# INFERRING ECOLOGICAL POPULATION STRUCTURE AND ENVIRONMENTAL ASSOCIATIONS THROUGH AUTOMATED ANALYSIS OF REPEAT-CONTAINING AND POLYMORPHIC DNA SEQUENCES

by

Luyao Zhan

Submitted in partial fulfilment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
May 2016

# DEDICATION PAGE

I dedicate this thesis to:

My mother, Xiaoli Rao, a loving and kind woman who always gives me endless love, support and encouragement throughout my life.

My father, Jianping Zhan, a strong, generous and caring man who always loves me, supports me and protects me. Thank you for calling me everyday and letting me feel your love always beside me. I am so honored to be your daughter.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABSTRACT

Biodiversity conservation plays an important role in the maintenance of a healthy ecosystem. Genetic diversity provides a foundation for understanding the diversity at the organism and population levels of organization. Genomic DNA markers offer the opportunity to identify genetic variations that distinguish populations, and can be used to investigate the underlying forces that drive adaptation to different environments. Short simple-repeat DNA sequences or *microsatellites* are one of the most popular genetic markers for many biological applications. However, microsatellite data require extensive manual checking for errors and characteristic signals, a laborious process that can take days or weeks for a single dataset. We have developed MEGASAT, a bioinformatics approach that automates microsatellite genotyping from DNA sequence data. MEGASAT uses fuzzy matches and counting of frequently observed sequences to distinguish true genotype signal from errors. We validated MEGASAT using microsatellite data from a population sample of 71 guppies from Trinidad, demonstrating a high level of reproducibility and accuracy of MEGASAT-called genotypes by a combination of genotyping error estimation methods.

We also developed a random-forest (RF) based method to identify adaptive gene variants and environmental factors associated with those adaptive variants in sea scallop data. Our approach uses the inverse Cholesky transformation to account for spatial autocorrelations in genetic and environmental data and ordination techniques to further explore the relationships between these two data sets. The variable importance ranked by RF models and ordination techniques were both used on corrected and uncorrected data to find which environmental variables play important role in shaping the genetic structure of sea scallop populations.

# LIST OF ABBREVIATIONS USED

| | |
|---|---|
| NGS | Next Generation Sequencing |
| SNPs | Single Nucleotide Polymorphisms |
| AFLPs | Amplified Fragment Length Polymorphisms |
| RFLPs | Restriction Fragment Length Polymorphisms |
| PCR | Polymerase Chain Reaction |
| SSR | Simple Sequence Repeats |
| VNTR | Variable Number Tandem Repeats |
| PCA | Principal Component Analysis |
| MEM | Moran's Eigenvector Map |
| $F_{ST}$ | Fixation Index |
| CCA | Canonical Correspondence Analysis |
| RF | Random Forest |
| ROC | Receiver Operating Characteristic |
| AUC | Area Under ROC Curve |
| RDA | Redundancy Analysis |
| MLR | Multiple Linear Regression |
| MSE | Mean Squared Error |

# ACKNOWLEDGEMENTS

First, I would like to express my deepest gratitude to my supervisor **Dr. Robert Beiko** for being a brilliant, patient and supportive supervisor. Thanks for your brilliant ideas, creative thinking, effective communication and tireless working, as well as your humorous, considerate and optimistic personality. Your patient guidance and invaluable comments greatly contribute to this thesis. I have learned a lot from your strategic vision and enlightening insights during the last two years.

My sincere gratitude also goes to my co-supervisor **Dr. Paul Bentzen** for his constructive suggestions and constant support. He has spent a lot of time reviewing and editing the manuscript, which has been a great help to me. Thank you for tolerating my poor writing skills.

I would like to thank my examing committee members **Dr. Christian Blouin** and **Dr. Daniel Ruzzante** for serving on my thesis and their guidance and invaluable feedbacks on my thesis.

I would also like to extend my gratitude to Ian G. Paterson, who is the lab manager of the Marine Gene Probe Laboratory in Dalhousie, for his advice and comments on the development of MEGASAT and his efforts on the generation of the guppy data set. A big thanks to Mallory Van Wyngaarden for prodiving the genetic and environmental data of sea scallop. Thank you for your detailed explanations on these data sets.

Thanks to all the lab members in Beiko lab: Alex, Mike, Dennis, Praveen, Nehil, Jonathan, Jyoti, Jessie, Kacper, Eyre and Akhilesh for all the fun we had in the last two years. I also would like to thank Dr. Segio Hleap Lozano for giving suggestions on the statistical methods.

# CHAPTER 1 INTRODUCTION

The variety of living organisms on Earth and diverse ecosystems in which they live and interact are referred to as biological diversity [1, 2, 3]. Biological diversity brings benefits to humans including food, recreation and health benefits such as medicinal resources [4, 5]. It is also essential to the maintenance of healthy ecosystems since it helps species to adjust to the changing environment and confers greater resilience after natural disasters. However, biodiversity has been threatened by rapid economic development in recent decades. Species are experiencing much higher extinction rate than the natural extinction rate (referred as the extinction rate without human interference) because of the environmental changes caused by human activities such as population growth, deforestation and over-exploitation of resources in global scales [6, 15]. The accelerated rate of species extinction could result in continued loss of biodiversity, which poses a risk to the health of the ecosystem and then could cause serious consequences to human health and the environment. Therefore, biodiversity conservation is an important task for scientists and the whole human population.

Biodiversity can be explored at several levels of organization [1, 6, 7]. *Genetic diversity* refers to the variation of genes within and between populations. Genetic diversity strengthens species by increasing their adaptibility to environmental changes and resistance to diseases, which reduces the risk of extinction. *Species diversity* focuses on the number of different species on the Earth, and *ecosystem diversity* is the variety of living and non-living organisms. All three levels of diversity play an important role in the exploration of the process and pattern of how species and ecosytems interact.

Genetic diversity provides a foundation for understanding diversity at higher levels of organization. Traditionally, genetic diversity has been approximated based on morphological characteristics or biochemical methods [8, 9, 10]. However, morphological methods are very time-consuming and expensive, and these characters are susceptible to environmental variations that may be at odds with genetic variation and evolutionary history [8, 10]. The genetic variability assessed via biochemical methods reflects only a small number of gene products which do not provide detailed information about variation [10, 106].

Molecular genetic markers, which capture variation at the nucleotide level, offer a more direct alternative to the morphological and biochemical approaches for the estimation of genetic diversity within a species. A molecular genetic marker is a DNA sequence with a specific location in the genome that varies among individuals in a useful way. Useful molecular markers are those that can be easily and reliably detected in DNA samples, show polymorphisms (variations) in DNA sequence and may or may not be under selection [13]. Detailed population studies often require many genetic markers, but until recently it was not feasible to characterize more than a few molecular markers using traditional DNA sequencing technologies (e.g. Sanger sequencing). With the development of next-generation DNA sequencing (NGS) technologies, the detection of molecular markers can be achieved at a much larger scale but in very short time and at lower cost. NGS techniques can generate millions of DNA fragments in a short period of time and massively decrease the time and costs to sequence any set from single markers to entire genomes.

Genetic diversity arises via the process of mutation [1], which provides the variations on which different evolutionary processes can act. Variants at a particular location in the genome (a *locus*) are termed *alleles.* Natural selection can act on alleles that modify the phenotype of an individual, such that individuals who are better adapted to their environment produce more offspring with a corresponding increase in the frequency of the fitter allele. Non-adaptive or neutral forces, by contrast, are not acted upon by selection, and thus may not correlate with environmental factors; however, these neutral alleles can still be used to differentiate populations based on differences in allele frequencies. The relationships between genetic diversity, population structure, and environmental factors are key to understanding biodiversity. It is therefore of great importance to explore these complex relationships to understand patterns of adaptation, migration, and population structuring, which can in turn be used to predict future patterns of biodiversity and inform management practices to preserve diversity.

Many types of molecular genetic markers have been developed to detect genetic variations, including microsatellites, single nucleotide polymorphisms (SNPs), amplified fragment length polymorphisms (AFLPs) and restriction fragment length polymorphisms (RFLPs) [8, 10]; in the sections below we focus on microsatellites and SNPs as targets for computational analysis. In this thesis we focus on *diploid* eukaryotic organisms that have two copies of each chromosome. Diploid individuals can have a maximum of two different alleles at a given locus: a *homozygous* individual has two identical alleles, whereas a *heterozygous* individual has different alleles on each of its chromosomes at the corresponding locus. Using the simple notation of “A” and “a” as two alleles at a particular locus, a homozygous individual can have genotypes “AA” or “aa”, while a heterozygous individual would be of type “Aa”. A desirable molecular marker should have the following attributes:

- Reliably detectable at a specific locus: *paralogous* sequences arise when genetic material is duplicated, leading to two separate loci that accumulate mutations independently. If marker detection has poor fidelity or cannot distinguish these two loci, then the inferred genetic information will be inaccurate.
- Highly polymorphic: if one allele is extremely rare (e.g., has a frequency < 1%) in one or more populations, it is unlikely to be useful for population studies. *Polymorphic* sites, typically defined as having at least two alleles with > 1% frequency in the population, are suitable candidate markers [107].
- *Co-dominant inheritance* [8]: homozygotes and heterozygotes are distinguishable at this type of marker. A co-dominant marker can differentiate AA from Aa, which is essential for estimating allele and genotype frequencies in a population.
- High abundance in the genome: markers should belong to a type that occurs at numerous locations throughout the genome. Such markers are easier to discover, and can aid in identifying areas for further genomic investigation.

## 1.1 Genetic Markers: Microsatellites And SNPs

This thesis develops new computational tools to infer genetic variation based on two types of genetic marker: microsatellites and SNPs. In this section we introduce these two types of marker and explain the key attributes that motivate the projects described in chapters 2 and 4.

