et al., 2005), in which it is assumed that populations are independent. Often these approaches provide the option of including sampling locations as priors in determining population structure. Hierarchical structure and genetic relationships between populations may be calculated using dendrograms, often through distance-based measures (such as UPGMA (Unweighted Pair Group Method with Arithmetic Mean) or Neighbour-Joining (NJ) trees), or maximum-likelihood approaches. As leaves or terminal nodes must be known to establish relationships between groups, tree-based methods require prior clusters or populations to be established.

As an alternative, model-free clustering approaches such as k-means clustering, which aims to find an optimal k at a local maximum may be implemented through software such as GENODIVE (Meirmans and Van Tienderen, 2004). In k-means clustering, increasing values of k are tested and compared using a Bayesian information criterion (BIC) score. The optimal k or smallest BIC score indicates the optimal number of clusters without overfitting the model. Discriminant analysis of principal components (DAPC) uses k-means clustering on PCA-transformed data to determine the optimal number of clusters. A linear discriminant analysis is conducted on the retained PCs, maximizing variance between clusters, with the minimum variance within clusters (Jombart et al., 2010).

We test three methods for determining population structure to cover a diverse approach to uncovering structure revealed by microsatellites and selected SNPs, and to compare clustering patterns across methods. We apply hierarchical STRUCTURE (Pritchard et al., 2000) without location priors, NJ trees based on Cavalli-Sforza and Edwards chord distance (Cavalli-Sforza and Edwards, 1967), and DAPC.

Table 3.1: List of selected microsatellite loci and primer sequences

Locus Name	Forward Primer (5’-3’)	Reverse Primer (5’-3’)
NGS-SsaD486	TGCAGTCCAATAATATCCCCGT	CCCTGCATGACTCGGATAAC
NGS-SSsp2210	CACATTCACTGCAAAATAAAGCT	TGGGATTCAATAAAGGTAAGTAAGT
Ssa-1.5	GCGTTATGTGCTTGCATGC	ACCACCGTACTCAGCTTATCC
Ssa-1.7	AGAACACAACAGAACCAGGTAC	CTCGAACACACTTCCAACCC
Ssa-1.8	AGGCCAAAGAAATCCTGCAC	ACTGACCCAAACACGCAAATAG
Ssa-10.1	GGTCCTCCAGTACCTCCAAC	AATCTGGTGAGTTCGTCCGG
Ssa-10.4	GGTGAAATGTAGCCTGCATG	ACACACTGCTATATGTGTGG
Ssa-11.2	AAAGTTTGTTTGTGGACCGC	CGGACAGTTTCTTGGACTTC
Ssa-11.3	AGCGTGTGTGTCGTTCAATAC	ATGTTTCACCTCTGCGTCAC
Ssa-11.5	GTGTGCCGTTCTATCGCTG	CCTAAAGAAATGCCAGAGTCCG
Ssa-11.6	TTAACCTGCTCTACCTCTCG	ACATCACCACACCTATCTTC
Ssa-12.5	TCTCCTTCCTCGATCAGCTC	AATGTGTCGCCTTCCCACC
Ssa-14.2	GGGCATGATCTCGACACC	AGGAATGAGTAAGCTGGCTAAG
Ssa-14.6	AGTCAAGAAAGTCACTGCCC	GGAATGGCAAACAGAAAGGG
Ssa-15.1	TTTCTTTGTGTGTTGTGCCC	CAGCTGTGGTTCCTCTGGG
Ssa-15.3	GCTAACGAATGACAGCTTGC	CATTAGTAAGACTGGCAGCAG
Ssa-15.7	GATGTGATGGCAGTGCTATG	CAGCAACAAGGTCAATCTCC
Ssa-19.1	TGTGCAAACGCCATGATACC	CCATGACAGCTCCATCCGG
Ssa-19.2	GTGACCCAAAGTGCTGCTG	CTCCAGACACCAGCACCTC
Ssa-19.3	ACGTCCTGACAGTTATCCTTG	GTCTTGTCATGGCTGTGCTC
Ssa-2.1	AGACTCCACCTGCCTTGTTC	CTCACTGTCAGAGCATGCG
Ssa-2.2	TGGCCATTCTCCAGAGCTAG	CCACCAAAGGAGAGTACGTG
Ssa-2.7	CCCAGACTTCCCACTCTCTATG	GGACACAGAACCTTGAACGG
Ssa-20.2	TCTTCCCTCTTCTGCAGCAG	AGCTCTGGACACCACACTG
Ssa-23.2	GGTGGTTGTTTCTAGTGAGGG	GCACCTCTAAAGCACCATGG
Ssa-25.2	TGCAGGAAGACTCTGAAAGG	AGGTGGGTGTTGTACATCAG
Ssa-5.2	AACTTGCGTGATGATGTGGC	GCTGGCCATGTTCTTCTGTG
Ssa-6.2	GGAGAAGAGGAGATGGAACTTG	ACACCTGACAATACCACACC
Ssa-7.1	CCACTCCCACGAATGATGTTC	GGAGGCCACATTGCAGTC
Ssa-9.3	GCCAACCACCGTTAAACCTC	TCAGCAGTTCCCAATATTTCCC
Ssa-9.8	GCGTCGACTGCCATTCAAC	TGTCCTTGCTTTCTCCGTGG
Ssa-1.10	TGGATGACAACCTCCGTTAAAC	CGGGAAGCCTGGTGAAGATC
Ssa-1.11	CTCATCAACGCTATCCTCTTCC	GTCTTTCATCTGTCCGCGTG
Ssa-1.14	TCGTATTTGTCAAGGATGTGCC	AGATGCCCATTGTATTGCCC
Ssa-11.11	CGGCATATACCTTTAACGTTGG	GAAGAAGCGATGCGAGAGG
Ssa-11.12	CGTTAGCACACATGGCAAATC	GGTGCTGTTTGGGATGCATC
Ssa-12.12	TTGCTGCTGGTTTGTGCTC	GGGACAGTGAAGTGGTATTGC
Ssa-12.13	ATCAGGCTCAGAGGTGGAAC	ACACAGTGGAGGTAGAGATAGC
Ssa-13.10	TGAAAGTTGGCTGCAATCCG	GGAACCTGTCTGCCCACAC
Ssa-13.12	AGTTTGGCGTAGTCTGGGAC	TCCCATCATCCTCCTCTGGG
Ssa-14.10	GGGAACGTGTGGAAGATTCAC	AAGGTATGGAGGGTGATGCC
Ssa-14.9	CCATAATGGCACTGCTTCTTC	GTGTTGCTTCATTACACTCCG
Ssa-16.5	CCGCTGGATTCCTCATTATGTC	GGACTGACAGGAAGAGAGACC
Ssa-19.7	CTCCTTCACACAACCACC	AAGTGCAGACCTACCTTGTG
Ssa-19.9	TCTGGTGCTGACGATGAGAG	GAAATCAGAGGTCATTGGCCC
Ssa-2.12	CAGTACAGAAGCAGTCATCGC	ATTGTTTGCGGACGGTCATG
Ssa-2.13	GCTCAGATCGCAACCTTGAC	TCTAAACCGACCAGACCGAG
Ssa-21.10	ACTGCTTAGCTAGATTTGGCC	TCTACAGACAGGTGAACATGC
Ssa-21.3	TTGAACCTGAACTGGAATCCC	ACCGGCCAGTCTGAAACAG
Ssa-21.5	CACTCCCTAACTCCATGGTC	TCATGGATGTCGTCACTGTG
Ssa-23.10	TGATTGTGAACGGCTTTGGG	ACAAGCAAGCACCCTTTGTC
Ssa-23.3	GGAGAAGTGATTATGGTTGTGC	GGACAACGGGTTCTACATGG
Ssa-23.9	ACGGATACAGAGAGACGCAC	ACAGCGAGGAGGACAAAGTC
Ssa-24.9	CACTCCATCTATCATCTGTGCC	GATGAGGAGCAGAAGAGGCC
Ssa-25.3	TTCCCACTGGCCAAGAACTG	GACATTCCCTTGTGTTGATGAC
Ssa-27.7	TCATCAGTGTGGAGGGAATC	TCTATCTTCCTCTGGCCTGG
Ssa-3.10	GACTGCAACTAACTGAATGACG	TCCATCATCCCTTTCAGCTG
Ssa-3.9	CACCTCCAACTGCTCAATTAGG	GAGGCCCGTGTTTCTCAAC
Ssa-5.11	CAACCGCCGTTAAACATCATC	GAGGCCCGTGTTTCTCAAC
Ssa-6.11	CCGTGGAAAGCACTTAACATG	GAACGCATGTCATGGCCTC

Locus Name	Forward Primer (5’-3’)	Reverse Primer (3’-5’)
Ssa-9.10	TCCATTGTTCCCTCAGACCC	GGTAACATGAAGGAGAGCTGG
Ssa-10.2	TGATCCTCTTCACCACCCTG	CTGAAGACTCCTCCCTCACC
Ssa-11.8	AAAGGACCCAGAACGTACAG	ACCACACAGTACCCTCAATG
Ssa-13.8	TGACGAGACAAGATTCAGGTTG	GACCTATGCAACCACCAACG
Ssa-14.3	TCAACCTAAACCCTCTGCCC	AATCATCACATTCCACAGCAAC
Ssa-14.5	CCAGGAGGCCTTCACATG	CCTCCTGGCAATGCTGTATAG
Ssa-14.8	AAACATTGATTTGGCTCTGTC	TATTGCACCATCCCGTTCTC
Ssa-17.1	CATCTTCCGGTTCGCTCAAC	GTCATGACCTGTGCAACCAG
Ssa-18.7	TGCAGGTTGTGGTCATGTTG	CACATTCTGTCCATTCGGCC
Ssa-20.d56	GAGGTCAAGGTTTCCACTGG	TAGCTGCTCTCTGTTCTGGG
Ssa-21.2	CTGTCCAAATTGCAGGCTTG	GCCTAATTTGCCTACTCCTGTC
Ssa-22.2	AGTGGTTGCTTTGGTTCTCC	GGATAAAGCGGACCAAGACG
Ssa-22.5	GTGACGTCTGGAATTGTGAC	GATCCAATCAACACCGGTAG
Ssa-22.9	CAAATGCCACACGACCTGAC	GGTCAACCGCTCTGCATATAG
Ssa-22.d31	AGTTTAGTAGGGCCTGCGTG	ACATTCTTCTGTCACAGCCTG
Ssa-25.11	GGGTCCATGAGAAAGGCAAC	TGGGATCCACACCTGACAAC
Ssa-26.1	TCACGCATAACCTTAGACAACC	AATGCCAACCCTGTTACAGC
Ssa-4.d44	TTGGGTCTTAATGGCACCTG	GCTTTGGTTCCCTGAGAGTG
Ssa-5.6	GTGCAGCTGTTCCTCACTTC	GGGACAGGCGTAGAAATCG
Ssa-6.7	GCAAATCAGCATTCAGGGC	CAGCTGATCGAACTGAATGGG
Ssa-7.12	CACTCCCTGACACGTTAACAC	CACTTCCTGACAAACATGCAC
Ssa-9.13	ATCCACACCTCTCTTGCCAC	GATCACCATCGTTACCATCCC
Ssa-1.9	CTGAGGAGCACAAAGGACAG	GTGTTGCTGGCTGTGTTCTC
Ssa-13.2	CTACACCAAGAGTCCAGTGTC	ACAATTTGTCTCCCTGTTGTTG
Ssa-17.2	ACCCATAGAATTACTGCACTGG	GTCGTACTGGCATAATGTCAAC
Ssa-22.d40	GCACAGAGGTAAGAGTTCAGC	CTCTGCTGCTGTGGGTGG
Ssa-22.d41	CTCTGTGGTCTGGGTCCTC	ACCTCGTACCCATGCACATC
Ssa-22.d44	GTACCTTTGAACATGCACACG	CATCTCCACATGATAACGTTGC
Ssa-24.d09	ACCGTAAGCAGCATCACTTTAG	GTTTGGGCTGTCTGGTACTG
Ssa-24.d24	CTGCCAACACACACTGCC	TTTGACTCTTCCTGTATGTCGG
Ssa-26.d06	CATAATCACCTTGCATGACACC	CCTGCTGCACCGCTAAATAC
Ssa-27.d46	TGGCTGGTGGTTATAGGAGC	ACCATGCCAAGACAGTGATG
Ssa-28.d01	ATTACTGCCCTATCGCCATG	TCACCTTCTTCACACACGATG
Ssa-29.d18	AGCTACCTATTCCTGGAGCG	AGAGATGTTAGCGGGTCAGG
Ssa-29.d33	TAACTGCTGAGCCGTGTGTC	GCAGTGAATTCTATCTTCGTCG
Ssa-3.2	GTCACCAATACCACGTCACC	TCGTCAAGGGATGTGGTCAC
Ssa-5.8	ACACAGCTCTTATTTAACCGTC	GAAGGAATCTCACTCGTCTAAG
Ssa-7.d33	AGCATAGCATAGGAACAGACAC	AGCACATCCTGACCTCATCG
Ssa-7.d47	TGGAATTGGGTCAGCAGTTC	AGGACAGGGTTGAGATCAGC
Ssa-8.d04	ACTGTGTGGACTGGGAGATC	CAGCAGCGTTGTCTTGTACC
Ssa-8.d07	GGGTGTGAGGGAGGACTTAAC	TGCTAGCTACACTCCTGTCC

3.1.2 Methods: Panel Establishment and Characteristics

A panel of 101 microsatellites (Table 3.1) distributed across the Atlantic salmon genome was selected for genotyping in 1558 individuals across 35 sampling sites. Loci were selected based on MEGASAT scores (Zhan et al., 2016) with automatic genotype scoring, with the additional criteria of an approximately even distribution of loci across the genome. Loci were amplified in 10-locus multi-plex PCRs. Multiplex PCRs were performed using Qiagen (Toronto, Ontario, Canada) Type-IT 29 Mastermix (1.75 μL), 0.2 μM each oligonucleotide and 0.7 μL genomic DNA. PCRs were conducted on

Eppendorf (Hamburg, Germany) Mastercycler ep384 PCR machines using the following parameters: 94°C for 15 min, followed by 20 cycles of 94°C for 30s, 57°C for 180s, 72°C for 60s, with a final extension at 68°C for 30 min.