### 1.1.1 Microsatellite Markers

Microsatellites, also called as variable number tandem repeats (VNTR), have been one of the most popular genetic markers in many applications of molecular ecology and other fields of biology such as population genetics, kinship analysis, association studies and genetic mapping due to their high degrees of polymorphism and ubiquity in eukaryotic genomes [16, 17, 18]. Tandem repeats consist of short *repeat motifs*, typically 1-6 nucleotides in length, that are repeated between 5 and 40+ times. Microsatellites mutate frequently due to errors in DNA replication, and the lack of selection at most tandem-repeat loci means that much of the introduced diversity is not detrimental to the organism, and is consequently more likely to be preserved. The resulting variation in the size of tandem repeat arrays contributes to large numbers of allelic variants among individuals within a population. A microsatellite-containing sequence consists of tandem repeats and flanking regions that are DNA sequence located at both sides of repeat array (as shown in Figure 1).

Microsatellites are highly abundant throughout many genomes: for example, microsatellites occur approximately once every 30,000 base pairs in the human genome [29, 30, 121]. Furthermore, the high levels of heterozygosity (i.e., a high incidence of heterozygotes in many populations) and polymorphism, high information content and relatively high mutation rates associated with microsatellites give this type of genetic maker the ability to provide contempory estimates of migration and the capability of detecting differences between closely related individuals [16, 17, 19]. Microsatellites were used as markers before the advent of NGS, but they remain a popular choice in many studies due to their positive attributes.

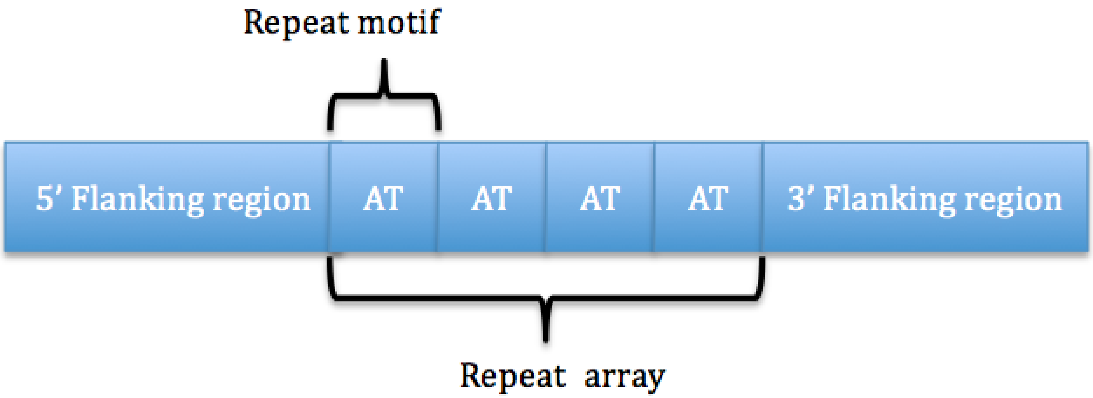


Figure 1 Components of a microsatellite-containing sequence. The dinucleotide repeat motif "AT" and the number of repeats is 4. Flanking regions of a microsatellite locus are less variable for all individuals of a species.

The traditional way to genotype microsatellites is to "amplify" microsatellites using the polymerase chain reaction (PCR), which generates many copies of the microsatellite region. The amplified DNA fragments are then subjected to electrophoresis, which separates DNA molecules by size based on different migration rates through a resistive medium such as a gel. Microsatellite genotyping suffers from several limitations, which stem primarilty from reliance on electrophoretic methods and the necessarily imperfect inference of genotypes from DNA fragment mobility data. Visual inference of the length distribution and inference of genotypes is very time consuming, owing to the difficulties in distinguishing true alleles from a variety of artifacts that could occur during PCR amplification and sequencing of microsatellites. In addition, the estimated sizes of microsatellite alleles may differ among different laboratories and electrophoretic platforms, which hinders the standardization of microsatellite data for some specific research objectives [17, 23].

### 1.1.2 Single Nucleotide Polymorphisms

A SNP is a variation at a single nucleotide in the DNA sequence among individuals, that occurs with a frequency greater than 1% in a population [110]. The nucleotides in the

DNA sequence have only four types: A, T, G, C. Therefore, in principle, SNPs could have at most four variations at each allele. However, in reality, large proportions of SNPs have only two types of variant [110], which indicates that only two alleles can occur at a given SNP. SNPs are therefore often treated as *bi-allelic* genetic makers.

Even though microsatellites have been used in molecular studies since the 1990s, SNPs have seen increasing use in many research fields of biology such as population genetics and genome-wide association studies. SNPs have a variety of advantages over microsatellites, including:

1) Greater abundance within genomes: SNPs are the most abundant genetic variants in the human genome and occur on average about once every 500 base pairs in many wild animal populations [29, 30].
2) Greater amenability to different high-throughput genotyping techniques [17].
3) Lower genotyping error rates: the error rate in SNP sequencing is typically < 1%, whereas for microsatellites, the error rate typically ranges from 5% to 10% [111].
4) Cost-effective: the genotyping cost ranges from $0.002 per SNP/genotype to around $0.15 per SNP/genotype.
5) SNPs are easier and cheaper to standardize between platforms and laboratories than microsatellites [35].

However, these advantages of SNPs are not always realized or relevant. For example, SNPs are typically *bi-allelic*, whereas microsatellite loci often have >2 alleles and can thus carry more information. Additionally, the heterozygosity of SNPs is much lower than microsatellites. A common estimate states that two to six times as many SNPs are required to gain the equivalent information content as microsatellites in individual identification, parentage and relatedness research [17, 31, 32]. This is because larger allelic diversity and higher levels of heterozygosity would have a larger possibility that genotypes of individuals in a population would differ, increasing the power to distinguish closely related individuals [33]. As an illustration, Herraez et al. [32] carried out a simulation study aimed at comparing the performance of microsatellites and SNPs for the individual identification and parentage analysis of a Galloway cattle population. The results showed that 33 SNP markers conferred approximately the same amount of information as 14 microsatellite markers.

Even though SNPs contain, on average, less information than microsatellites, many researchers still prefer to use SNPs in genetic analyses. This is because recent studies have demonstrated that SNPs could have better or similar performance in many broad-scale applications of population structure assessment, pedigree assignments and parentage analysis by increasing the total number of SNPs to approximately 6 to even 10 times the number of microsatellites [28, 34, 36, 37]. Furthermore, owing to the lower rates of genotyping error and lower costs of genotyping for SNPs, using large sets of SNPs is still a better option when compared to a small set of microsatellites. However, the efficiencies and economies associated with SNP genotyping are best realized at large scales: many loci (minimally, hundreds) genotyped in many individuals. Such large-scale genotyping efforts require large initial investments in set-up costs, which may not be cost-effective when experimental needs require only more modest numbers of loci [40]. Small-scale genotyping of SNPs can be more costly than genotyping of microsatellites, particularly when the lower information content per locus is considered. Although the use of technologies such as microfluidic devices can lower the cost of small-scale SNP assays, these technologies require access to expensive and specialized instrumentation [41].

Although SNPs have been reported to be superior to microsatellites in many applications, SNPs are not suitable for all applications. For example, Hess et al. [38] illustrated that microsatellites outperformed SNPs in uncovering fine-scale relationships for salmon genetic stock identification. Forstmeier et al. [42] reported better performance of microsatellites in inferring correlations between heterozygosity and fitness-related traits. Some studies suggested that a combination of microsatellite and SNP markers could be more effective than using these two markers independently [28, 38]. In addition, with the advent of NGS technologies, many disadvantages of microsatellites such as high development cost and low genotyping throughput will be overcome by emerging methods. Therefore, microsatellites will continue to be the genetic marker of choice for some specific reseach objectives in the future.

The recent advent of NGS technologies has also mitigated some of the challenges of microsatellite development and use, including the detection of microsatellite loci that are good candidates for PCR and the analysis of microsatellite variations using NGS data [22, 24, 25, 26]. NGS technologies can also directly read microsatellite genotypes from raw sequence data without the need for electrophoresis. The sequencing read lengths (currently 85 to 900 base pairs per read) [113, 114] for several NGS systems can encompass the range of allele sizes for large numbers of microsatellite loci across any genome of interest in a single run. NGS also offers a variety of additional benefits for microsatellite genotyping [52, 53]:

1) It directly sequences the fragments and provides nucleotide sequence data rather than estimating the total length of repeats by electrophoresis.
2) Microsatellite sequences can have SNPs in repeats or the flanking region of identical-length alleles, adding to their discriminative power if DNA sequencing is used in place of electrophoresis.
3) Any length artifacts due to PCR errors can be analyzed and filtered by investigating the NGS reads. This is contrast to the electrophoretic method, which does not provide sequence reads that can be used to analyze the reason why length artifacts occur.

Before this potential can be achieved, suitable software is needed to automatically convert raw amplicon sequence data to multilocus microsatellite genotypes. MicNeSs [22] is the only program that can infer microsatellite genotypes from NGS data. However, this program suffers from several disadvantages including long running time and relatively high genotyping error rate owing to the length artifacts that occur during PCR amplification. Furthermore, this program does not offer a user-friendly interface. Therefore, better software is required to automate the genotyping process. Before developing this software, several challenges must be addressed in order to enable automation.

## 1.2 Automating The Genotyping Of Microsatellites From NGS Data

In order to gain these advantages associated with NGS, suitable software needs to be developed to deal with several challenges that impede the inference of microsatellite genotypes from NGS sequence data. One of the main challenges has to do with the sequencing errors that could arise at any step throughout the laboratory processes including sample preparation, library construction and sequencing [54, 55]. Among sequencing errors, some of the most prevalent types are nucleotide substitution errors and insertion/deletion (indel) errors. The overall error rate ranges from 0.001 to 0.15 across current NGS platforms [112]. Furthermore, the lengths of NGS sequence reads are usually shorter than those generated by traditional sequencing platforms. The number of tandem repeats in microsatellites is highly variable; larger microsatellite-containing sequences may even be longer than the length of the sequencing read. In this case, precise sizing of microsatellite alleles can be impossible owing to the possibility that the contents of microsatellites cannot be fully present in the NGS sequence reads.

Another challenge is a variety of errors or experimental artifacts that can occur during PCR amplification and sequencing of microsatellites. Chief among these are *stutter artifacts,* in which replication “slippage” during PCR amplification generates products that differ in size by multiples of the microsatellite repeat unit from the ‘true’ allele length. Replication slippage is the misalignment of two DNA strands during replication, which leads to deletions or insertions of repeat units in the microsatellite sequence [56]. Stutter artifacts increase the difficulties of scoring alleles accurately in the case when the two alleles in a heterozygous individual differ in size by only a few nucleotides [57]. PCR, which is used to generate many allele copies for downstream analysis, can also preferentially amplify certain alleles, leading to different relative abundances of products. In the extreme, PCR can fail to amplify one or both alleles (termed as allelic dropout); this can be caused by amplification bias favouring small alleles, or low starting DNA template quantity or quality. Allelic dropout may result in a false assignment in which a heterozygous individual may be inferred as being a homozygote.

The scale of microsatellite NGS data creates the need for automated techniques for allele and genotype assignment from noisy sequencing data, and inference of population structure from these assignments. String representations simplify the analysis of DNA sequences, but string-comparison algorithms must accommodate the possibility of SNPs and indels in the interpretation of microsatellite data. Automated algorithms must also be able to handle experimental artifacts, applying knowledge of the most common erroneous patterns to identify the correct alleles. Finally, interactive visualizations are needed to allow researchers to manually curate and interpret results at the level of the entire population, which in current studies can comprise hundreds or even thousands of individuals.

## 1.3 Machine-learning Analysis Of Large SNP Datasets In Landscape Genetics

### 1.3.1 Introduction To Landscape Genetics

Landscape genetics aims to understand how features such as environmental and geographical variables affect genetic variations in concert with evolutionary and ecological processes [15, 67, 68]. This research area integrates methodological developments in population genetics, landscape ecology, and statistics to better understand patterns of genetic variation within species, and consequently provide critical information and implications for wildlife conservation and management [69, 70, 71]. As an example, Manel et al. [118] used a simple statistical model (linear regression) that regressed the allele frequencies on environmental variables to identify environmental factors that are the main drivers of adaptive genetic variation and also the associated loci while accounting for spatial effects using Moran’s eigenvector maps (MEM).

However, the processes that drive patterns of genetic variability are complicated (as shown in Figure 2). Although they are not subject to selection, neutral genetic variations play a central role in understanding spatial relationships among populations [117]. Adaptive genetic variations, which do affect organism fitness [72], can differentiate individuals inhabiting a heterogenous environment. Individuals with mutations in genes that provide better adaptation to environmental factors are more likely to survive in that environment. This evolutionary process is known as *local adaptation*. These mutations convert to adaptive genetic variation through the process of natural selection. Therefore, adaptive variants present different patterns from neutral variations, which show genetic differentiations in some individuals but not in a whole population.

The preservation of these adaptive genetic differentiations and the understanding of local adaptation provide valuable insights in the maintainance of endangered species' potential in response to ecological pressures. In general, the key tasks of landscape genetics are to uncover the underlying processes that drive these two types of genetic differentiations, distinguish local adaptation from other non-adaptive patterns, and correlate the detected adaptive signatures with landscape parameters, especially the environmental conditions of greatest interest.

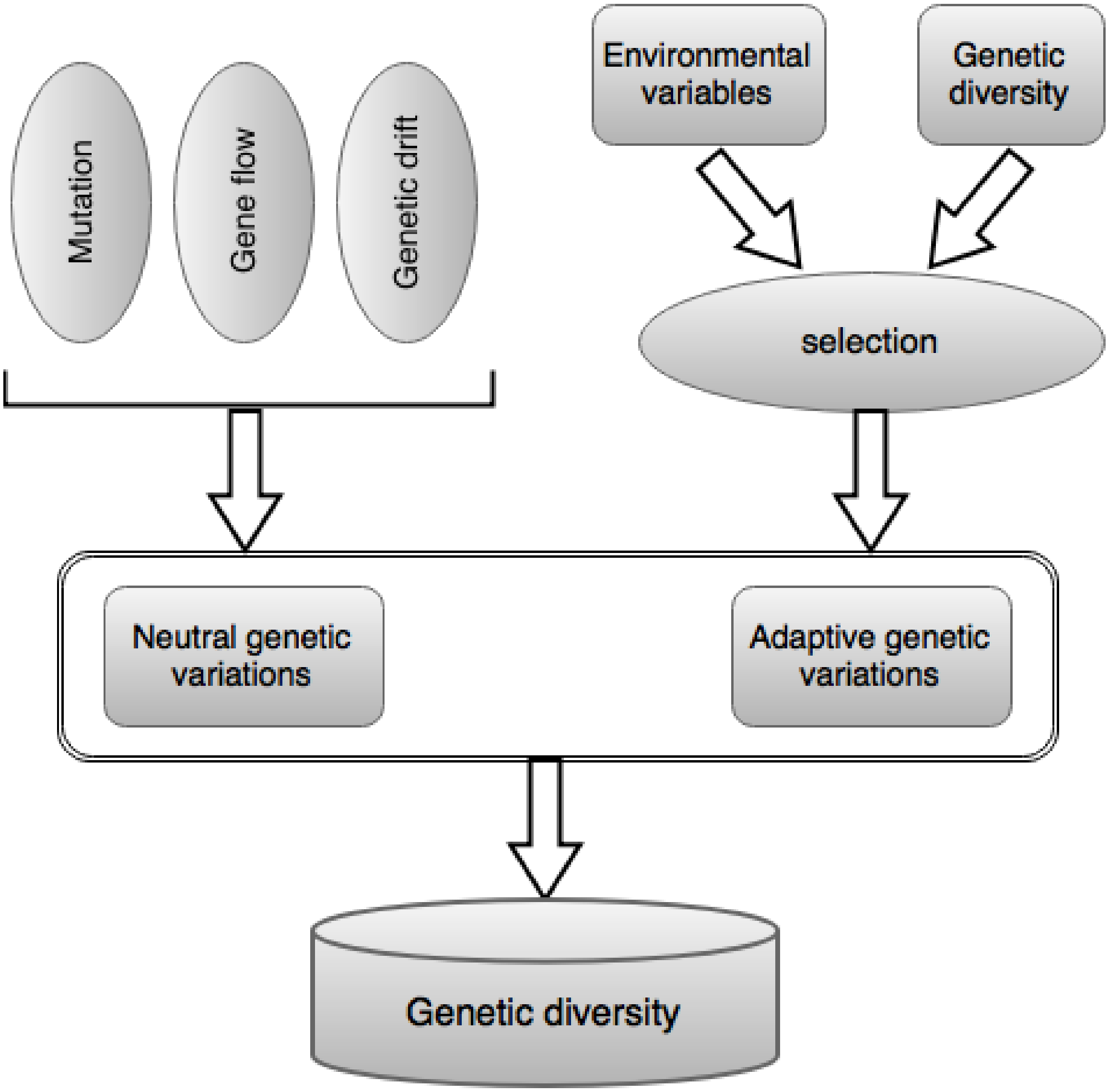


Figure 2 The role of evolutionary driving forces in generating patterns of genetic variation (see [122] for more details of how these evolutionary processes contribute to genetic variations).

### 1.3.2 Environmental Associations Analysis In Seascape Genetics

It has long been recognized by researchers that environmental features such as ocean temperature, salinity and surface ocean pH are among the main driving factors that influence the genetic structure of marine populations [73, 74]. In order to identify genetic differences that drive local adaptation, marine researchers use *seascape genetics* (the application of landscape genetics approaches to marine populations). Environmental association analysis commonly uses a three-step approach, which includes sampling design, the collection of genetic and environmental data, and the correlation of genetic data with environmental data using statistical models [77]. The choice of sampling method depends on the motivations of the proposed landscape genetics study. More specifically, the choices of which pattern of genetic variations to investigate determines the sampling strategy.

Elucidation of key underlying population and environmental patterns is a challenging problem, and many statistical methods have been developed to distinguish different types of variation and explain them in terms of geographic and environmental factors. However, the measurement of gene-environment interactions might be complicated by several factors, including:

- Sampling design will impact the successful detection of adaptive genetic differentiations. Genetic and environmental data are the most fundamental components for building a good gene-environment association model. A strategic sampling design can give a good presentation of genetic and environmental data, which helps to uncover an unbiased sign of selection.
- A large number of environmental parameters may be required to uncover loci subjected to local adaptation; many of these environmental variables will be highly correlated and many that show correlations with genetic data will not be causal. For example, salinity can be highly correlated with temperature and precipitation; therefore, genetic variation that is driven by salinity will also show high correlations with the other factors. These highly correlated environmental

variables increase the difficulty of investigating the contributions of different variables driving adaptive gene variants.

- The statistical models used to identify associations between adaptive loci and environmental parameters need to account for other driving factors that could cause genetic differentiations, including non-adaptive forces that have no strong relationships with environmental factors. However, the appropriate estimation of the effects of these processes remains a challenge. As a result, it is possible that correlations with some environmental factors will be identified as significant, even where no causal relationship exists.

Many methods have been used to overcome these challenges. For example, a common technique to address the effects of collinearity among environmental factors is to use principal component analysis (PCA) to extract highly correlated aspects of environmental variables as principal components or *metavariables* that capture maximal amounts of covariance. Some studies use the transformed PCA metavariables in place of the original variables to eliminate redundancy in the data, but at the expense of losing the direct associations with specific environmental factors.