Indexing sequences were added to the PCR products using a second PCR. The index PCR used oligonucleotides composed of Illumina annealing adapter sequences, a 6b barcode and the Illumina sequencing primers. Indexing PCRs were performed in 5 µL total volume with 0.25 U *Taq* DNA polymerase (New England Biolabs, Ipswich, MA, USA), 0.5 µL Thermopol 109 buffer (NEB), 0.2 mM each dNTP, 0.2 µM Index_1 oligo, 0.2 µM Index_2 oligo and 0.3 µL of 20-fold diluted multiplex-PCR product. Cycling parameters were as follows: 95°C for 2 min, followed by 18 cycles of 95°C for 20s, 60°C for 60s, 72°C for 60s with a final extension at 72°C for 10 min. Indexed PCR products were pooled and cleaned using Ampure XP (Beckman Coulter, Pasadena CA, USA) or Sera-Mag Speedbeads (GE Healthcare, Little Chalfont, UK) magnetic beads (1.8:1 bead:DNA library ratio). Libraries were quantified using Kapa (Wilmington, MA, USA) Library Quantification for Illumina on a Roche (Basel, Switzerland) LC480 qPCR instrument following manufacturers' protocols. Libraries were sequenced at 10–12 pM concentration using MiSeq v3 chemistry with 150 cycles in one direction and dual indexing. Indexed individuals were demultiplexed with the MiSeq sequence analysis software. After filtering at a maximum of 40% missing data, 1485 individuals remained in the data set for downstream analysis.

As described in Chapter 2, a preliminary GRRF run was applied to select SNPs from a filtered panel of 93,703 loci. SNPs were selected based on their informativeness for individual assignment within Lake Melville, as applied in the previous chapter, to select a candidate panel for assessing the structure of greater coastal Labrador. At a gamma value (see description of the RRF and GRRF algorithm and parameters in section 2.1.2) of 0.1, a panel of 443 SNPs was selected. Amplicon-based sequencing (Table 3.1) was conducted in 1559 individuals across 35 sites (Fig. 1.2). The protocol was implemented as described for microsatellites, using five PCR multiplexes. Of these 443 targeted amplicons, after filtering for quality and MAF of 0.05, 557 SNPs were successfully genotyped across 381 amplicons. After sequence alignment, the single SNP associated with the original Cigene SNP chip was retained per locus. After comparing allele frequency between samples in both the original data set (analysed in Chapter 2) and the amplicon data set, SNPs outside of two standard deviations of the linear trend line were removed, resulting in a final panel of 376 SNPs across 1389 samples (Table 3.2).

3.1.3 Method Application

Nei's pairwise population F_{ST} was calculated for both molecular marker sets using the R package 'hierfstat' (Goudet and Jombart, 2015) (Table 3.3). STRUCTURE (Pritchard et al., 2000) was run using a burn-in of 100,000 and 500,000 iterations for both the amplicon and microsatellite data for K = 1-35. Often, STRUCTURE runs conducted across the entire data set result in an optimal k=2 (Janes et al., 2017). Known as the K = 2 conundrum, investigators are encouraged to conduct hierarchical structure analyses, in which each of the two clusters are tested for finer-scale structure through subsequent STRUCTURE runs. When our analyses resulted in an optimal K = 2, we re-ran STRUCTURE using a burn-in of 500,000 and 1,000,000 iterations for K = 1 - n+1 (where n is equal to the number of sampling sites within the subset) up to 3 levels of hierarchy. Mixed populations were assigned to 1 of 2 clusters at a criterion of at least 50% of individuals having a q-value greater than 0.5 for that cluster. For SNP data, as high rates of admixture made splitting the data for hierarchical analysis difficult, we also inspected alternative optimal values of K. Bar plots for the determined optimal K were created in Excel 2016.

NJ trees were created using Populations v. 1.2.31 (Langella, 1999), and edited using FigTree v. 1.4.2 (Rambaut, 2006), using 1000 bootstrap replications.

DAPC was run using the R package 'adegenet' (Jombart et al., 2008). For the microsatellite and amplicon data, 500 and 400 PCs, and 2 and 4 linear discriminants were retained, respectively. The optimal k for each data set was determined through BIC plots. DAPC scatterplots were created in R v. 3.3.2; NJ trees and DAPC clustering figures were created in GenGIS v. 2.5 (Parks et al., 2013).

Table 3.2: Site ID, location and sample size for both datasets used for Chapter 2 analyses.

River Name	Site ID	Latitude(N)	Longitude(W)	Sample Size (SNP)	Sample Size (Microsats)
Alexis River	ALX	52°6'	56°53'	29	29
Big River	BIG	54°84'	58°94'	27	29
Cape Caribou River	CCR	53°32'48,8"	60°36'270"	39	39
		53°34'26,76"	60°42'2960"		
		53°37'166"	60°25'594"		
Caroline Brook	CAR	53°15'232"	60°31'899"	17	25
Charles River	CHA	52°23'	55°84'	37	50
Crooked River	CRO	53°50'991"	60°48'863"	51	50
		53°52'768"	60°49'946"		
Double Mer	DBM	54°02'296"	59°65'24"	50	50
Eagle River	EAG	53°53'333"	57°46'67"	44	50
English River	ENG	53°53'1519"	58°50'39'71"	58	59
Forteau River	FOR	51°48'467"	56°94'48"	41	50
Hunt River	HUN	55°56'836"	60°66'97"	46	46
Kenamu River	KMU	52°50'952"	60°08'279"	18	18
Kenemich River	KEN	53°31'864"	59°82'05"	30	30
L'anse au Loup River	LLO	51°52'628"	56°81'68"	34	50
Main Brook	MNB	54°04'355"	57°52'374"	42	42
		54°04'189"	57°52'306"		
Mary's Harbour	MAH	52°31'317"	55°82'43"	38	50
Middle Bay Brook	MID	54°41'463"	58°09'15"	49	49
Muddy May Brook	MBB	53°64'	57°07'	50	50
Mulligan River	MUL	53°52'138"	60°05'392"	43	47
Paradise Brook	PAB	53°42'305"	57°23'72"	42	41
Paradise River	PAR	53°42'408"	57°24'95"	39	40
Partridge Point	PPT	54°09'918"	59°47'51"	50	48
Peter's River	PTR	53°20'104"	60°47'153"	51	46
		53°20'345"	60°37'293"		
Pinware River	PIN	51°63'	56°69'	50	49
Port Marnum	POM	52°4'	55°74'	33	31
Pottle's Bay	POT	54°48'043"	57°72'98"	13	13
Red Wine River	RWR	53°52'764"	61°27'976"	50	50
		53°52'928"	61°28'730"		
Sand Hill River	SAH	53.56'667"	56°35'	48	47
Sebaskachu River	SEB	53°47'397"	60°08'523"	30	30
		53°46'10"	60°10'575"		
Shinny's River	SHI	52°59'	56°34'	50	50
St. Lewis River	SLR	52°44'	56°19'	27	48
Susan River	SUS	53°44'365"	61°3'275"	34	50
		53°44'184"	61°02'216"		
Tom Luscombe River	TLU	54°34'106"	58°55'04"	50	50
Traverspine River	TSP	53°08'853"	60°27'769"	48	48
		53°07'319"	60°30'380"		
West Brook	WST	54°39'706"	58°10'31"	31	31
	TOTAL			**1389**	**1485**

Table 3.3: Nei’s pairwise population F_{ST} for microsatellite data (bottom left) and SNP data (top right). Colour intensity indicates magnitude of $F_{ST.}$

	ALX	FOR	HUN	KMU	KEN	LLO	MNB	MAH	MID	MBB	MUL	BIG	PAB	PAR	PPT	PTR	PIN	POM	POT	RWR	SAH	SEB	CAR	SHI	SLR	SUS	TLU	TSP	WST	CCR	CHA	CRO	DBM	EAG	ENG
ALX	-	0.047	0.038	0.066	0.053	0.051	0.039	0.032	0.031	0.028	0.075	0.041	0.036	0.027	0.067	0.054	0.020	0.035	0.032	0.047	0.021	0.073	0.050	0.038	0.029	0.055	0.029	0.064	0.033	0.063	0.028	0.039	0.055	0.025	0.033
FOR	0.031	-	0.048	0.069	0.070	0.030	0.054	0.037	0.039	0.040	0.073	0.048	0.046	0.040	0.076	0.069	0.028	0.042	0.036	0.060	0.042	0.080	0.064	0.048	0.045	0.066	0.044	0.078	0.046	0.063	0.030	0.052	0.069	0.038	0.048
HUN	0.043	0.045	-	0.038	0.049	0.054	0.044	0.042	0.037	0.041	0.059	0.029	0.043	0.039	0.068	0.047	0.033	0.046	0.031	0.046	0.042	0.067	0.037	0.055	0.042	0.048	0.037	0.052	0.035	0.038	0.037	0.028	0.055	0.032	0.031
KMU	0.043	0.051	0.027	-	0.055	0.077	0.062	0.068	0.055	0.056	0.083	0.068	0.061	0.060	0.085	0.049	0.050	0.075	0.085	0.043	0.054	0.099	0.039	0.068	0.076	0.050	0.055	0.036	0.066	0.058	0.064	0.040	0.072	0.052	0.050
KEN	0.035	0.045	0.040	0.049	-	0.074	0.048	0.058	0.047	0.050	0.071	0.062	0.060	0.050	0.089	0.053	0.047	0.063	0.051	0.041	0.046	0.086	0.046	0.066	0.058	0.058	0.044	0.054	0.055	0.058	0.056	0.042	0.070	0.045	0.046
LLO	0.030	0.025	0.040	0.046	0.046	-	0.058	0.035	0.043	0.040	0.080	0.059	0.045	0.047	0.081	0.068	0.031	0.042	0.041	0.061	0.042	0.088	0.065	0.046	0.048	0.067	0.049	0.078	0.048	0.072	0.034	0.055	0.068	0.040	0.050
MNB	0.029	0.038	0.040	0.049	0.040	0.036	-	0.049	0.033	0.044	0.041	0.046	0.046	0.039	0.069	0.060	0.041	0.046	0.042	0.050	0.037	0.051	0.060	0.055	0.045	0.058	0.043	0.076	0.041	0.063	0.050	0.054	0.053	0.048	0.041
MAH	0.021	0.029	0.033	0.042	0.034	0.030	0.031	-	0.027	0.029	0.070	0.043	0.033	0.032	0.067	0.058	0.019	0.025	0.022	0.056	0.025	0.080	0.056	0.030	0.030	0.056	0.032	0.070	0.029	0.062	0.016	0.046	0.052	0.028	0.040
MID	0.027	0.031	0.037	0.049	0.031	0.033	0.030	0.022	-	0.023	0.062	0.032	0.035	0.027	0.060	0.053	0.017	0.026	0.015	0.050	0.024	0.067	0.050	0.035	0.029	0.049	0.017	0.066	0.013	0.051	0.018	0.040	0.044	0.030	0.031
MBB	0.032	0.039	0.045	0.055	0.042	0.038	0.030	0.030	0.032	-	0.063	0.036	0.026	0.020	0.061	0.060	0.026	0.027	0.023	0.048	0.024	0.068	0.063	0.038	0.027	0.054	0.036	0.071	0.031	0.061	0.026	0.047	0.053	0.030	0.036
MUL	0.036	0.049	0.036	0.047	0.049	0.047	0.029	0.042	0.043	0.056	-	0.069	0.064	0.062	0.096	0.073	0.061	0.071	0.057	0.065	0.066	0.052	0.066	0.074	0.070	0.078	0.061	0.080	0.059	0.073	0.068	0.062	0.079	0.063	0.068
BIG	0.038	0.034	0.035	0.044	0.038	0.032	0.038	0.024	0.030	0.037	0.036	-	0.045	0.036	0.066	0.056	0.031	0.045	0.039	0.050	0.037	0.079	0.052	0.051	0.047	0.056	0.029	0.066	0.038	0.065	0.035	0.034	0.052	0.036	0.027
PAB	0.020	0.040	0.029	0.043	0.041	0.038	0.040	0.029	0.031	0.037	0.043	0.027	-	0.025	0.067	0.059	0.023	0.035	0.030	0.051	0.031	0.077	0.055	0.040	0.033	0.060	0.036	0.070	0.036	0.056	0.029	0.049	0.058	0.026	0.041
PAR	0.027	0.036	0.042	0.054	0.039	0.034	0.032	0.028	0.027	0.035	0.040	0.037	0.033	-	0.060	0.054	0.016	0.028	0.025	0.046	0.024	0.071	0.052	0.037	0.028	0.056	0.029	0.067	0.027	0.057	0.022	0.040	0.060	0.019	0.035
PPT	0.052	0.065	0.071	0.075	0.063	0.072	0.064	0.058	0.051	0.063	0.079	0.060	0.060	0.064	-	0.086	0.051	0.065	0.046	0.087	0.066	0.101	0.085	0.079	0.065	0.078	0.061	0.099	0.052	0.076	0.057	0.071	0.071	0.060	0.065
PTR	0.048	0.065	0.049	0.052	0.046	0.058	0.049	0.047	0.052	0.059	0.056	0.045	0.056	0.052	0.090	-	0.056	0.062	0.048	0.047	0.056	0.080	0.052	0.067	0.058	0.053	0.067	0.065	0.056	0.050	0.064	0.049	0.079	0.059	0.055
PIN	0.019	0.029	0.037	0.043	0.033	0.030	0.030	0.016	0.020	0.027	0.042	0.025	0.028	0.023	0.053	0.052	-	0.019	0.024	0.043	0.018	0.064	0.053	0.028	0.019	0.045	0.028	0.070	0.026	0.053	0.021	0.045	0.041	0.027	0.030
POM	0.030	0.035	0.041	0.065	0.050	0.034	0.038	0.024	0.030	0.040	0.046	0.033	0.039	0.028	0.063	0.057	0.026	-	0.025	0.055	0.025	0.080	0.063	0.033	0.030	0.063	0.031	0.076	0.031	0.072	0.020	0.047	0.052	0.032	0.042
POT	0.032	0.031	0.039	0.083	0.040	0.028	0.030	0.021	0.021	0.033	0.044	0.039	0.039	0.033	0.044	0.050	0.020	0.041	-	0.038	0.019	0.070	0.057	0.024	0.033	0.050	0.028	0.059	0.027	0.090	0.026	0.038	0.039	0.028	0.025
RWR	0.034	0.044	0.030	0.037	0.032	0.040	0.034	0.032	0.031	0.041	0.037	0.030	0.038	0.033	0.071	0.039	0.032	0.041	0.037	-	0.046	0.071	0.039	0.060	0.049	0.038	0.049	0.048	0.048	0.040	0.048	0.023	0.072	0.041	0.048
SAH	0.023	0.033	0.035	0.045	0.034	0.030	0.032	0.017	0.022	0.030	0.041	0.029	0.029	0.024	0.055	0.046	0.019	0.024	0.028	0.030	-	0.064	0.051	0.030	0.021	0.054	0.023	0.064	0.025	0.051	0.018	0.042	0.055	0.027	0.034
SEB	0.047	0.055	0.049	0.072	0.061	0.054	0.037	0.047	0.050	0.058	0.030	0.048	0.056	0.049	0.077	0.067	0.049	0.057	0.055	0.047	0.052	-	0.079	0.075	0.073	0.079	0.066	0.087	0.069	0.090	0.074	0.069	0.087	0.070	0.069
CAR	0.033	0.037	0.023	0.003	0.021	0.033	0.025	0.028	0.030	0.036	0.020	0.021	0.021	0.031	0.063	0.030	0.028	0.032	0.026	0.012	0.024	0.034	-	0.066	0.059	0.045	0.059	0.049	0.062	0.045	0.060	0.037	0.070	0.057	0.045
SHI	0.029	0.036	0.047	0.050	0.041	0.037	0.039	0.023	0.027	0.039	0.050	0.031	0.038	0.034	0.066	0.058	0.027	0.031	0.027	0.042	0.025	0.056	0.036	-	0.030	0.062	0.038	0.079	0.040	0.063	0.025	0.053	0.060	0.042	0.048
SLR	0.023	0.034	0.038	0.048	0.032	0.029	0.035	0.020	0.024	0.033	0.039	0.031	0.031	0.027	0.051	0.048	0.022	0.030	0.027	0.035	0.023	0.048	0.032	0.025	-	0.056	0.027	0.067	0.032	0.072	0.025	0.046	0.051	0.030	0.038
SUS	0.015	0.023	0.007	-0.03	0.002	0.020	0.010	0.016	0.018	0.018	0.009	0.004	0.005	0.019	0.033	0.016	0.017	0.014	0.006	0.004	0.013	0.015	0.008	0.024	0.012	-	0.048	0.053	0.053	0.058	0.055	0.035	0.060	0.050	0.046
TLU	0.022	0.031	0.037	0.047	0.031	0.030	0.030	0.022	0.017	0.034	0.040	0.028	0.031	0.027	0.057	0.053	0.021	0.030	0.022	0.034	0.023	0.051	0.028	0.028	0.022	0.018	-	0.072	0.026	0.061	0.031	0.048	0.036	0.037	0.030
TSP	0.035	0.044	0.039	0.036	0.035	0.042	0.037	0.035	0.039	0.044	0.049	0.034	0.041	0.042	0.079	0.043	0.037	0.043	0.036	0.030	0.033	0.056	0.022	0.044	0.038	0.011	0.042	-	0.065	0.034	0.074	0.051	0.085	0.068	0.062
WST	0.023	0.028	0.035	0.053	0.045	0.030	0.026	0.019	0.010	0.031	0.038	0.028	0.029	0.027	0.046	0.049	0.019	0.032	0.019	0.033	0.022	0.049	0.027	0.022	0.019	0.010	0.013	0.036	-	0.070	0.031	0.045	0.038	0.035	0.031
CCR	0.045	0.050	0.025	0.058	0.051	0.050	0.048	0.043	0.051	0.055	0.052	0.032	0.050	0.054	0.080	0.052	0.047	0.061	0.066	0.037	0.047	0.070	0.005	0.052	0.046	-0.02	0.050	0.034	0.053	-	0.068	0.044	0.068	0.057	0.040
CHA	0.017	0.030	0.027	0.046	0.040	0.031	0.034	0.017	0.025	0.033	0.036	0.020	0.027	0.028	0.058	0.054	0.020	0.024	0.028	0.036	0.022	0.047	0.023	0.027	0.022	0.008	0.022	0.042	0.024	0.047	-	0.045	0.046	0.029	0.031
CRO	0.054	0.051	0.038	0.036	0.045	0.048	0.046	0.043	0.048	0.055	0.041	0.040	0.040	0.052	0.076	0.050	0.046	0.053	0.054	0.028	0.044	0.056	0.014	0.056	0.047	-0.01	0.046	0.042	0.047	0.025	0.039	-	0.056	0.041	0.039
DBM	0.042	0.049	0.056	0.066	0.050	0.051	0.041	0.043	0.032	0.051	0.054	0.045	0.051	0.055	0.052	0.074	0.040	0.054	0.036	0.051	0.043	0.061	0.048	0.050	0.042	0.029	0.034	0.063	0.030	0.071	0.044	0.064	-	0.054	0.054
EAG	0.020	0.030	0.031	0.039	0.029	0.030	0.029	0.021	0.023	0.028	0.040	0.025	0.025	0.025	0.057	0.044	0.016	0.029	0.024	0.028	0.022	0.045	0.022	0.031	0.026	0.014	0.025	0.032	0.021	0.039	0.025	0.042	0.041	-	0.035
ENG	0.028	0.036	0.028	0.040	0.037	0.036	0.034	0.025	0.027	0.037	0.041	0.021	0.035	0.031	0.062	0.050	0.027	0.032	0.025	0.034	0.028	0.048	0.027	0.033	0.025	0.018	0.027	0.038	0.024	0.048	0.028	0.043	0.044	0.027	-

3.1.4 Results: Microsatellites

Hierarchical STRUCTURE analysis on the microsatellite data determined an optimal K = 2 at all levels (Fig. 3.1). At the broadest level, sites were split into a west Lake Melville group, with the additional northern site Hunt River, and a second cluster containing sites outside of Lake Melville, as well as sites located at the mouth of Lake Melville, or on the east-northern shore. Hierarchical structure revealed a further division separating Sebaskachu River, Main Brook, Mulligan, and 23 individuals from Peter's River from the rest of Lake Melville populations. Of the former three, Main Brook formed a separate group from Mulligan River and Sebaskachu River in the next level of STRUCTURE runs. Within the west Lake Melville cluster, Peter's River was divided into two groups, in which individuals caught in the upstream sampling site (Table 3.2) cluster separately from those caught near the river mouth, consistent with patterns of mis-assignment in Chapter 2. Populations from the north shore of Lake Melville (Double Mer and Partridge Point) clustered separately from rivers located near the mouth, or outside of Lake Melville in the second level of hierarchy (Fig. 3.1). These rivers each formed their own distinct cluster in the next level of STRUCTURE runs, with Forteau River and L'anse au Loup, the southern-most sites, forming a discrete group from the rest of Labrador populations.

Using DAPC, resolution was not as fine as the hierarchical STRUCTURE approach, however delineations in clusters were consistent. An optimal K = 3 was identified in the microsatellite DAPC analysis (Fig. 3.2A), separating west Lake Melville, north shore Lake Melville populations, and populations outside of Lake Melville into distinct groups. The proportion of individuals from each site assigned to a given cluster (Fig. 3.2B) revealed a clinal pattern in admixture, where mixing increased with latitude. Like the STRUCTURE results, DAPC showed an association between Hunt River and the west Lake Melville cluster. This pattern was also evident in the NJ tree (Fig. 3.2B), which clustered west Lake Melville and Hunt River together, distinct from all other populations. West Lake Melville and northern Labrador appeared more closely related compared with west Lake Melville and southern Labrador. North shore Lake Melville

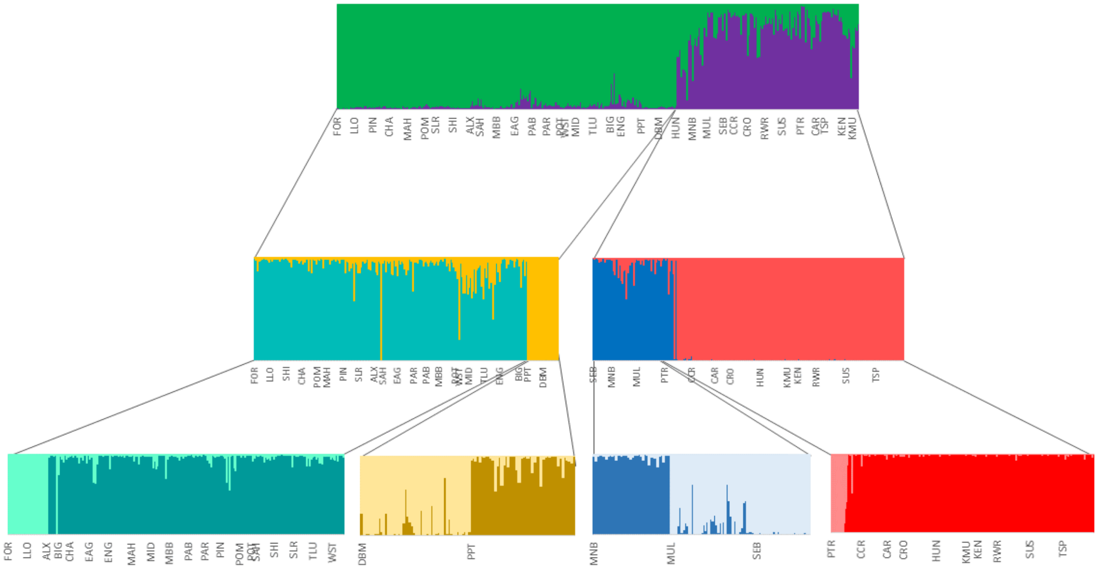

Figure 3.1: Hierarchical STRUCTURE (Pritchard et al., 2000) analyses barplots of q-values using microsatellite data. All analyses resulted in an optimal K = 2.

populations clustered with both branches, consistent with admixture shown with DAPC, and with splits in the hierarchical STRUCTURE analysis. However, bootstrap values for deep splits were quite low, thus failing to provide strong genetic evidence of these relationships. Within Lake Melville, our NJ tree showed high bootstrap support for branches containing Caroline River and Traverspine River, and Crooked River and Red Wine river, the same rivers that showed high rates of paired mis-assignment in Chapter 2.

3.1.5 Results: SNPs

STRUCTURE (Pritchard et al., 2000) resulted in an optimal K = 2 or K = 11 (Fig. 3.3), however both showed a high degree of admixture across populations. With K = 2, we observed the same split, generally, as observed with microsatellite data at the broadest level of K = 2. There was one cluster consisting of west Lake Melville populations, north shore populations (Mulligan River, Main Brook and Sebaskachu River) and northern-most sites, and a second cluster containing sites south of Lake Melville, as well as some located along the north shore and near the mouth of the embayment. With microsatellites, only Hunt River was found to cluster with this group. With SNPs, Hunt River showed the

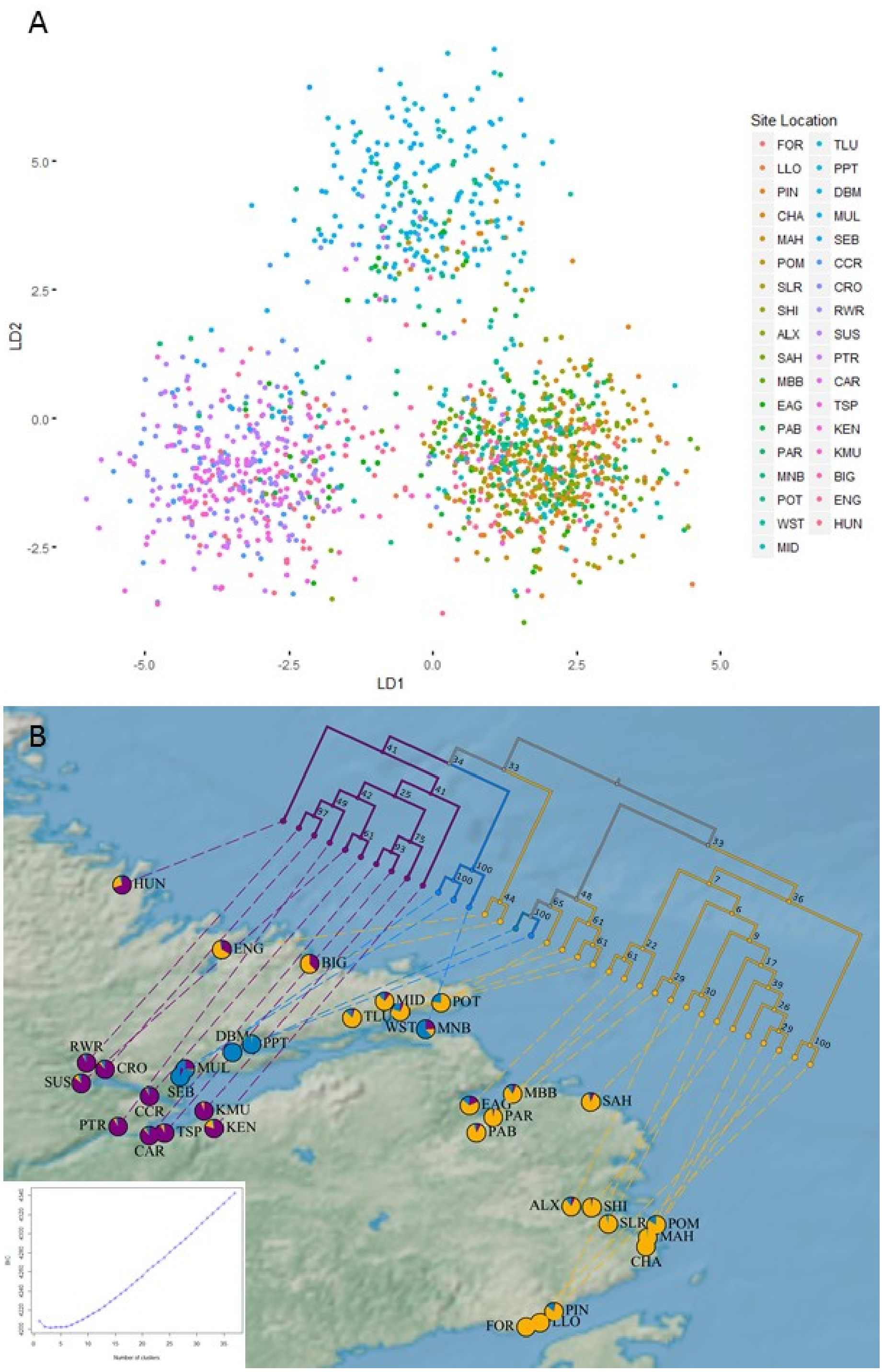

Figure 3.2: Population structure detected by microsatellite analysis as indicated by scatter plot (A) showing three distinct clusters from DAPC analysis. Pie graphs in (B) indicate proportion of individuals from each sampling site associated with each cluster. The neighbour-joining cladogram indicates hierarchical relationships, coloured according to the DAPC assignment of majority at a given site. Node labels indicate bootstrap values. Colours approximately correlate across (A) and (B). BIC plot indicating optimal K in the DAPC analysis is in the bottom left of (B).