Many statistical models have been developed to investigate the effects of environmental heterogeneity on genetic population structure. One of the most common methods is the fixation index ($F_{ST}$) [123] which identifies loci that show strong differences in allele frequencies among populations, and then correlate these loci ('outliers') with environmental variables to check the existence of some strong relationships with the environment. However, $F_{ST}$ may not detect some adaptive genetic variations in some cases when local adaptation has few impacts on allele frequencies [77]. Coop et al. [75] proposed a Bayesian approach, which is different from $F_{ST}$ method, to identify environment-associated adaptive loci. It builds a null model that esimates the allele frequency differences between populations, and then uses the null model against a model that searches for relationships between allele frequencies and environment to account for neutral effects [75]. Machine-learning methods are emerging as a new set of approaches to build complex, non-linear models that link environmental and genotype data.

Our proposed machine-learning method differs from other approaches. It uses environmental variables as predictors and genotypes at all loci as outputs to construct a classification model, and chooses loci that show good predictions as possible adaptive loci. Then we use statistical models to explore the correlations with environment while controlling for some neutral effects and spatial effects in particular. As an example, Holliday et al. [124] used random forests to identify subsets of adaptive loci that can optimally predict the phenotype. In order to control for the effects of population structure on the identified associations, genoypes and phenotypes were both regressed on the estimated cluster membership, and residuals from the two regressions were used as predictors and output to build another random forest model.

## 1.4 Contributions And Thesis Outline

The remainder of this thesis is organized into four chapters that describe two novel contributions.

Our first contribution is MEGASAT, new software that allows the automated and rapid inference of multilocus genotypes from microsatellite NGS data. MEGASAT has three primary functions: (i) separate highly complicated NGS data encompassing large amounts of loci into locus-specific files, based on primer and flanking sequences; (ii) automate the scoring of microsatellite genotypes, using customized decision rules to account for amplification artifacts; and (iii) generate plot files (histograms of sequence length-frequency distributions) for manual verification and updating of genotypes. MEGASAT outputs predicted multilocus genotypes to tab-delimited text files that can be imported into spreadsheets. In **Chapter 2** we describe the algorithms in MEGASAT, and in **Chapter 3** we apply it to a large dataset of guppies collected from Trinidad using 43 microsatellites. We further demonstrate a high level of reproducibility and accuracy of

MEGASAT-called microsatellite genotypes by a combination of genotyping error estimation methods.

Our second contribution, described in **Chapter 4**, is a series of new random-forest-based techniques for seascape genetic analysis. We propose a novel workflow to identify environment-associated adaptive variations using random forest while utilizing some statistical methods to control for spatial effects. And then we use some ordination techniques to explore which environmental variables present strong correlations with identified adaptive variations. We apply this workflow into genetic and environmental data collected from the sea scallop *Placopecten magellanicus* across 12 locations from Newfoundland to the Mid-Atlantic Bight.

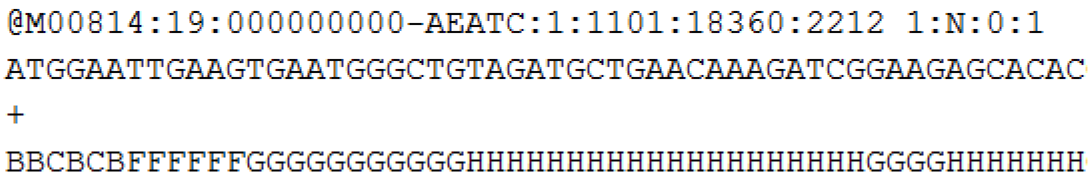
In **Chapter 5** we conclude with a summary of the thesis and propose future work.

# CHAPTER 2 INFERENCE OF MICROSATELLITE GENOTYPES FROM NGS DATA

In this chapter we describe MEGASAT, a software package to automatically and rapidly infer microsatellite genotypes from NGS data. MEGASAT has two main algorithms and one script to obtain the three main functions as we stated in section 1.4. In this chapter, we describe the required input files, workflow and algorithms implemented in MEGASAT.

## 2.1 INPUT FILES FOR MEGASAT

MEGASAT accepts FASTQ files and a tab-delimited text file as input files. FASTQ is a text-based file that stores nucleotide sequences and the corresponding quality scores. Every read in a FASTQ file consists of four lines: sequence identifier, sequence, quality score identifier line and quality scores (as shown in Figure 3). MEGASAT reads only the DNA sequence line which consists of the nucleotide characters "A T G C" to identify microsatellite-containing sequences.



```
@M00814:19:000000000-AEATC:1:1101:18360:2212 1:N:0:1
ATGGAATTGAAGTGAATGGGCTGTAGATGCTGAACAAAGATCGGAAGAGCACAC
+
BBCBCBFFFFFFGGGGGGGGGGHHHHHHHHHHHHHHHHHHHGGGGHHHHHHH
```

Figure 3 Example of a sequence record in a FASTQ file. The four lines represent sequence identifier, sequence, quality score identifier line and quality scores.

The other input file is a user-definable text file that stores the component information for all the microsatellite loci (as shown in Table 1). Microsatellite-containing amplicon sequences have the following components: forward primer (FP), forward flank (FF), microsatellite repeat array (MRA), reverse flank (RF) and reverse-complement primer

(RP) (as shown in Figure 5). The FP and RP are the designed primers for PCR to amplify the microsatellites. Once microsatellites have been amplified, the primers have no biological meaning apart from a correctness check. MRA is the region of tandem repeats. The FF and RF are flanking regions located on each side of the MRA. These flanking regions are important for designing locus-specific primers, and some studies illustrated their critical role in applications such as phylogenetic inference [19]. The first row in the input file specifies the microsatellite locus name and the first column represents the required column name. Each subsequent row in the text file corresponds to a microsatellite locus, with each column containing the expected sequence for each of the five components of that locus. The last column in the text file is optional, and allows users to define specific parameters for the genotype assignments. If these are not specified, then default parameter settings will be used. MEGASAT uses the locus-specific information as reference data to search for reads that contain the microsatellites and call genotypes based on the identified reads.

Table 1 Format of the input tab-delimited text file.

| Locus Name | Forward Primer | Reverse-complement Primer | Reverse Flank | Forward Flank | Repeat Unit | Ratios group (optional) |
|---|---|---|---|---|---|---|
| Locus 1 | AACCTG | GGCCTA | GGCC | CATGCT | AC | |
| Locus 2 | CCTGAC | TTAACG | ATAG | TGACC | TG | |
| … | | | | | | |
| Locus n | TCGACT | ACCTGC | TAGCC | CCT | ACG | |

## 2.2 CORE FUNCTIONS INCORPORATED IN MEGASAT

MEGASAT uses the information provided in the text file to search for patterns in the sequence reads. However, sequencing errors in the NGS reads can cause a microsatellite-containing sequence to be discarded because of one or more mismatches between the reference primer or microsatellite flank sequences and the user-provided reference sequence used to identify the locus. The sequencing error rate on the Illumina platform (the NGS platform used to generate our trial data) is around 0.1% [112]. Even though the error rate is low enough to be useful in genotyping, billions of NGS reads will contain a large number of errors. The primary error type is nucleotide substitution errors, which can impact any component of the microsatellite-containing reads. To overcome this problem, MEGASAT has a function to allow a tolerance for mismatches when matching reference sequences for primers and flanking regions with observed sequences. The number of allowed mismatches (parameter *m*) is a user-controlled variable. The function traverses each sequence in the FASTQ files and searches for the starting position of a near-exact match in the target sequence, which can be used to enable trimming of primers.

Another important function in MEGASAT helps to find the end of the MRA. The function searches for the longest continuous MRA while also incorporating tolerance for sequencing errors (i.e., “fuzzy” matching) in one or two microsatellite repeat units. This function can also be used to find the end of the reverse flank (RF) when only a few bases of the reverse-complement primer (RP) are present in the sequence.

The third function deals with cases when the complete reverse primer complement of a large microsatellite allele is not present in the NGS read at all. This function uses the Hamming distance (the number of differences between two strings of equal length) to find the starting point of any incomplete reverse primer complement. Based on the length of the reference reverse-complement primer (RP), it progressively removes the last few bases of the observed reads and calculates the Hamming distances between the resulting truncated sequence and all or part of reference reverse primer complement. Then the Hamming distance is divided by the length of the removed sequence to calculate an error ratio. The truncated sequence with the smallest error ratio will be the incomplete RP and then it can be easily trimmed off from the sequence. We use this function because it allows the primers to be trimmed off at the correct position even if there are length variations in the RF. Figure 4 demonstrates this procedure.

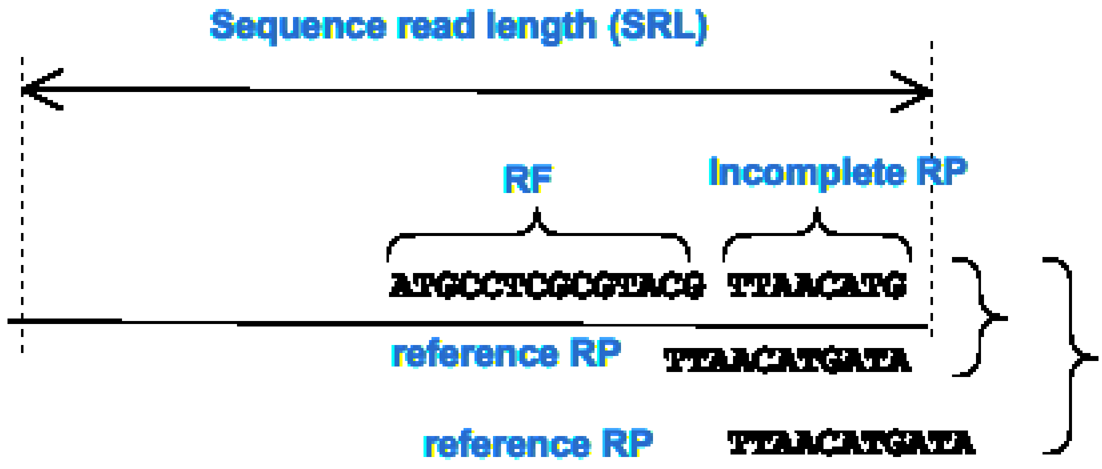


Figure 4 Using the Hamming distance to locate the starting point of an incomplete reverse-complement primer. The last bases of the reads are progressively removed and the Hamming distance computed between the truncated sequence and the all or part of reference RP. The truncated sequence with the smallest error ratio (Hamming distance divided by the length of subtracted string) will be the correct incomplete RP.