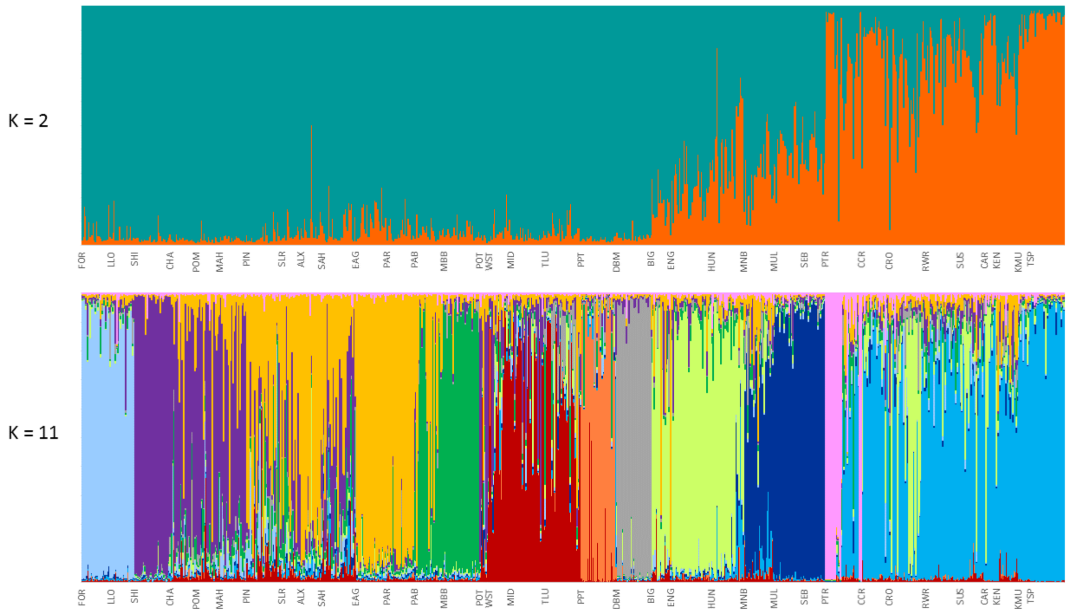

Figure 3.3: STRUCTURE analyses of SNP data showing optimal K = 2 or K = 11. X-axes for both plots are identical for direct comparison.

highest association, but all three sites located north of Lake Melville showed a high degree of admixture with the Lake Melville cluster.

Because of the high degree of admixture across the two clusters, we also inspected an alternative, K = 11. In this case, there was a distinct cluster of southernmost sites (Forteau River and L'anse au Loup River), and three other clusters of rivers south of Lake Melville. Shinny's River and Charles River, as well as Eagle River and Paradise River were somewhat distinct, with high admixture in Port Marnum, Mary's Harbour, Pinware River, St. Lewis River, Alexis River and Sandhill River between these three southern groups. Paradise Brook and Muddy Bay Brook formed a fourth group of southern sites, with high admixture with the Eagle River and Paradise River group. A discrete group of populations near the mouth of Lake Melville, consisting of Pottle's Bay, West Brook, Middle Bay Brook and Tom Luscombe River also showed high rates of admixture with other clusters. Partridge Point and Double Mer each comprised a distinct group, with other north shore populations (Main Brook, Mulligan River and Sebaskachu River) forming a separate cluster. Northern-most rivers clustered together, with the highest rate of admixture seen in Hunt River, mostly with the Lake Melville group, as observed in our microsatellite analyses. Lastly, west lake Melville populations formed a

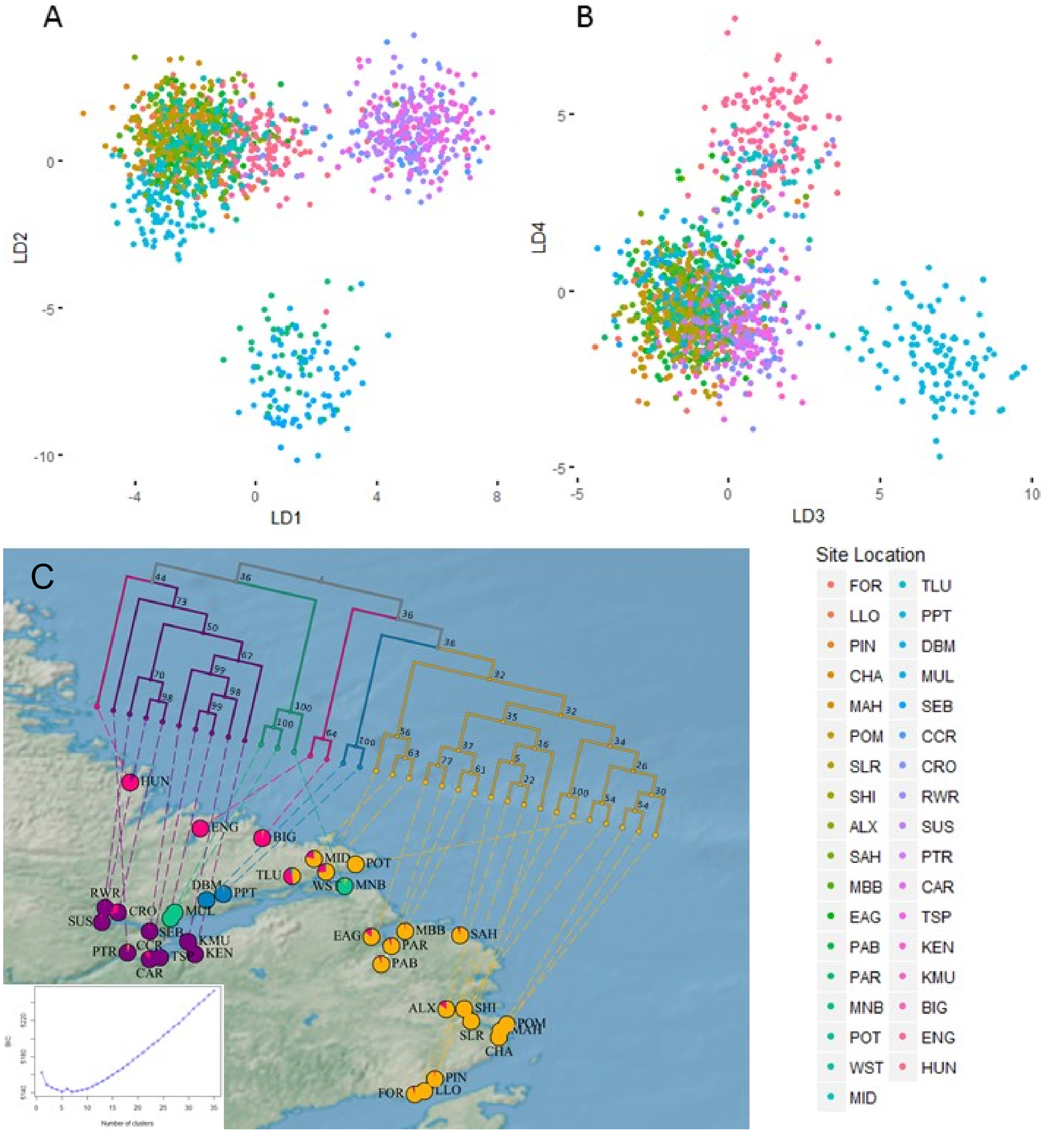

Figure 3.4: Population structure detected by SNP analysis as indicated by a scatter plot of (A) linear discriminants 1 and 2 and (B) 3 and 4, showing 5 clusters from DAPC analysis. Pie graphs in (C) indicate proportion of individuals from each sampling site associated with each cluster. The neighbour-joining tree indicates hierarchical relationships, coloured according to the DAPC assignment of majority at a given site. Node labels indicate bootstrap values. Colours approximately correlate across A and B with C. BIC plot indicating optimal K in the DAPC analysis is in the bottom left of (C).

distinct group with high admixture with the northern cluster, particularly in Crooked River and Susan River. Here, Peter's River also grouped into two separate clusters, with individuals from the sampling site closest to the river mouth clustering with other Lake Melville populations. Though there was a high rate of admixture, or uncertain clustering within some populations, general findings from this STRUCTURE analysis were consistent with the microsatellite analyses.

The DAPC BIC plot showed an optimal K = 5, though only four clusters were clear through the scatter plot of linear discriminants 1 and 2, and 3 and 4 (Fig. 3.4). These clusters generally consisted of west Lake Melville, northern Labrador, north shore Lake Melville (north shore 1), and southern Labrador. Less clear in the scatter plot was a fifth cluster containing two other populations of north shore Lake Melville rivers, Double Mer and Partridge Point (north shore 2). The distinction between these two groups of north shore Lake Melville populations is consistent with structure observed with microsatellite analysis, and a deep split in the NJ tree. Though the DAPC analysis grouped Hunt River with other northern Labrador sites, the NJ tree, as with the microsatellite data, grouped Hunt River with west Lake Melville populations. With both the microsatellite and SNP data, however, there was a clear signal of admixture between Hunt River and west Lake Melville, regardless of actual cluster assignment. With the SNP data, other northern populations (Big River and English River) appeared more closely related to south Labrador populations than west Lake Melville and north shore 1 populations. However, bootstrap values for deep splits were again quite low (ranging from 32 to 44%). Again, there was strong support for pairwise relationships within Lake Melville as seen in Chapter 1 and the microsatellite NJ tree. Also consistent with the microsatellite NJ tree was the clinal pattern of admixture in southern Labrador populations, increasing with latitude. Although we did not observe a high degree of admixture in southern populations, relative to that found in the STRUCTURE analysis, this is likely due to the grouping of southern populations into a single cluster using DAPC, as the observed admixture was generally limited to this region.

3.1.6 Discussion

Both SNP and microsatellite marker panels revealed clear and largely congruent patterns of population structure across Labrador. Using STRUCTURE, DAPC and NJ trees, we found distinct evidence of differentiation between Lake Melville, particularly west Lake Melville, and all other Labrador populations, across all methodologies. Further, all analyses showed evidence of differentiation within Lake Melville, with distinct splits between west Lake Melville and north-shore populations, and further potential genetic differentiation within the north shore of Lake Melville. STRUCTURE analyses showed splits between these groups at the second level of microsatellite hierarchy and in the K = 11 SNP analysis. DAPC clustering with both SNPs and microsatellites indicated the presence of at least one north-shore Lake Melville group. Further, north shore populations diverge from deep splits with NJ trees, and have the highest reported pairwise F_{ST} values.

Pairwise population patterns within Lake Melville were consistent with rates of incorrect assignment in Chapter 2. Although this was not surprising for the SNP data as these were originally selected for informativeness of population structure within Lake Melville sites, we found further support of these relationships using microsatellites.

Here we find clear evidence of genetic differentiation of Lake Melville Atlantic salmon populations. That these findings are consistent across molecular marker types, particularly when assessing SNPs selected for resolving population structure at a finer scale, demonstrates the high power of both marker panels to differentiate population structure and the high degree of information inferred by RF-selected SNPs. While this may suggest altering management strategies to conserve this genetic diversity, the ecological importance and source of this diversity is unknown.

3.2 Landscape Genetics

3.2.1 Techniques to Identify Landscape Associations

Landscape genetics aims to identify underlying genetic patterns of populations or sub-populations that are influenced by variation in environmental and geographic characteristics (Rellstab et al. 2015). Environmental association analysis (EAA) refers to the general statistical approaches involved in identifying these correlations, based on the principle that local adaptation evolves due to environmental or ecological heterogeneity across a species or population range (Lotterhos and Whitlock, 2015). Effective EAAs identify genetic variation that exists across populations due to adaptation while accounting for neutral genetic effects.

Many approaches identify outlier loci based on F_{ST} values as potentially adaptive (Coop et al. 2010). While F_{ST}-based outlier methods have been widely used (Zueva et al., 2014; Lotterhos and Whitlock, 2014; Foll and Gaggiotti, 2008), outlier methods alone do not incorporate an underlying neutral genetic structure into the analysis. Furthermore, measures of F_{ST} rely on *a priori* knowledge of subpopulation structuring, necessitating additional analyses using software such as STRUCTURE (Pritchard et al., 2000) or relying on potentially over-simplified grouping of individuals based on geographic location (Zueva et al., 2014).