MEGASAT employs two different algorithms to allow the automated scoring of microsatellite alleles from the sequences in the FASTQ files. These files contain reads from all targeted microsatellite loci, which are distinguished by the PCR primers that were used for amplification. The first algorithm sorts the input reads (all sequences for a given sample, found in one FASTQ file) into locus-specific files containing only the relevant reads of interest by discarding reads which do not contain the locus-specific priming and flanking sequence. This algorithm also locates the correct boundary of the microsatellite flanking regions. The second algorithm takes as input the trimmed reads generated by the first algorithm and uses scoring parameters to deal with errors such as stutter that arise from PCR amplification. These algorithms are described in the following sections.

## 2.3 SORTING THE INPUT SEQUENCES BY LOCUS

MEGASAT uses reference data for each microsatellite locus to identify sequences associated with individual loci and to remove primer sequences. The FF and RF portions of the microsatellite amplicon are retained as part of the allele, for two reasons: (i) The flanking sequences may themselves contain insertion or deletion mutations that contribute to allelic diversity. (ii) The boundaries of the MRA may not be clear in some loci; retaining the two flanking sequences avoids the need in most cases to define exact boundaries for the MRA, although our script includes the ability to define the boundary of the MRA when needed (see below).

The process of identifying and trimming off primers may be complicated by one or more factors, including sequencing errors and some cases that the size of the amplified microsatellite allele exceeds the length of NGS reads (Figure 5 b-d). Nowadays, the read lengths of NGS data typically range from 125 to 350 bases, which are longer than most, but not all, microsatellite repeat arrays [113, 114]. It is possible that all or part of the reverse-complement primer, or even the reverse flank, may be absent from the sequenced portion of the amplicon. Therefore, in order to correctly locate the boundary of microsatellite flanking regions, we implemented the following algorithm in MEGASAT. The detailed algorithm flowchart is shown in Appendix Figure A.1.

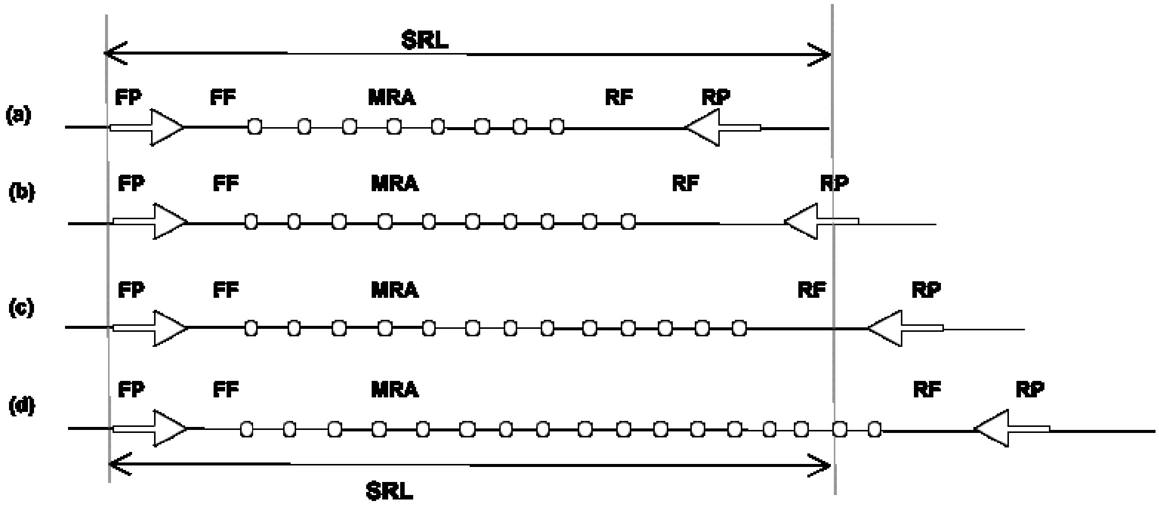


Figure 5 Loss of information in short DNA sequence read data. (a) The forward primer (FP), forward flank (FF), microsatellite repeat array (MRA), reverse flank (RF) and reverse-complement primer (RP) are present within the sequence read length (SRL). (b) A longer MRA leaves only a few bases of the 3' end of the RP within the SRL. (c) An even longer MRA causes only part of the RF to be present with the SRL. (d) An MRA that extends past the end of the SRL. In cases (a) and (b), MEGASAT is able to

detect the 3’ end of the RP, and directly ascertain the length of the amplified microsatellite allele, which consists of FF+MRA+RF. In case (c), MEGASAT detects the end of the MRA and adds the reference length for the RF to infer the allele length. In case (d), MEGASAT detects the length of the FF and adds an integer value, to denote alleles that exceed the length of the SRL.

First, MEGASAT reads each sequence in the FASTQ or FASTA file and checks if there is a fuzzy match for the forward primer of any locus in each sequence. If there are no matches, it will discard the sequence. If there is a match, MEGASAT will delete the forward primer from the sequence and check if there is fuzzy match for the corresponding RP of that locus. If so, it will delete the RP from this sequence and then check the trimmed sequence to see if it contains a fuzzy match for 5’ and 3’ flanking sequences (FF and RF), and for the repeat array of that locus (Figure 5a). If it does, MEGASAT will retain the trimmed sequence in a hash table, henceforth referred to as the genotyping set. If not, it will add the deleted RP sequences back to the reads and put the sequences into a hash table that contains all the discarded sequences for each locus.

In some cases the amplified sequence does not contain the complete RP, because the amplicon sequence exceeds the length of the sequence read (Figure 5b). In such cases, MEGASAT checks the sequence to see if it contains a fuzzy match for the FF and RF sequences, and if the end of the microsatellite repeat array occurs before the end of the sequence. If this is the case, MEGASAT checks if the length of FP + FF + MRA $\leq$ the length of SRL – RF – *l*. The “*l*” is an integer variable that should be in the range between the half of the length of RP and the total length of RP. This determines if the sequence contains at least *l* bases of the incomplete RP; if it does, MEGASAT identifies the starting point of the incomplete RP. In this case the incomplete RP is removed, and the trimmed sequence is added to the genotyping set.

If at least *l* bases of the RP are not present, MEGASAT searches for the end of the reverse flank (RF) using the end point of the MRA plus the length of RF. In this case the incomplete RP is removed, and the trimmed sequence is added to the genotyping set.

In some cases the sequence does not contain the complete reverse flank (RF). If the RF is not detected, MEGASAT checks if the sequence has a fuzzy match for the FF, and if the end of the MRA occurs before the end of the sequence (Figure 5c). In such cases, MEGASAT locates the end point of the MRA. The incomplete amplicon RF sequence is then deleted, and the reference RF sequence is added to the end point of the MRA, to produce an untruncated sequence, which is added to the genotyping set.

Failure to detect the end of the MRA implies the presence of an allele that is too large to contain identifiable RF sequence (Figure 5d). In this case, an integer variable which is the sum of the length of input sequence plus 100 is added to the observed length of the FF. In our guppy experiment described in Section 3.1 below, the value of the integer variable was 250 since our trial data had a read length of 150 bp, which is a common length generated by NGS sequencing runs. The addition of the large integer value clearly distinguishes such large alleles from those that can be accurately scored within the bounds of the read length. The inclusion of the FF length in the allele length value allows different classes of large alleles to be distinguished if there is variation in the length of the FF. The sequences are added to the genotyping set.

## 2.4 INFERENCE OF MICROSATELLITE GENOTYPES

The second major step in MEGASAT is to predict microsatellite genotypes. Once the genotyping set is complete, MEGASAT determines the lengths of all the trimmed sequences for each locus and each individual, and records the count of sequence length variants for each locus in each individual. These records will be printed into text files and used as input files for another script to visualize the microsatellite data used to infer genotypes. When the hash tables containing the distributions of length variants for each locus in each individual are constructed, the next step is to infer allele sizes, and then call genotypes from those length variants. However, the process of inferring alleles and genotypes is complicated by the stutter, allele dropout and indel length artifacts introduced in Section 1.2 that can arise during PCR amplification and sequencing.

Amplification stutter and sequencing indels could result in ‘false’ length variants being misinterpreted as microsatellite alleles, and length artifacts that arise during amplification can cause heterozygotes to be mis-scored as homozygotes. MEGASAT employs a number of rules to distinguish true alleles from those complicated length artifacts, and infer genotypes. The process of allele and genotype inference is described briefly below; the process is outlined in more detail in Algorithm 1.

The first task is to determine how many length variants are present. If there is no length variant, the genotypes will be scored as "0 0". If there is only one length variant present, which means that only one allele is amplified, then the next step is to check if the number of copies (read depth) of this length variant is larger than a user-definable minimum threshold of observations *n*. If so, the locus is scored as homozygous. If two or more length variants are present, MEGASAT first identifies those variants with a minimum read depth of *n*. MEGASAT compares the size of the second most abundant length variant (A2) with the most common length variant (A1) (Figure 6). Based on the relative length of A1 and A2, it uses a set of criteria to calculate the depth ratios of as many as four (i.e., from A1 to A4) of the most common length variants relative to A1. Then it compares the depth ratios with user-specified thresholds to determine which amplification products should be called as real alleles. This procedure minimizes the risk of miscalling stutter artifacts as real alleles.