To date, most studies (e.g. Hecht et al., 2015; Zueva et al., 2014; Bourret et al., 2013) use several statistical approaches to accommodate for underlying neutral genetic structure and demographic characteristics (Hecht et al., 2015; Bourret et al., 2013). While potentially effective, these approaches may suffer due to an accumulation of underlying assumptions in sequential analyses and have been criticized as potentially lacking in power, and may be biased due to early selection of a subset of loci (Frichot et al., 2013). This may include using software such as STRUCTURE (Pritchard et al., 2000) to provide an approximation of the number of subpopulations occupying the area of interest, or methods to identify and reduce environmental variables with which to investigate genetic association (e.g. principal component analysis (PCA)). Then, using F_{ST}-based methods outlier loci are identified from the whole array (e.g. Ferchaud and Hansen, 2016; Bradbury et al., 2014; Zueva et al., 2014). From here further analysis may include a method of incorporating underlying neutral genetic structure into the test for environmental association.

A variety of statistical approaches may be used to conduct an association analysis with these prepared data. Logistic approaches have been developed to incorporate spatial autocorrelation as a proxy for neutral genetic data (Rellstab et al., 2015). Both the spatial analysis method (SAM) (Joost et al., 2007) and generalized estimating equations (GEEs) test the association between a single environmental factor and the presence or absence of a certain allele. This is an obvious limitation for studies that wish to incorporate a variety of environmental variables. Furthermore, spatial autocorrelation as a proxy for neutral genetic structure may not detect true underlying structure, resulting in a high false positive rate (Rellstab et al., 2015). Joost et al. (2007) suggested that spatial autocorrelation methods may be most appropriate for population studies across a broad geographic scale.

Bayesian, F_{ST}-based methods such as BayEnv (Coop et al., 2010) and BayScenv (de Villemereuil and Gaggiotti, in press) compare the fit of the allele frequency data as a covariance matrix with environmental variables to a model using only neutral structure, or a null model. Issues arise in ensuring only neutral loci are incorporated into the null model; cross-method comparison studies have identified a high rate of false positives and false discovery rate in BayEnv, particularly in populations exhibiting hierarchical population structuring (de Villemereuil et al. 2014). As BayEnv is computationally intensive and relatively slow, faster methods have been developed based on a Bayesian framework. BayScenv (de Villemereuil and Gaggiotti, in press) includes neutral genetic effects in the model as random factors and an alternative model containing environmental factors. Like logistical approaches, only one environmental variable can be tested simultaneously, severely reducing run-time.

Latent factor mixed models (LFMM) (Frichot et al., 2013), integrate environmental variables into a PCA framework of a Bayesian regression mixed-model as latent factors. This approach requires the estimation of the number of populations (k) (Frichot et al., 2013). Simulated data showed LFMM to perform well compared to Bayenv and linear regression models (higher detection of true positives, reduced error and well-calibrated p-values); however, this was also highly dependent on the selected value of K (Frichot et al., 2013). While Bayesian approaches allow for the incorporation of various, flexible sources of uncertainty (Lemey et al., 2010), establishing appropriate priors and posteriors for inference can be obscured by complex demographic characteristics, resulting in an imperfect null model (Rellstab et al., 2015). Though using additional tests to determine appropriate input for these analyses may be a more sophisticated approach to this problem, the benefit of reducing the number of tests (and their associated assumptions) inferred through these mixed-methods may be consequently nullified.

Multiple linear regression models have been developed to test the effects of multiple environmental variables on loci simultaneously. Canonical correlation analysis (CCA) and redundancy analysis (RDA) implement regression to find an optimal relationship of orthogonal sets of variables to test for significance against explanatory variables (Hecht et al., 2015; Bradbury et al., 2014; Bourret et al., 2014). These methods are advantageous as they allow for neutral genetic structure or geographic distance to be incorporated as an explanatory variable and allow for numerous environmental factors to be tested simultaneously. Hecht et al. (2015) incorporated neutral genetic factors through underlying genetic patterns into RDA and identified precipitation and stream conditions (water flow) as a likely driver in adaptive divergence in populations of Chinook salmon. Bourret et al. (2014) used RDA to regress allele frequency data with principal components of environmental factors from a PCA in Atlantic salmon, identifying 12 SNPs associated with three environmental PCs. Using a similar approach with microsatellite data, Bradbury et al. (2014) identified the importance of watershed size as a driving force in genetic divergence of populations. As habitat size greatly influences carrying capacity and therefore population size, this is likely indicative of strong neutral forces shaping the detected structure. Using a PCA for explanatory variables reduces collinearity but may cloud interpretation of the influence of environmental factors, depending on the loadings of a given PC.

To avoid reducing the size of our selected data set, we applied a partial RDA on all SNPs constrained against PCs of each category of environmental data (precipitation, temperature, habitat), while controlling for geographic distance across sites, allowing for clearer interpretation of correlated vectors. Although the use of sampling site as a proxy for neutral genetic structure may be overly simplified given the population structure previously revealed, RDAs consider collinearity between both independent and dependent variables (environmental parameters and genetic structure, respectively). For the purposes of gaining initial understanding of environmental parameters influencing population structure without further reductions to the size of the data set, we believe this approach to be sufficient.

To assess the importance of each environmental parameter on genetic population structure, we also ran random forest regression analysis (also called regression forests) as an alternative regression approach (Breiman, 2001). Like the classification approach discussion in Chapter 2, RF uses a subset of features, or in this case, environmental variables, as predictors of a model. Instead of assigning to a class, decision trees regress each feature to a continuous response variable (here, allele frequency). In RF classification, a split at a node occurs with discordant 'votes' for a given class, across the subset of features. In RF regression, features are used to predict the value of the response variable at a given feature value. As multiple predictions are made, the prediction with the minimum residual squared error (RSE), the squared difference between the predicted value and the actual value, of the dependent variable is essentially equivalent to a correct classification in a classification tree. Features can likewise be ranked by MDA or increase in error to the model when a feature is removed based on the overall contribution to obtaining the minimum RSE across nodes. The successful application of RF regression to identify SNPs associated with phenotypic traits in plants (Holliday et al., 2012) and Chinook salmon (Brieuc et al., 2015), suggest that the predictive power of explanatory variables (here environmental data) on genetic data may be detected with greater power than simpler approaches such as linear regression. In her Master's thesis, Zhan (2016) applied RF to select SNPs associated with important environmental parameters. Fourteen SNPs were found to associate with 10 of 90 environmental parameters. Although we are not seeking to further reduce data sets, the environmental parameters most related to genetic variance may still be identified through the MDA across all loci. By assessing the average RSE for each site, it is also possible to determine site-specific effects on the regression model.

3.2.2 Method Application

We sourced 19 BioClim (WorldClim) variables (Fick and Hijmans, 2017), derived from interpolated models at 1-km spatial resolution, of monthly rainfall and temperature data, for all 35 sites (Table 3.4). To assess the accuracy of these data we visually compared downloaded data from three locations to nearby weather station data accessed from

Environment Canada's climate normals online. An additional 10 habitat variables (number of obstructions, number of complete obstructions, drainage area, mean width, axial length, basin perimeter, maximum basin relief, length by meander, total length and number of tributaries) were included for 29 sites (Anderson, 1985). Missing data (for Caroline River, Muddy Bay Brook, Main Brook, Port Marnum, Pottle's Bay, and Red Wine River) for the number of tributaries were approximated using Google Earth. Remaining habitat missing data were replaced with the variable median. All environmental data were standardized and normalized in R.

The first PC was retained from a PCA for each category of environmental data. Loadings of each variable on their respective PCs are available in Table 3.4. These PCs were then used as constraining variables, explaining allele frequency, conditioned on geographic distance from the northernmost site (Hunt River) for RDAs for SNP and microsatellite data. Rare microsatellite alleles (present in less than 5% of all individuals) were removed for RDA analysis. RDAs were run using the R package 'vegan' (Oksanen et al., 2016).

We ran multiple RF regression using the R package 'randomForest', using SNP and microsatellite allele frequency for each locus as individual response vectors, with standardized and normalized environmental data as the predictor set of variables. Distance from the northernmost river was also included in this predictor variable set to compare associations with environmental parameters relative to geographic distance. Parameters were implemented as described in Chapter 2, apart from the m_{try} parameter, which was set to default for RF regression. MDA for each environmental variable and RSE for each site was averaged across all loci.

3.2.3 Results

RDA analyses from both SNP and microsatellite data indicated that habitat parameters explained the most genetic variance associated with environmental variables (Fig. 3.5). Precipitation, temperature, and habitat PC vectors were found to be significant ($p<0.05$), with the temperature vector the most significant ($p<0.001$) in an ANOVA analysis for both the microsatellite and SNP data. The total proportion of the variance explained by

Table 3.4: Environmental variables: associated PC1 loadings for each categorical PCA used in RDA, sum of mean decrease in accuracy (MDA) associated with microsatellite and SNP allele frequencies from random forest regression analysis, in percent (%).

Environmental Variable		PC1 Loading	RF MDA (micros)	RF MDA (SNPs)
Temperature PC				
Annual mean Temperature	BIO1	-0.235	0.007	0.016
Mean diurnal range (mean of monthly (max temp - min temp))	BIO2	0.343	0.024	0.067
Isothermality (mean diurnal range/temperature annual range) * 100	BIO3	-0.058	0.001	0.002
Temperature seasonality (standard deviation * 100)	BIO4	0.350	0.023	0.055
Max temperature of warmest month	BIO5	0.329	0.024	0.063
Minimum temperature of coldest month	BIO6	-0.341	0.017	0.039
Temperature annual range (BIO5 - BIO6)	BIO7	0.351	0.023	0.057
Mean temperature of wettest quarter	BIO8	0.274	0.016	0.055
Mean temperature of driest quarter	BIO9	-0.288	0.015	0.056
Mean temperature of warmest quarter	BIO10	0.289	0.018	0.060
Mean temperature of coldest quarter	BIO11	-0.335	0.016	0.037
Precipitation PC				
Annual precipitation	BIO12	-0.466	0.010	0.031
Precipitation of wettest month	BIO13	-0.032	0.009	0.029
Precipitation of driest month	BIO14	-0.453	0.008	0.026
Precipitation seasonality (coefficient of variation)	BIO15	0.373	0.006	0.014
Precipitation of wettest quarter	BIO16	-0.098	0.007	0.012
Precipitation of driest quarter	BIO17	-0.466	0.014	0.033
Precipitation of warmest quarter	BIO18	-0.034	0.007	0.011
Precipitation of coldest quarter	BIO19	-0.458	0.012	0.028
Habitat PC				
Number of obstructions		-0.191	0.003	0.012
Number of complete obstructions		-0.086	0.002	0.006
Drainage area		-0.383	0.003	0.012
Mean width		-0.370	0.004	0.011
Axial length		-0.386	0.003	0.010
Basin perimeter		-0.384	0.003	0.011
Maximum basin relief		-0.189	0.003	0.009
Length by meander		-0.273	0.005	0.016
Total length		-0.358	0.007	0.017
Number of tributaries		-0.371	0.006	0.018
Geographic distance		NA	0.015	0.032

the constraining (environmental PCs) variables was 13.89% and 17.4%, with 6.5% and 8.1% of the total proportion of variance explained by conditioning variables (distance) for microsatellites and SNPs, respectively. A greater proportion of the variance can be explained by environmental data in the SNP data set compared to microsatellites, supporting the idea that selection is more likely to be detected in SNPs. The relatively low variance explained by environmental PCs overall indicates that while there might be some noticeable and significant effect of environmental vectors, particularly for temperature and habitat-related variables, the majority of the genetic variance (79.0% and 73.7% in microsatellites and SNPs, respectively) is not explainable by constrained factors. This is likely due to neutral forces influencing genetic variation, though there may be additional parameters not considered in these analyses that affect the observed genetic structure. Despite a relatively low proportion of variance explained by the RDA axes (Fig. 3.5), vectors associating genetic frequency data with temperature and habitat and to a lesser degree, precipitation, were evident with both genetic data sets. Habitat appeared most influenced by southernmost sites (Forteau River and L'anse au Loup River) and inner north shore Lake Melville populations (Double Mer, Partridge Point River, Sebaskachu River and Mulligan River). Temperature was most influenced by rivers draining into west Lake Melville. Closer inspection of habitat variables revealed that these sites are smaller than other rivers included in the analysis (Table 3.4). Further, habitat parameters indicative of river size (drainage area, mean width, axial length, basin perimeter, total length and number of tributaries) had the greatest load on the habitat PC1 used in the RDA (Table 3.4).