As an illustration, for the case when length (A1) is smaller than length (A2) (Figure 6a), the ratio *q* of the second most abundant length variant to the most abundant variant (A2/A1) is calculated. If *q* is less than a user-defined threshold (default value R1= 0.15), the genotype is scored as homozygous (in this case, A1 A1). If *q* is greater than or equal to the threshold value, additional scenarios are considered that take into account the possibility of complex stutter patterns associated with the larger allele. The script checks for the presence of a next-most abundant length variant that is one microsatellite repeat unit larger than the second most abundant variant (A3 in Figure 6a). The script calculates the ratio A3/A2, and compares it to a user-defined threshold value (default value R3 = 0.7). If the ratio is greater than or equal to R3, then the third most abundant variant (A3) is considered the large allele (the genotype is A1 A3). Therefore, genotype "A1 A3" will be called in the case of Figure 6(a).

If the length (A1) is larger than the length (A2) (as shown in Figure 6 (b)), the script next checks if the difference between A1 and A2 is larger than or equal to one microsatellite repeat unit and the ratio depth(A2)/depth(A1) is larger than or equal to a user-defined threshold (default value R4= 0.6). In Figure 6b, the ratio is smaller than the threshold value. The script then checks for the presence of a next-most abundant length variant that is larger than the most abundant variant (A3 in Figure 6b) and compares the ratio depth(A3)/depth(A1) to a user-defined threshold (R6; default value = 0.2). Since the ratio in Figure 6b is larger than R6, the third-most abundant variant (A3) is considered the large allele (the genotype is A1 A3).

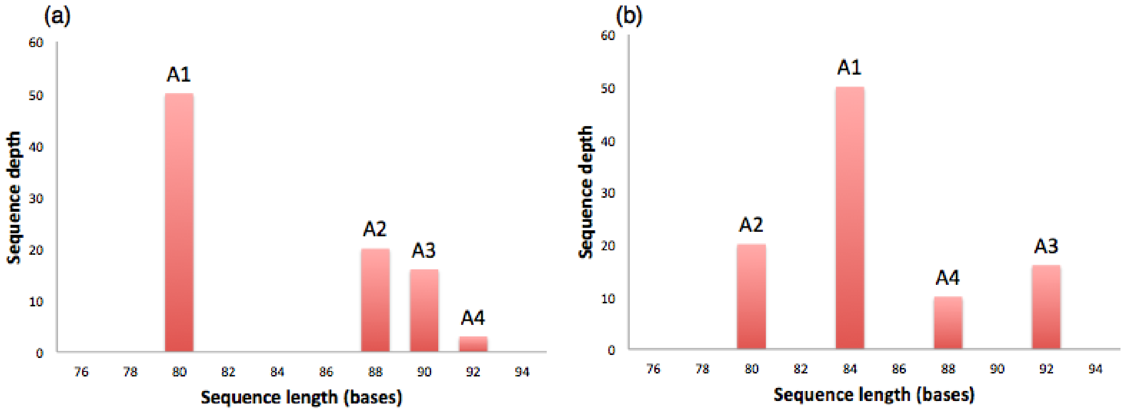


Figure 6 Examples of two common modes of length-variant distributions. Both distributions lead to genotype calls of "A1 A3" via different paths in Algorithm 1.

Algorithm 1 presents more details about how to infer microsatellite genotypes based on those trimmed data sets we got from the previous step. The variable "$l_{RU}$" in this algorithm refers to the length of one repeat unit of a microsatellite locus.

**Algorithm 1** Inference of microsatellite genotypes

```
Input: the first four abundant length variants (A1, A2, A3, A4)
Input: six user-definable minimum ratio thresholds (R1, R2, R3, R4, R5, R6)
Input: a minimum read depth threshold (n)

if there are two or more length variants present then
    if depth (A1) + depth (A2) $\geq$ n then
        if A1 < A2 then
            if depth (A2)/depth (A1) $\geq$ R1 then
                if there is A3 or A4 and depth (A3 or A4)/depth (A2) $\geq$ R2 then
                    if A3-A2 is equal to $l_{RU}$ and depth (A3)/depth (A2) $\geq$ R3 then
                        Genotype is “A1 A3”
                    else if A4-A2 is equal to $l_{RU}$ and depth (A4)/depth (A2) $\geq$ R3 then
                        Genotype is “A1 A4”
                    else
                        Genotype is “unscored unscored”
                else
                    Genotype is “A1 A2”
            else
                Genotype is “A1 A1”
        else (A1>A2)
            if A1-A2 $\geq$ $l_{RU}$ and depth (A2)/depth (A1) $\geq$ R4 or A1-A2 is equal to 2 and depth (A2)/depth (A1) $\geq$ R5 (this means that A1 A2 are real alleles) then
                if there is A3 and depth (A3)/depth (A1) $\geq$ R6 then
                    Genotype is “unscored unscored”
                else
                    Genotype is “A2 A1”
            else if there is A3 and depth (A3)/depth (A1) $\geq$ R6 then
                if there is A4 and A4-A3 is equal to $l_{RU}$ and depth (A4)/depth (A3) $\geq$ R3 then
                    Genotype is “A1 A4”
                else
                    Genotype is “A1 A3”
            else
                Genotype is “A1 A1”
    else
        Genotype is “0 0”
else if there is only one length variant present then
    if depth (A1) $\geq$ n then
        Genotype is “A1 A1”
    else
        Genotype is “0 0”
else (there is no length variant present, which means that this locus was not identified in this individual)
    Genotype is “X X”
```

## 2.5 Visual Representation Of Genotype Calls

The third major function of MEGASAT enables visual review of the sequence length distributions for each locus and each individual. This function allows users to quickly verify the accuracy of automatically scored genotypes and hence offers the opportunity for reviewers to correct the allele calls in the text data file generated by MEGASAT. Although validation trials of MEGASAT showed a high accuracy of allele calls, user review may reveal cases where a locus signal was misinterpreted. To accomplish this, histograms of length and depth count information are displayed using the R "ggplot2" package [124]. The plots are a graphical representation of the allele calls MEGASAT has made, and are an important tool for quickly reviewing the predicted genotypes. These plots are presented in PDF format.

These histogram plots are colour coded for easy review (as shown in Figure 7). Histogram bars are either grey, pink or blue based on the depth sum of the first two abundant length variants per locus. Grey indicates a sample below the minimum depth threshold which is marked with a "0 0" genotype in the data file. Blue histograms indicate a high depth, and pink bars indicate that the depth is just marginally (in the range between minimum threshold "*n*" and *n*+10) above the minimum threshold. Predicted alleles are plotted in dark blue bars and length artifacts are coloured in light blue bars. These color-coded histograms allow the reviewers to scan quickly over the plot files to see if MEGASAT has scored alleles correctly. The interface allows the reviewer to correct the genotype calls in the text data file.

MEGASAT also generates text files that contain all the discarded sequences (sequences with FP present but otherwise fail to meet MEGASAT requirements), which allows reviewers to check if the customized scoring parameters are set correctly and MEGASAT is not filtering out reads that contain true alleles. These discard files are important early in a project when one is still characterising the loci.

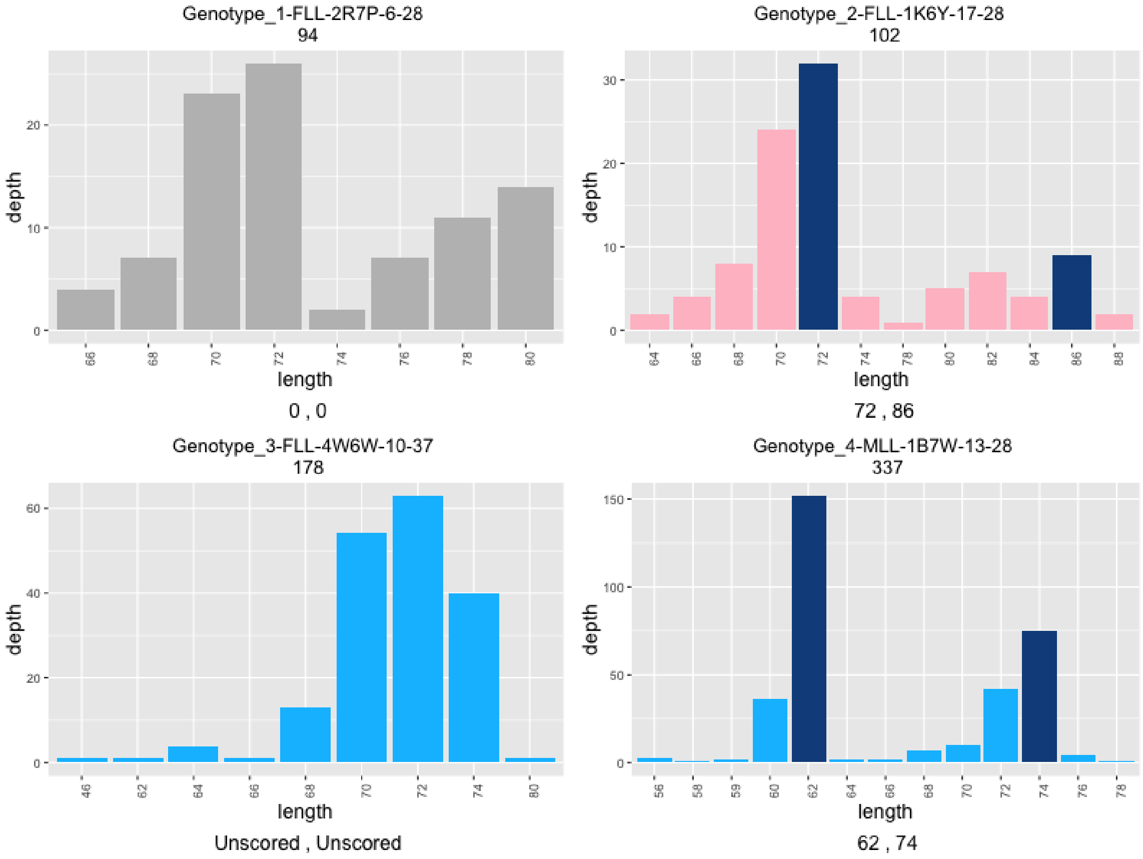


Figure 7 Example of MEGASAT histograms that show the frequencies of sequence length variants per microsatellite locus per individual. Sample IDs title each plot, followed by the total depth. Genotypes are listed under the x-axis. a) Grey bars indicate samples below the minimum depth threshold, with no alleles called (genotype is represented as “0 0”). b) Pink indicates that the depth is between the minimum read depth *n* and *n*+10, dark blue bars indicate allele calls (72/86) c) Blue bars indicate sufficient read depth, but with no alleles called (genotype is represented as “unscored unscored” due to complicated length artifact pattern). d) Dark blue at length 62 and 74 indicate real alleles and a heterozygous genotype.