For both molecular marker types, MDA values calculated from RF regression were quite small, suggesting that environmental variables correlate little with allele frequency across populations. Consistent with the proportion of variance explained by RDAs, microsatellite-associated MDAs were consistently lower than SNP-associated MDAs, suggesting that SNPs are more effected by particular environmental parameters. For both types of molecular marker, temperature-associated variables had the greatest impact on the accuracy of the model (largest MDA). The highest MDA values were mostly associated with variables relating to temperature ranges (mean diurnal range and temperature annual range) and warmth (maximum temperature of warmest month and

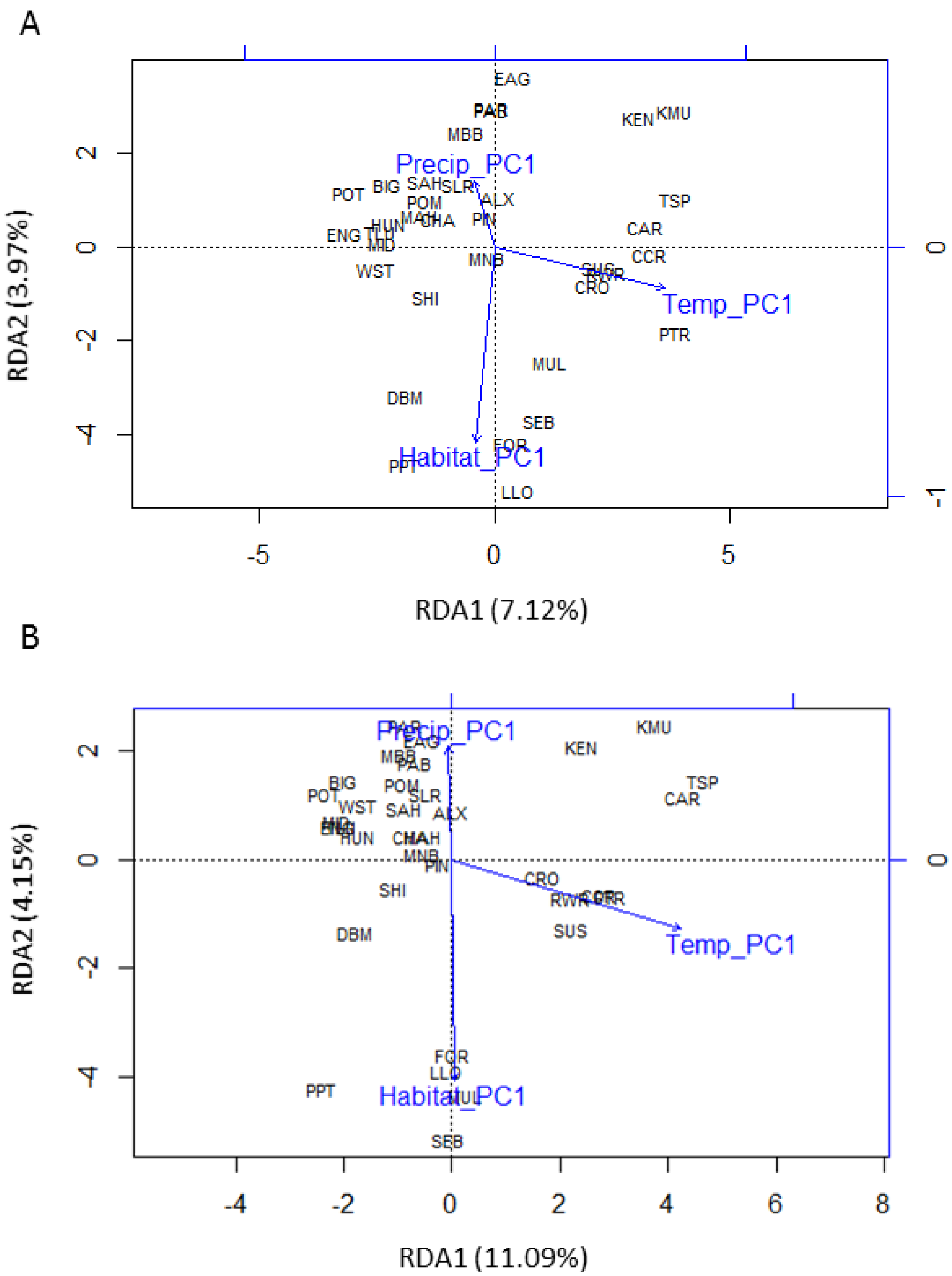

Figure 3.5: RDA plot of (A) microsatellites and (B) SNPs indicating distribution of sites explained by first PCs of three categories of environmental variables (Precipitation, Temperature and Habitat), conditioned on geographic location. Axis labels indicate the proportion of variance explained by each RDA axis out of the total variance.

Table 3.5: Mean and standard error of all habitat variables for sites most influenced by habitat variables according to RDA analyses (Forteau River, L'anse au Loup, Double Mer, Partridge Point River, Sebaskachu River and Mulligan River), relative to all other sites.

Variable	Habitat-influenced sites		Non-habitat influenced sites	
	Mean	SE	Mean	SE
No. of obstructions	2.50	0.81	2.74	0.51
No. of complete obstructions	1.00	0.62	0.78	0.29
Drainage area (km^2)	609.33	200.02	2059.87	510.45
Mean width (km)	13.33	2.39	23.78	2.93
Axial length (km)	34.00	5.96	68.35	7.23
Basin Perimeter (km)	111.33	24.04	228.00	29.08
Max. basin relief (m)	295.00	43.56	436.96	38.42
Length by Meander (km)	41.33	9.49	86.00	10.99
Total Length (km)	158.17	33.52	782.18	223.19
No. of tributaries	24.00	4.40	36.52	5.50

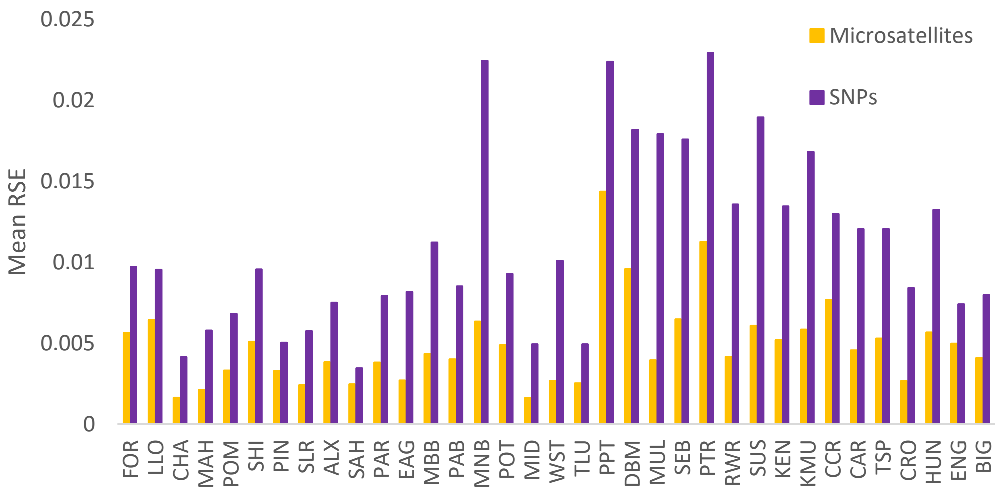

Figure 3.6: Site-specific mean residual squared error (RSE), averaged across all regression forests created for each microsatellite allele and SNP locus. Sites are clustered generally by geographic location (see Fig. 1.2).

mean temperature of warmest quarter). Unlike RDA findings, habitat variables had very little impact on either RF regression analysis.

Average RSE across all sites (Fig. 3.6) was lower for microsatellites indicating that overall, allele frequency was better predicted than that of SNPs. While this may seem contradictory with the lower MDA values for microsatellites, this is not necessarily the case. As MDA gives us relative importance of each independent variable to the overall model, the regression overall may still be accurate, but not affected by any one variable in particular. With SNPs, however, temperature variables were considerably higher, suggesting that the inclusion of these parameters improves the accuracy of the regression model. The highest RSE for both microsatellites and SNPs was observed in sites in Lake Melville, both west and north-shore populations. These allele frequencies at these sites are therefore more difficult for the regression forests to predict, but this accuracy increases with the variables with highest MDA.

3.2.4 Discussion

The results of both EAA approaches are consistent with salmonid landscape associations observed in previous studies. That microsatellites are less associated with environmental variables than SNPs suggests that microsatellite structure is more affected by genetic drift (Bradbury et al., 2014; Ozerov et al., 2012; Dillane et al., 2008), likely due to the higher mutation rate in microsatellites resulting in multi-allelism, indicating a stronger influence of drift within subpopulations (Nishant et al., 2009). Both the SNP RDA and regression forests indicate that Labrador Atlantic salmon may be influenced by local adaptation, or that temperature difference may restrict gene flow. Both RDAs and elevated site-specific RSE suggest that west Lake Melville populations are most affected by this phenomenon. As water temperatures within the embayment are higher than in the rest of Labrador, that variables related to temperature range and extreme (relative to the area) warm temperatures are most associated with allele frequency in the regression forests also supports this hypothesis. Over time, the exposure of juveniles to warmer conditions may result in selection of temperature-associated genes involved in regulatory or developmental processes. Presumably because of the difficulty of dealing with extreme temperature regimes, both cold and warm temperatures have been shown to relate to genetic structure in salmonids (Larson et al., 2016; Kovach et al., 2015; Chang and Psaris, 2013). If an individual or its offspring is less adapted to these extreme temperatures, the likelihood of survival is severely decreased, effectively reducing gene flow. As temperature is often correlated with distance along a latitudinal gradient, it can be difficult to distinguish temperature associations from isolation by distance along a latitudinal gradient. Given that our temperature associations are not along a latitudinal gradient, and that there is little covariance between latitude and temperature-related variables in our data (Fig. 3.7), it is unlikely that temperature associations are an artefact of latitudinal effect. However, this is only true for our comparison involving Lake Melville. Exploring the environmental associations of coastal Labrador sites alone may reveal a stronger latitudinal and temperature gradient associated with the observed admixture in our population analyses that is not detectable here. It may also be that different environmental conditions affect these populations unevenly. Population structure of steelhead trout (*Oncorhynchus mykiss*) from coastal populations has been found to correlate strongest with precipitation and the distance of the spawning site to the ocean, while inland sites are more influenced by precipitation and temperature (Matala et al., 2014). A deeper analysis of coastal Labrador structure and landscape associations may reveal incongruent trends across regions. The MDA associated with geographic distance is high, relative to environmental variables, for both the SNP and microsatellite data, suggesting that geographic distance can explain a great deal of genetic variance across populations. Neutral or random factors clearly influence genetic structure, but this does not eliminate the possibility of selection also influencing local adaptation.

In this chapter, we show evidence of temperature-driven structure. As temperature has been found to influence spawning time in Pacific salmon (Lisi et al., 2013), there may be the development of temporal isolation over time, reinforcing the current structure. Alternatively, in climate change conditions resulting in warming ocean temperatures (IPCC, 2013), warm temperature-intolerant populations may become less productive, leading to replacement by warm-tolerant populations to these rivers, leading to a loss of overall genetic diversity. This, of course, would depend on the specific conditions to which individuals are exposed, and the plasticity of warm-tolerant populations. Predicting changes in population structure requires interdisciplinary approaches to understand responses of species to projected climate conditions.

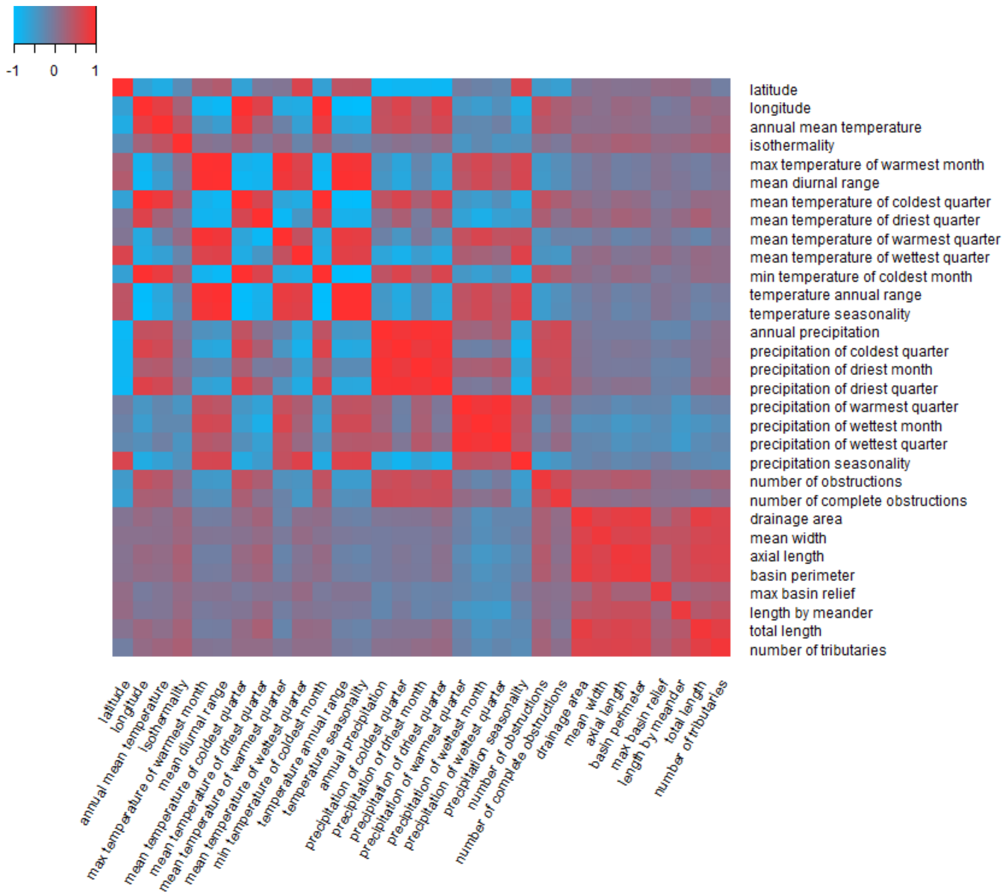

Figure 3.7: Heatmap of covariance of all environmental variables used for EAA.
Point, Mulligan River and Sebaskachu River, sometimes (in the case of the SNP DAPC) with a further division of these sites into two groups. Here, smaller river size may affect population size, influencing genetic drift in these areas.

Though there was no clear indication of habitat influencing population structure in the RF regression analysis, RDAs indicated that habitat size may influence population structure of southernmost populations, and north shore Lake Melville populations. North-shore Lake Melville populations also had relatively high RSEs, though MDAs do not indicate a strong habitat influence on improving the regression models. MDAs alone indicate this structure may also be temperature driven, but this is not supported by the RDA vectors. Despite the lack of support through MDAs, habitat parameters may help to explain the population structure observed within Lake Melville. Our population structure analyses with both molecular marker types identified one or two separate genetic clusters consisting of Double Mer, Partridge Point, Middle Bay Brook, Sebaskachu River and Mulligan River.

Our population structure analyses consistently identify west Lake Melville populations as a discrete genetic cluster. To effectively assess the importance of Lake Melville for genetic diversity, we sought evidence of local adaptation within the region. We found evidence of temperature-regulated selection acting on both microsatellite and SNP variation through linear constrained ordination and non-linear regression forests. Genetic associations with geographic distance and the large amount of unexplained variance in the data set indicate neutral factors influence both microsatellites and SNPs. Despite these promising findings, we were limited by the environmental data available for us to explore. Environmental parameters taken at the time of sampling and covering a wider description of the habitat and water conditions (e.g. salinity, water temperature, chemical constituents, chlorophyll concentration) would likely be better suited as they more accurately describe the environmental conditions faced by salmon in these habitats. Lastly, as we did not identify outlier SNPs, neutral SNPs in our dataset may reduce the likelihood of detecting true selective forces, and may result in false positive associations. The work presented here provides a starting point in uncovering the factors affecting genetic diversity of Labrador Atlantic salmon.

Chapter 4

Conclusion

In this work, we describe a novel application of random forest for SNP selection for use in individual assignment. By comparing RF and RF variations with an established method (F_{ST} rank) in two SNP data sets, we demonstrate consistent improvement in assignment accuracy for two species of high commercial interest. We apply these techniques for referencing individuals to spawning site across a small-scale, within a single marine embayment for Atlantic salmon and across a broader scale of five populations of Chinook salmon in Alaska. The RF methods outperform F_{ST} rank for SNP selection in both applications. For the published data set our application of RF methods shows improved assignment relative to the original publication (Larson et al., 2014a), comparable to subsequent analyses utilizing the information gain of haplotype genotypes for assignment (McKinney et al., 2017). We show consistent patterns of population structure across methods, further supporting the accuracy of RF for classifying individuals to populations.

In Chapter 3, we further demonstrate the utility of RF-selected SNPs by testing the resolution of population structure achieved using a panel of 376 SNPs selected for importance in classification within Lake Melville. Including a total of 35 rivers, we conduct a Labrador-wide study of population structure. Our approaches enabled the exploration of the genetic distinctiveness of Lake Melville populations using a variety of methods to support overall structure. Comparing our SNP data set with a panel of 101 microsatellites, we provide strong support for Lake Melville populations as a unique genetic cluster. We also show evidence of within-region delineation between west and north shore Lake Melville populations, and a latitudinal cline in admixture along costal Labrador, consistent across statistical methods and molecular marker sets.

Both RDA and regression forests provide putative evidence of temperature and, to a lesser degree, habitat effects on population genetic variation. Most promisingly, west Lake Melville populations show the strongest association with temperature, consistent with warmer temperature regimes found in the area. It is highly likely that both neutral and adaptive forces influence genetic structure across the genome. Although the low MDA overall indicates that genetic differentiation is mostly unrelated to the environmental conditions that were tested, there were consistent associations identified through both methods, and a stronger environmental association with SNPs compared to microsatellites.

4.1 Future Work

Our methods have the potential to be effective for numerous other applications. By testing not only random forest, but additional machine-learning algorithms for genetic marker selection, population classification, landscape associations, and extending its use to other genetic and ecological questions, there is great potential for developing novel techniques, and improving upon existing approaches. Fast, efficient techniques with easy implementation are not only important within academia, but also for increasing the knowledge base on which adequate management policies rely. Here, we show a successful application of random forest for feature-selection on hierarchical populations that are likely strongly delineated. We focused on the utility of RF for classification for feature-selection and regression due to its optimal use with many features relative to the number of samples (Breiman, 2001). However other machine-learning algorithms may also be highly suitable for feature-selection of genetic markers, individual assignment, or inferring ecological associations with genetic structure (Guinand et al., 2002). Genetic algorithms are designed to function based on principles of natural selection and have been successfully applied for use in ranking SNPs for importance in drug resistance development (Shah and Kusiak, 2004). Support vector machines can also be implemented for classification or regression, and are able to detect complex relationships through the use of customisable kernels. SVMs have been widely used to classify genes based on cancer classification, gene expression data (e.g. Vanitha et al., 2015; Guyon et al., 2002) and in ecological genetics, to assign individuals to a genetic group based on a phenotype (Grbic et al., 2015). K-nearest neighbour clustering may also be highly useful, particularly for genetic individual assignment, as it has been applied to assess the genetic origin of trees (Degen et al., 2017), and humans (Huckins et al., 2014) using SNP panels.

Interdisciplinary approaches to develop deterministic algorithms that are highly suited for genetic assignment and classification problems may prove to be highly applicable across study systems.

Within the context of Atlantic salmon in Labrador, we have provided consistent, initial evidence of population structure at a finer resolution than addressed by the management strategies currently in place. As previously discussed, additional criteria are needed to establish the need for a population to be treated as a separately considered DU. With both microsatellites and SNPs, we found evidence of a genetic split between Lake Melville and greater Labrador equal to or greater than that between populations north and south of Lake Melville, suggesting that updating DUs to reflect this genetic diversity may be appropriate. Evidence of temperature and habitat-related structure suggests the need for a more in-depth exploration of potentially adaptive loci, as well as their associated genes with accurate, local, aquatic environmental parameters included in the analyses.

We further hope to investigate the information gained using haplotype genotypes by identifying SNPs within the flanking regions of microsatellites, and in the microsatellites themselves, as well as identifying multiple SNPs per locus in our amplicon-based detection approach. Without incurring additional laboratory costs, this may allow for finer resolution of population structure, and improvement in genetic population assignment.

Though no single approach is likely to provide an over-arching solution, even for a single species, by working towards the development of evidence-based strategies aimed at conserving genetic diversity, we hope to contribute to the recovery of threatened wildlife and promote lasting conservation.

References

Anderson, E. C. (2010). Assessing the power of informative subsets of loci for populations assignment: standard methods are upwardly biased. Molecular Ecology Resources, 10, 701-710.

Anderson, E. C., Waples, R. S. and Kalinowski, S. T. (2008). An improved method for predicting the accuracy of genetic stock identification. Canadian Journal of Fisheries and Aquatic Sciences, 65, 1475-1486.

Anderson, T. C. (1985). The rivers of Labrador. Department of Fisheries and Oceans. Canadian Special Publication of Fisheries and Aquatic Sciences. 81: 389p.

André, C., Svedäng, H., Knutsen, H., Dahle, G., Jonsson, P., Ring, A.-K., Sköld, M. and Jorde, P. E. (2016). Population structure in Atlantic cod in the eastern North Sea-Skagerrak-Kattegat: early life stage dispersal and adult migration. BMC Research Notes, 9, 1.

Bekkevold, D., Helyar, S. J., Limborg, M. T., Nielsen, E. E., Hemmer-Hansen, J., Clausen, L. A. and Carvalho, G. R. (2015). Gene-associated markers can assign origin in a weakly structured fish, Atlantic herring. ICES Journal of Marine Science, 72, 1790-1801.

Boulesteix, A., Janitza, S., Kruppa, J. and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computation biology and bioinformatics. WIREs Data and Mining Knowledge Discover, 2, 493-507.

Bourret, V., Dionne, M., Kent, M. P., Lien, S. and Bernatchez, L. (2013). Landscape genomics in Atlantic salmon (*Salmo salar*): searching for gene–environment interactions driving local adaptation. Evolution, 67, 3469-3487.

Bradbury, I. R., Hamilton, L. C., Chaput, G., Robertson, M. J., Goraguer, H., Walsh, A., Morris, V., Reddin, D., Dempson, J. B. and Sheehan, T. F. (2016). Genetic mixed stock analysis of an interceptory Atlantic salmon fishery in the Northwest Atlantic. Fisheries Research, 174, 234-244.

Bradbury, I. R., Hamilton, L. C., Rafferty, S., Meerburg, D., Poole, R., Dempson, J. B., Robertson M. J., Reddin D. G., Bourret, V., Dionne, M., Chaput, G., Sheehan T. F. and King, T. L. (2015a). Genetic estimates of local exploitation of Atlantic salmon in a coastal subsistence fishery in the Northwest Atlantic. Canadian Journal of Fisheries and Aquatic Sciences, 72, 83-95.

Bradbury, I. R., Hamilton, L. C., Dempson, B., Robertson, M. J., Bourret, V., Bernatchez, L. and Verspoor, E. (2015b). Transatlantic secondary contact in Atlantic Salmon, comparing microsatellites, a single nucleotide polymorphism array and restriction-site associated DNA sequencing for the resolution of complex spatial structure. Molecular Ecology, 24, 5130-5144.

Bradbury, I. R., Hamilton, L. C., Robertson, M. J., Bourgeois, C. E., Mansour, A. and Dempson, J. B. (2014). Landscape structure and climatic variation determine Atlantic salmon connectivity in the Northwest Atlantic. Canadian Journal of Fisheries and Aquatic Sciences, 71, 246-258.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Brieuc, M. S., Ono, K., Drinan, D. P. and Naish, K. A. (2015). Integration of Random Forest with population-based outlier analyses provides insight on the genomic basis and evolution of run timing in Chinook salmon (*Oncorhynchus tshawytscha*). Molecular Ecology, 24, 2729-2746.

Bromaghin, J. F. (2008). BELS: backward elimination locus selection for studies of mixture composition or individual assignment. Molecular Ecology Resources, 8(3), 568-571.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P. and Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. Genetic Epidemiology, 28, 171-182.

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 analysis using linkage disequilibrium. American Journal of Human Genetics, 74(1), 106-120.

Cavalli-Sforza, L. L., Edwards, A. W. F. (1967). Phylogenetic analysis: models and estimation procedures. Evolution, 21(3), 550-570.

Chang, H. and Psaris, M. (2013). Local landscape predictors of maximum stream temperature and thermal sensitivity in the Columbia River Basin, USA. Science of the Total Environment, 461, 587-600.

Corander, J., Sirén, J., & Arjas, E. (2008). Bayesian spatial modeling of genetic population structure. Computational Statistics, 23(1), 111-129.

COSEWIC. (2011). COSEWIC assessment and status report on the Atlantic salmon *Salmo salar* in Canada. Ottawa Committee on the Status of Endangered Wildlife in Canada.

DeFaveri, J., Viitaniemi, H., Leder, E. and Merilä, J. (2013). Characterizing genic and nongenic molecular markers: comparison of microsatellites and SNPs. Molecular Ecology Resources, 13(3), 377-392.

Degan, B., Blanc-Jolivet, C., Stierand, K. and Gillet, E. (2017). A nearest neighbor approach by genetic distance to the assignment of individual trees to geographic origin. Forensic Science International: Genetics, 27, 132-141.

Dempson, B., Schwartz, C. J., Bradbury, I. R., Robertson, M. J., Veinott, G., Poole, R. and Colbourne, E. (2017), Influence of climate and abundance on migration timing of adult Atlantic salmon (*Salmo salar*) among rivers in Newfoundland and Labrador. Ecology of Freshwater Fish, 26, 247-259.

Deng, H. and Runger, G. (2013). Gene selection with guided regularized random forest. Pattern Recognition, 46, 3483-3489.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7, 1.

Dillane, E., McGinnity, P., Coughlan, J. P., Cross, M. C., De Eyto, E., Kenchington, E., Prodöhl, P., and Cross, T. F. (2008). Demographics and landscape features determine intrariver population structure in Atlantic salmon (*Salmo salar* L.): the case of the River Moy in Ireland. Molecular Ecology, 17(22), 4786-4800.

Dunfield, R. W. (1985). The Atlantic salmon in the history of North America. Canadian Special Publication of Fisheries and Aquatic Sciences. 81: 181p.

DFO. (2016). Atlantic salmon (*Salmo salar*) stock status update in Newfoundland and Labrador for 2015. DFO Canadian Sciences Advisory Secretariat Science Response.

ESRI. (2011). ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.

Ferchaud, A. and Hansen, M. M. (2016). The impact of selection, gene flow and demographic history on heterogenous genomic divergence: threespine sticklebacks in divergent environments. Molecular Ecology, 25, 238-259.

Fick, S. E. and Hijmans, R. J. (2017). Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology.

Foll, M. and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and co-dominant markers: a Bayesian perspective. Genetics, 180, 977-993.

Foll, M. and Gaggiotti, O. (2006). Identifying the environmental factors that etermine the genetic structure of populations. Genetics, 174, 875-891.

Frichot, E., Schoville, S. D., Bouchard, G. and Francois, O. (2013). Testing for associations between loci and environmental gradients using Latent Factor Mixed Models. Molecular Biology and Evolution, 30(7), 1687-1699.

Glover, K. A., Hansen, M. M., Lien, S., Als, T. D., Høyhein, B., and Skaala, Ø. (2010). A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. BMC Genetics, 11(2), doi: 10.1186/1471-2156-11-2.

Gosselin, T., Benestan, L. and Bernatchez, L. (2015). assigner: Assignment Analysis with GBS/RAD Data using R. R package version 0.1.4.

Goudet, J. and Jombart, T. (2015). Hierfstat: estimate and tests of hierarchical F-statistics. R package version 0.04-22.

Grbic, D., Saenko, S. V., Randriamoria, T. M., Debry, A., Raselimanana, A. P., Milinkovitch, M. C. (2015). Phylogeography and support vector machine classification of colour variation in panther chameleons. Molecular Ecology, 24(13), 3455-3466.

Greig, C., Jacobson, D. P. and Banks, M. A. (2003). New tetranucleotide microsatellites for fine-scale discrimination among endangered chinook salmon (*Oncorhynchus tshawytscha*). Molecular Ecology Resources, 3(3), 376-379.

Guinand, B., Topchy, A., Page, K. S., Burnham-Curtis, M. K., Punch, W. F., Scribner, K. T. (2002). Comparisons of likelihood and machine learning methods of individual classification. The American Genetic Association, 93, 260-269.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46, 389-422.

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

Hecht, B.C., Matala, A.P. Hess, J.E. and Narum, S.R. (2015). Environmental adaptation in Chinook salmon (*Oncorhynchus tshawytscha*) throughout their North American range. Molecular Ecology. doi: 10.1111/mec.13409

Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M., Ogden, R., Limborg, M., Cariani, A., Maes, G., Diopere, E. and Carvalho, G. (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Molecular Ecology Resources, 11, 123-136.

Hendry, A. P., Castric, V., Kinnison, M. T., Quinn, T. P., Hendry, A. and Stearns, S. (2004). The evolution of philopatry and dispersal. Evolution illuminated salmon and their relatives, 52-91.

Hess, J. E., Matala, A. P. and Narum, S. R. (2011). Comparison of SNPs and microsatellites for fine-scale application of genetic stock identification of Chinook salmon in the Columbia River Basin. Molecular Ecology Resources, 11, 137-149.

Hilborn, R., Quinn, T. P., Schindler, D. E. and Rogers, D. E. (2003). Biocomplexity and fisheries sustainability, PNAS, doi_10.1073_pnas.1037274100

Holliday, J.A., Wang, T. and Aitken, S. (2012). Predicting adaptive phenotypes from multilocus genotypes in Sitka spruce (*Picea sitchensis*) using random forest. G3 (Bethesda). 2(9), 1085-1093. doi: 10.1534/g3.112.002733

Huckins, L. M., Boraska, V., Franklin, C. S., Floyd, J. A. B., Southam, L., GCAN, WTCCC3, Sullivan, P. F., Bulik, C. M., Collier, D. A., Tyler-Smith, C., Zeggini, E. and Tachmazidou, I. (2014). Using ancestry-informative markers to identify fine structure across 15 populations of European origin. European Journal of Human Genetics, 22, 1190-1200.