# CHAPTER 3 APPLICATION OF MEGASAT TO GUPPY MICROSATELLITE DATA

In this chapter, we first describe the source of our validation data set. Then we demonstrate MEGASAT performance via a combination of repeated genotyping of independently extracted and amplified duplicate samples, and examine known pedigrees of guppies to identify genotyping errors in the validation dataset. Further, we compare the genotyping performance of MEGASAT with MicNeSs [22], which is the only other program that can call microsatellite genotypes from NGS reads, and demonstrate MEGASAT's improved performance relative to MicNeSs. The gold standard on all the genotype calls is the manual validation by an experienced researcher that was performed based on the histograms of sequence length variations produced by MEGASAT. The generated text files containing all the discarded sequences were used as a tool to ensure the histograms can correctly represent the length variants and corresponding depth counts. Although an experienced human curator can reliably identify the majority of genotypes, in some extreme cases when stutter artifacts present rare patterns, those rare genotyping errors may not be correctly assigned by MEGASAT or the human curator.

## 3.1 VALIDATION DATA SET

In order to demonstrate the application of MEGASAT for microsatellite genotyping, we used 43 microsatellite loci obtained from the guppy (*Poecilia reticulata*) genome. The majority of the 43 loci are di- and trinucleotide microsatellite loci. Di- and trinucleotide microsatellites tend to generate more stutter artifacts than longer repeats [16, 46], therefore high accuracy on these 'difficult' microsatellite loci may generalize well to microsatellites with longer repeat units. The abbreviated workflow of primer design and DNA sequencing that was performed to generate the guppy dataset is illustrated in Figure 8.

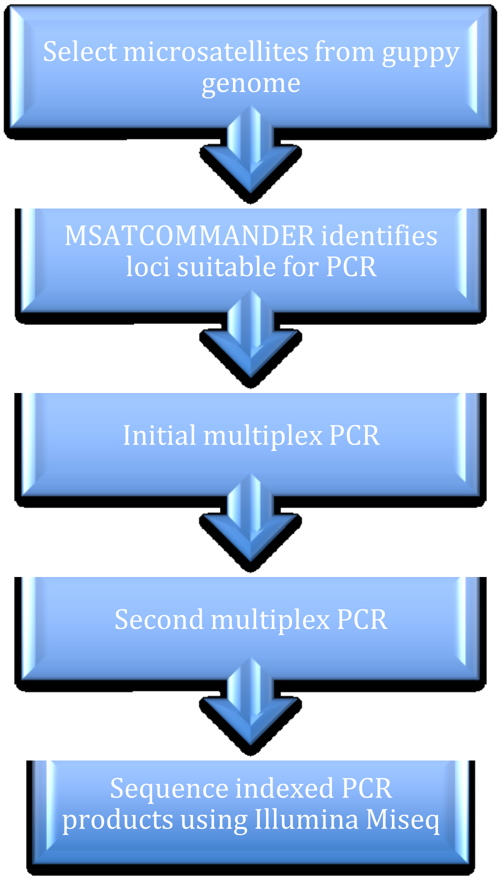

Select microsatellites from guppy genome
MSATCOMMANDER identifies loci suitable for PCR
Initial multiplex PCR
Second multiplex PCR
Sequence indexed PCR products using Illumina Miseq

Figure 8 Abbreviated workflow of guppy dataset generation. Microsatellites containing sequences were obtained from the guppy genome and put into MSATCOMMANDER [59] to choose loci suitable for PCR. Two multiplex PCRs were performed on the selected microsatellite loci and then the generated PCR products were sequenced using Illumina Miseq.

### 3.1.1 Microsatellite Selection

An initial 2659 di- and tri-nucleotide microsatellite-containing sequences were retrieved from a guppy genome project (e.g. [115, 116]) database, targeting sequences with a minimum of 8 microsatellite repeats and approximately 150 bases of sequence on either side of the microsatellite array. The initial microsatellite-containing sequences ranged in length from 258-359 bases. MSATCOMMANDER [59], a program to automatically detect microsatellite arrays and design locus-specific primers, identified 5818 uninterrupted microsatellite arrays within these sequences, 82.2% dinucleotide and 17.8% trinucleotide. Constraining the number of repeat units to >6 and product size to 60-150 bases, MSATCOMMANDER identified 2915 loci suitable for PCR. A total of 468 loci with >7 repeat units were selected for further analysis. Loci with limitations such as low information content, high error rates, evidence of nulls or inability to multiplex well were discarded; a final set of 43 loci was selected for long-term data collection.

### 3.1.2 Laboratory Method

Details of the laboratory workup of our guppy dataset, including PCR, attachment of Illumina adapters and indices and sequencing methods are provided in Appendix A.2. The following is an abbreviated outline. Microsatellite loci (n=43) were amplified in two multiplex PCRs per individual DNA sample. The amplicons were diluted and used as template for a second PCR which added Illumina annealing adapter sequences, a 6 base-pair index (barcode) and the Illumina sequencing primers. Purified indexed PCR products were sequenced in one direction using Illumina MiSeq v3 reagents, which give sequence read lengths of 150 bases. Miseq Reporter software [132] demultiplexed the dual-indices to create separate FASTQ files for each individual.

### 3.1.3 MEGASAT Runs

After the collection of validation data, we customized the parameters that need to be set in MEGASAT. All the primer and flanking information of the 43 microsatellite loci were stored in the input text file. The number of mismatches (*m*) was defined as 2 and the minimum depth threshold *n* was set to 50, values which were found to generate the most accurate genotyping outputs on our guppy data according to the manually curated calls that served as our gold standard. The integer variable "*l*" we specify in the section 2.3 was set to 4. The primer lengths of all of our 43 microsatellite loci are around 20 base pairs. Four is a suitable value that ensures at least four bases of incomplete RP are present in the read. All of the analyses were performed on a single CPU core.

## 3.2 Results And Discussion

### 3.2.1 Frequency Of Genotype Assignments

In order to present some basic information associated with our NGS data and demonstrate the performance of MEGASAT genotyping, we examined the effects of properties that can impact the genotype call rate (referred as the proportion of called informative genotypes as a proportion of the total number of loci). The analysis of the percentage of the targeted genotypes variations among individuals was performed on 2048 guppy individuals (FASTQ files) by counting the number of informative genotypes inferred by MEGASAT (excluding those genotypes scored as “0 0” due to low sequence depth and genotypes not scored because of the identification of more than two alleles) over the 43 loci (Figure 9a, b). The genotyping success rate varied among individuals (Figure 9a). Approximately 86% of individuals were genotyped at a high percentage (over 90% of all loci for that individual were assigned an informative genotype), and ~70% of individuals were genotyped at all 43 loci. The average percentage of collected genotypes was 92.15% among individuals. However, 6% of individuals had genotype call rates <50%, which may be due to the failure of PCR amplification at some individuals or the DNA degradation of some individual samples. Since trimmed sequences are the most important factor that contributes to the generation of genotypes [40], the percentage of trimmed sequences was calculated based on the ratio between the counts of trimmed sequences and total counts of raw reads for each individual sample. The average percentage of trimmed sequences was 64.82% among all the samples. Figure 9b shows the percentage of informative genotypes versus the number of individual trimmed reads. It is obvious that the percentage of informative genotypes collected increases with the trimmed sequence counts, with genotype calling reaching approximately 100% with 5000 trimmed reads.

The analysis of trimmed read distributions among all the 43 loci was performed by counting the average percentage of trimmed reads at each locus for all 2048 individuals (Figure 10). In a random recovery model, each locus would be expected to account for 2.33% (=100%/43) of all trimmed reads, but we observed high variability, with values for individual loci ranging from 1.06% to 4.59%. Around half of the loci (21 loci) have the percentage of trimmed reads larger than the expected percentage (2.33%).