ICES. (2013). Report of the Working Group on North Atlantic Salmon (WGNAS). 3-12 April 2012. Copenhagen, Denmark. ICES CM

IPCC (2013) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (ed. by T. F. Stocker, D. Qin, G. K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P. M. Midgley), 148 pp. Cambridge University Press, Cambridge, U.K. and New York, New York.

Janes, J. K., Miller, J. M., Dupuis, J. R., Malenfant, R., M., Gorrel, J., C., Cullingham, C. I., Andrew, R. L. (2017). The K=2 conundrum. Molecular Ecology, 26(14), 3594-3602.

Jombart T., Devillard S., Balloux F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. BMC genetics, 11, 94.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics**,** 24, 1403-1405.

Joost, S., Bonin, A., Bruford, M. W., Despres, L., Conord, C., Erhardt, G. and Taberlet, P. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. Molecular Ecology. 16, 3955-3969. doi: 10.1111/j.1365-294X.2007.03442.x

Karlsson, S., Moen, T., Lien, S., Glover, K. A. and Hindar, K. (2011). Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP-chip. Molecular Ecology Resources, 11, 247-253.

King, T. L., Kalinowski, S. T., Schill, W. B., Spidle, A. P. and Lubinski, B. A. (2001). Population structure of Atlantic salmon (*Salmo salar* L.): a range-wise perspective from microsatellite DNA variation. Molecular Ecology, 10, 807-821.

King, T. L., Spindle, A. P., Eackles, M. S., Lubinski, B. A. and Schill, W. B. (2000). Mitochondrial DNA diversity in North American and European Atlantic salmon with emphasis on the Downeast rivers of Maine. Journal of Fish Biology, 57, 614-630.

Kursa, M. B. (2014). Robustness of Random Forest-based gene selection methods. BMC Bioinformatics, 15, 1.

Langella, O. (1999). Populations 1.2.31. CNRS UPR9034.

Larson, W. A., Lisi, P. J., Seeb, J. E., Seeb, L., W. and Schindler, D. E. (2016). Major histocompatibility complex diversity is positively associated with stream water temperatures in proximate populations of sockeye salmon. Journal of Evolutionary Biology, 29(9), 1846-1859.

Larson, W. A., Seeb, L. W., Everett, M. V., Waples, R. K., Templin, W. D. and Seeb, J. E. (2014a). Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). Evolutionary Applications, 7, 355-369.

Larson, W. A., Seeb, L. W., Everett, M. V., Waples, R. K., Templin, W. D. and Seeb, J. E. (2014b). Data From: Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.rs4v1.

Larson, W. A., Seeb, J. E., Pascal, C. E., Templin, W. D. and Seeb, L. W. (2014c). Single-nucleotide polymorphisms (SNPs) identified through genotyping-by-sequencing improve genetic stock identification of Chinook salmon (*Oncorhynchus tshawytscha*) from western Alaska. Canadian Journal of Fisheries and Aquatic Sciences, 71, 698-708.

Lemay, M. A. and Russello, M. A. (2015). Genetic evidence for ecological divergence in kokanee salmon. Molecular Ecology, 24, 798-811.

Lemey, P., Rambaut, A., Welch, J.J. and Suchard, M.A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. Molecular Biology and Evolution, 27(8), 1877-1885.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. R News, 2, 18-22.

Lisi, P. J., Schindler, D. E., Bentley, K. T., Pess, G. R. (2013). Association between geomorphic attributes of watersheds, water temperature, and salmon spawn timing in Alaskan streams. Geomorphology, 185(1), 78-86.

Lotterhos, K. E. and Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Molecular Ecology, 24, 1031-1046. doi: 10.1111/mec.13100

Lou, W., Wang, X., Chen, F., Chen, Y., Jiang, B. and Zhang, H. (2014). Sequence based prediction of DNA-binding proteins based on hybrid feature-selection using random forest and Gaussian naive Bayes. PLoS ONE, 9, e86703.

Manel, S. and Holdregger, R. (2013). Ten years of landscape genetics. Trends in Ecology and Evolution, 28(10), 614-621.

Manel, S., Gaggiotti, O. E. and Waples, R. S. (2005). Assignment methods: matching biological questions with appropriate techniques. Trends in Ecology and Evolution, 20, 136-142.

Manel, S., Schwartz, M. K., Luikart, G. and Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. Trends in Ecology and Evolution, 18(4), 189-197.

Martinsohn, J. T., Ogden, R. and Consortium, F. (2009). FishPopTrace—Developing SNP-based population genetic assignment methods to investigate illegal fishing. Forensic Science International: Genetics Supplement Series, 2, 294-296.

Matala, A. P., Ackerman, M. W., Campbell, M. R. and Narum, S. R. (2014). Relative contributions of neutral and non-neutral genetic differentiation to inform conservation of steelhead trout across highly variable landscapes. Evolutionary Applications, 7(6), 682-701.

McKinney, G. J., Seeb, J. E. and Seeb, L. W. (2017). Managing mixed-stock fisheries: genotyping multi-SNP haplotypes increases power for genetic stock identification. Canadian Journal of Fisheries and Aquatic Sciences, 74, 429-434.

Meirmans, P. G. and P. H. Van Tienderen. (2004). GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. Molecular Ecology Notes, 4, 792-794.

Meng, Y. A., Yu, Y., Cupples, L. A., Farrer, L. A. and Lunetta, K. L. (2009). Performance of random forest when SNPs are in linkage disequilibrium. BMC Bioinformatics, 10, 1.

Mills, K. E., Pershing, A. J., Sheehan, T. F. and Mountain, D. (2013). Climate and ecosystem linkages explain widespread declines in North American Atlantic salmon populations. Global Change Biology, 19(10), 3046-3061.

Moore, J. S., Bourret, V., Dionne, M., Bradbury, I., O'Reilly, P., Kent, M., Chaput, G. and Bernatchez, L. (2014). Conservation genomics of anadromous Atlantic salmon across its North American range: outlier loci identify the same patterns of population structure as neutral loci. Molecular Ecology, 23, 5680-5697.

Neville, H., Isaak, D., Dunham, J., Thurow, R. and Rieman, B. (2006). Fine-scale natal homing and localized movement as shaped by sex and spawning habitat in Chinook salmon: insights from spatial autocorrelation analysis of individual genotypes. Molecular Ecology, 15, 4589-4602.

Nilsson, J., Gross, R., Asplund, T., Dove, O., Jansson, H., Kelloniemi, J., Kohlmann, K., Löytynoja, A., Nielsen, E. E., Paaver, T., Primmer, C. R., Titov, S., Vasemägi, A., Veselov, A., Öst, T. and Lumme, J. (2001). Matrilinear phylogeography of Atlantic salmon (*Salmo salar* L.) in Europe and postglacial colonization of the Baltic Sea area. Molecular Ecology, 10, 89-102.

Ning, J. and Beiko, R. G. (2015). Phylogenetic approaches to microbial community classification. Microbiome, 3, 47.

Nishant, K. T., Singh, N. D. and Alani, E. (2009). Genomic mutation rates: What high-throughput methods can tell us. Bioessays, 31(9), 912-920.

O'Connell, M. F., Dempson, J. B., Reddin, D. G. (1992). Evaluation of the impacts of major management changes in the Atlantic salmon (*Salmo salar* L.) fisheries of Newfoundland and Labrador, Canada, 1984-1988. ICES Journal of Marine Science, 49: 69-87.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legender, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E. and Wagner, H. (2016). Vegan: community ecology package. R package version 2.4-1.

Oliveros, J. C. (2007 – 2015). Venny: an interactive tool for comparing lists with Venn's diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html

Ozerov, M., Vasemägi, A., Wennevik, V., Diaz-Fernandez, R., Kent, M., Gilbey, J., Prusov, S., Niemelä, E. and Vähä, J. P. (2013). Finding markers that make a difference: DNA pooling and SNP-arrays identify population informative parkers for genetic stock identification. PLoS ONE, 8, e82434. doi:10.1371/journal.pone.0082434.

Ozerov, M. Y., Veselov, A. E., Lumme, J., Primmer, C. R. (2012). "Riverscape" genetics: river characteristics influence the genetic structure and diversity of anadromous and freshwater Atlantic salmon (*Salmo salar*) populations in northwest Russia. Canadian Journal of Fisheries and Aquatic Sciences, 69, 1947-1958.

Palstra, F. P., O'Connell, M. F. and Ruzzante, D. E. (2007). Population structure and gene flow reversals in Atlantic salmon (*Salmo salar*) over contemporary and long-term temporal scales: effects of population size and life history. Molecular Ecology, 16, 4504-4522.

Parks, D. H., Mankowski T., Zangooei S., Porter M. S., Armanini D. G., Baird D. J., Langille M. G. I., Beiko R. G. (2013). GenGIS 2: Geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework. PLoS One, 8(7), e69885

Pavey, S. A., Gaudin, J., Normandeau, E., Dionne, M., Castonguay, M., Audet, C. and Bernatchez, L. (2015). RAD sequencing highlights polygenic discrimination of habitat ecotypes in the panmictic American eel. Current Biology, 25, 1666-1671.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, *155*(2), 945-959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. and Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81, 559-575.

R Development Core Team (2012). R: A language and environment for statistical computing. Retrieved from http://www.R-project.org/.

Rambaut, A. (2006). FigTree. website: http://tree.bio.ed.ac.uk/software/figtree/

Reiss, H., Hoarau, G., Dickey-Collas, M. and Wolff, W. J. (2009). Genetic population structure of marine fish: mismatch between biological and fisheries management units. Fish and Fisheries, 10, 361-395.

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M. and Holderegger, R. (2015). A practical guide to environmental association analsis in landscape genomics. Molecular Ecology. 24, 4348-4370.

Rosenberg, N. A. (2005). Algorithms for selecting informative marker panels for population assignment. Journal of Computational Biology, 12(9), 1183-1201.

Rougemont, Q. and Bernatchez, L. (2017). Reconstructing the demographic history of Atlantic Salmon (Salmo salar) across its distribution range using Approximate Bayesian Computations. bioRxiv preprint, doi: dx.doi.org/10.1101/142372.

Shah, S. C. and Kusiak, A. (2004). Data mining and genetic algorithm based gene/SNP selection. Artificial Intelligence in Medicine, 31, 183-196.

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

Stanley, R. R., Jeffery, N. W., Wringe, B. F., DiBacco, C. and Bradbury, I. R. (2016). genepopedit: a simple and flexible tool for manipulating multilocus molecular data in R. Molecular Ecology Resources, 1755-0998.

Taylor, V. R. (1985). The early Atlantic salmon fishery in Newfoundland and Labrador. Canadian Special Publication of Fisheries and Aquatic Sciences. 76: 7p.

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.

Toloşi, L. and Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics, 27, 1986-1994.

Topchy, A., Scribner, K. and Punch, W. (2004). Accuracy-driven loci selection and assignment of individuals. Molecular Ecology Resources, 4(4), 798-800.

Vähä, J. P., Erkinaro, J., Fålkegard, M., Orell, P. and Niemelä, E.E. (2016). Genetic stock identification of Atlantic salmon and its evaluation in a large population complex. Canadian Journal of Fisheries and Aquatic Sciences, doi:10.1139/cjfas-2015-0606.

Vanitha, C., Devaraj, D. and Venkatesulu, M. (2015). Gene expression data classification using support vector machine and mutual information-based gene selection. Procedia Computer Science, 47, 13-21.

Verspoor, E., Beardmore, J. A., Consuegra, S., García de Leaniz, C., Hindar, K., Jordan, W. C., Koljonen, M. -L. Mahkrov, A. A., Paaver, T., Sánchez, J. A. Skaala, O., Titov, S. and Cross, T. F. (2005). Population structure in Atlantic salmon: insights from 40 years of research into genetic protein variation. Journal of Fish Biology, 67, 3-54.

de Villemereuil, P., Frichot, E., Bazin, E., Francois, O. and Gaggiotti, O. E. (2014). Genome scan methods against more complex models: when and how much should we trust them? Molecular Ecology. 23, 2006-2019.

de Villemereuil, P. and Gaggiotti, O. E. (2015). A new FST-based method to uncover local adaptation using environmental variables. Methods in Ecology and Evolution, 6(11), 1248-1258.

Wang, J., Xue, D., Zhang, B., Li, Y., Liu, B. and Liu, J. (2016). Genome-wide SNP discovery, genotyping and their preliminary applications for population genetic inference in spotted sea bass (Lateolabrax maculatus). PloS One, 11(6), e0157809. doi:10.1371/journal.pone.0157809.

Waples, R. S. and Anderson, E. C. (2017). Purging putative siblings from population genetic data sets: a cautionary view. Molecular Ecology, 26, 1211-1224.

Waples, R. S. (1987). A multispecies approach to the analysis of gene flow in marine shore fishes. Evolution, 385-400.

Weir, B. S. and Cockerham C. C. (1984) Estimating F-Statistics for the Analysis of Population Structure. Evolution, 38,1358-1370

Zhan, L. (2016). Inferring ecological population structure and environmental associations through automated analysis of repeat-containing and polymorphic DNA sequences. (Master's Thesis). Available from DalSpace Institutional Repository (dalspace.library.dal.ca).

Zhan, L., Patterson, I. G., Fraser, B. A., Watson, B., Bradbury, I. R., Ravindran, P. N., Reznick, D., Beiko, R. G. and Bentzen, P. (2016). MEGASAT: automated inference of microsatellite genotypes from sequence data. Molecular Ecology Resources, 17(2), 247-256.

Zueva, K. J., Lumme, J., Veselov, A. E., Kent, M. P., Lien, S., Primmer, C. R. (2014). Footprints of directional selection in wild Atlantic salmon populations: evidence for parasite-driven evolution? PloS One, 9(3), e91672. doi: 10.1371/journal.pone.0091672