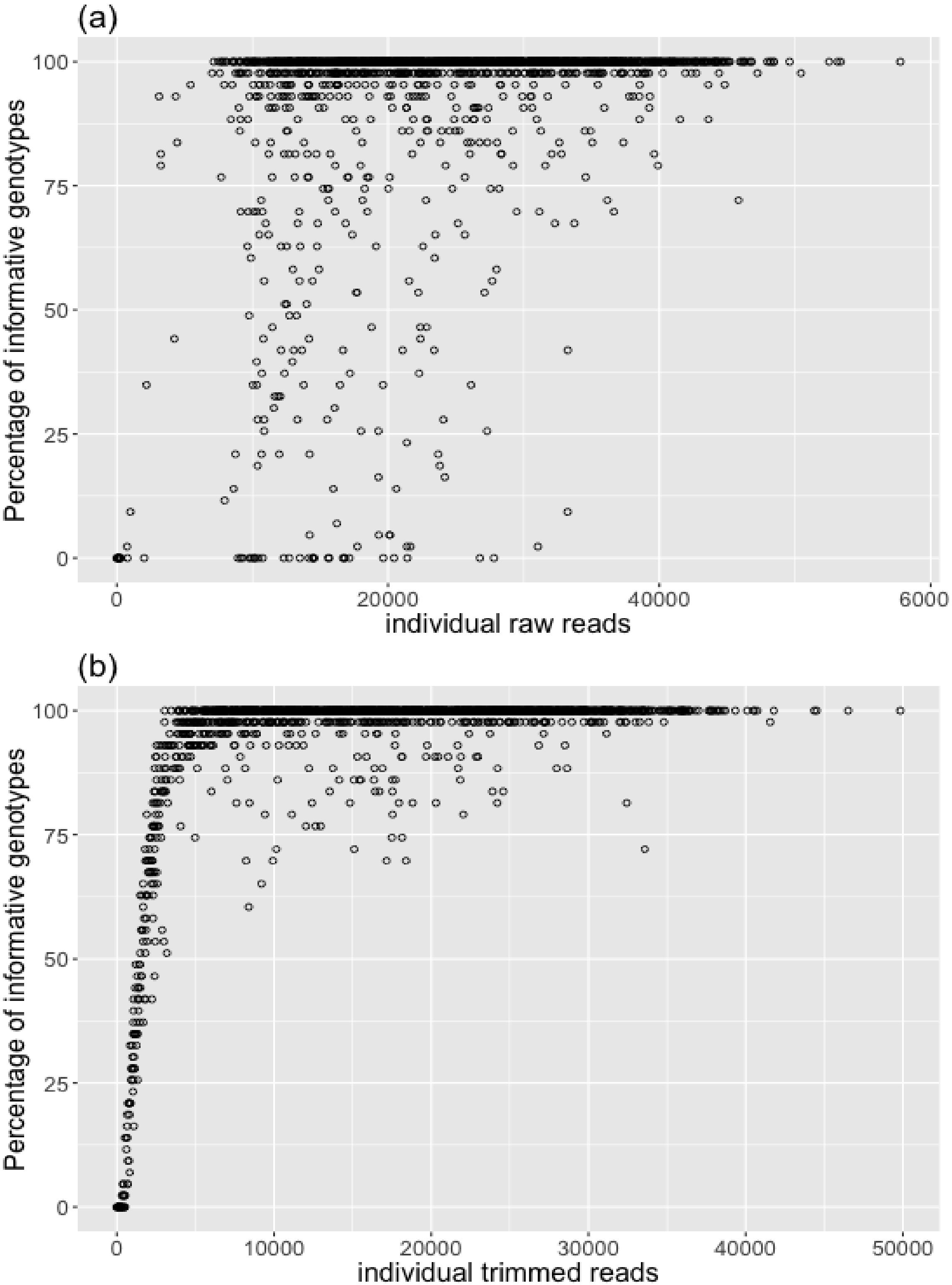


Figure 9 (a) The percentage of informative genotypes collected vs. the number of individual raw reads. (b) The percentage of informative genotypes vs. the number of individual trimmed reads.

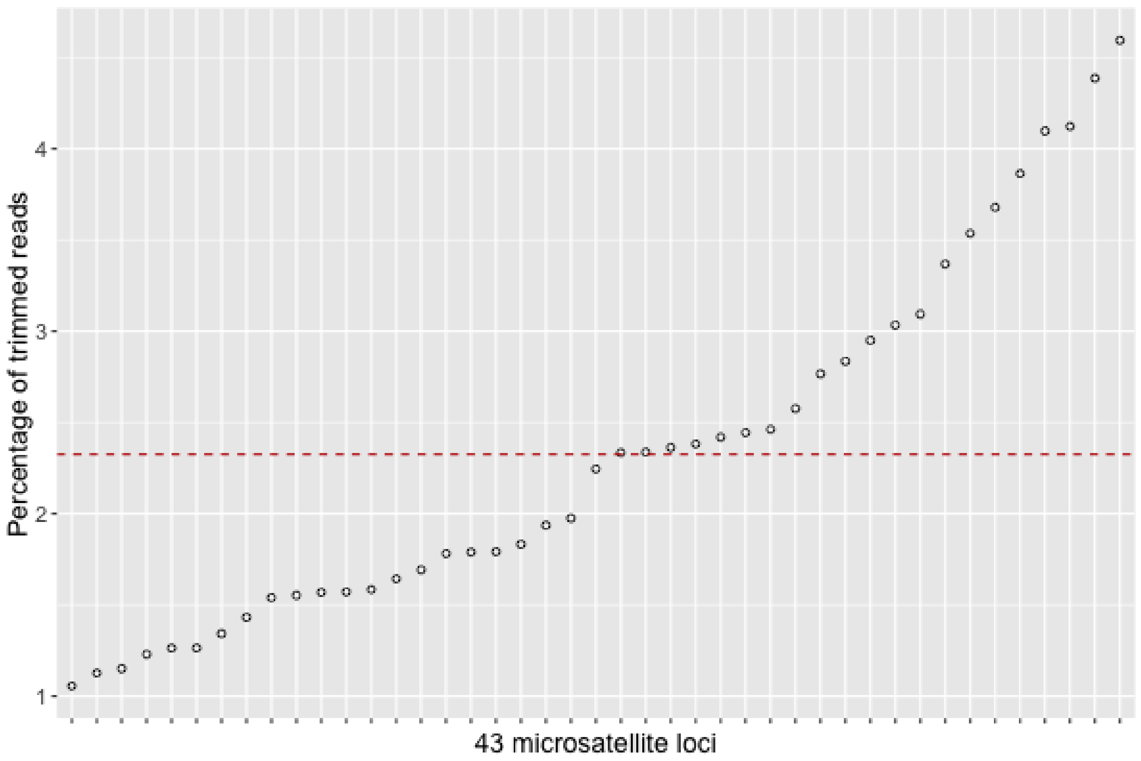


Figure 10 The average percentage of trimmed reads among the 43 loci for all the 2048 individual samples. The percentage of trimmed reads is the number of trimmed reads at each locus over the total trimmed read counts of each individual. The dashed horizontal line represents the expected percentage of trimmed reads (2.33%) for each locus.

Sequence mismatch tolerance allows for the retention of reads with a small number of sequencing errors or mutations within the microsatellite locus. In some cases, if large amounts of microsatellite-containing sequences are discarded because of sequencing errors, it will lead to some amplification artifacts being called as real alleles; in extreme cases, it will result in no alleles being called due to the low amounts of trimmed reads. As an illustration, we chose the 'difficult' microsatellite locus BF-272 that presents more substitution errors to assess the impact of error tolerance. Figure 11 shows the number of retained reads with a mismatch tolerance *m* of 0 (i.e., exact matches required) and 2 for a small set of 23 individuals) for the locus BF-272. Error tolerance yielded a dramatic increase in the number of matched reads for most individuals, and in some cases there were no exact matched reads at all. We also varied the allowed number of mismatches on a sample data set for all 43 loci with $m$ = 0, 1, 2, 3, 4, 5, 6 (see Table 2). Values of $m$ = 3 or 4 gave similar results as $m$ = 2, but larger variations were seen for $m$ <= 1 and $m$ >= 5. These results demonstrate the number of mismatches can affect the accuracy of automatically scored genotypes.

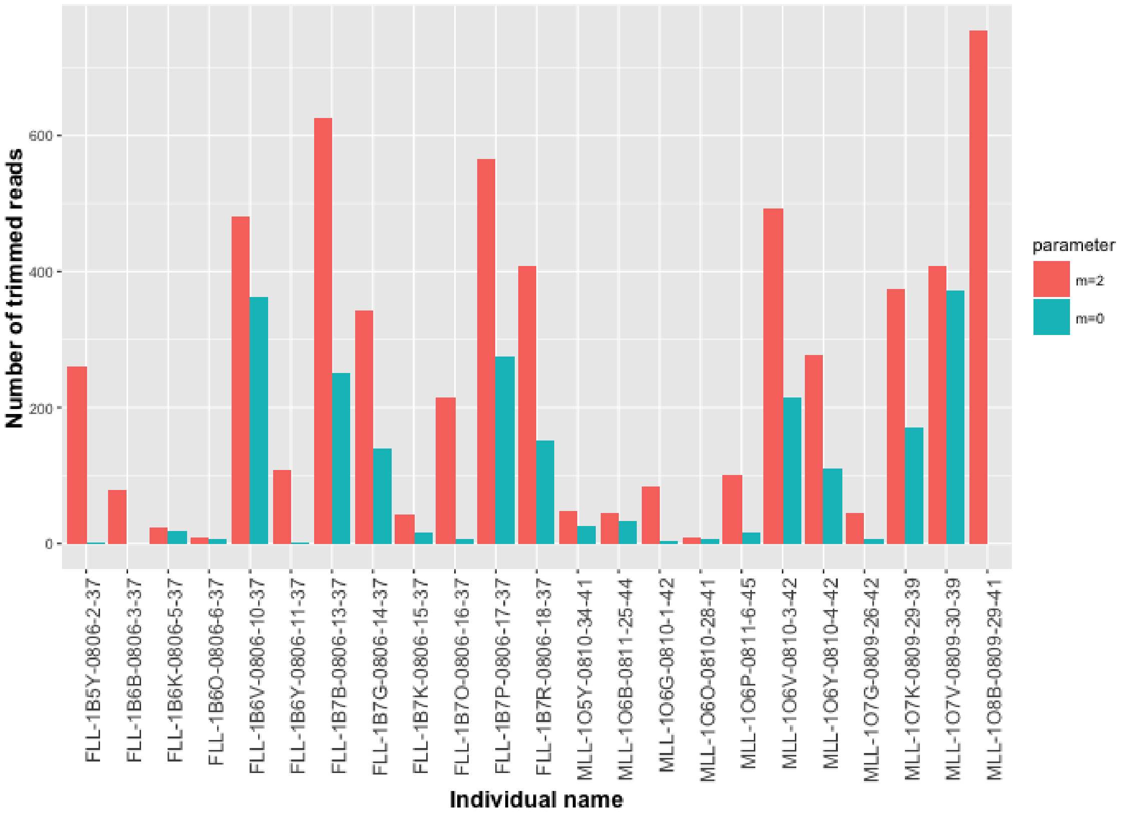


Figure 11 Number of trimmed sequences identified by MEGASAT with and without an error tolerance for locus BF-272 in selected subset of 23 individuals.

Table 2 Comparison of MEGASAT runs with different $m$ values and the run with $m$ set to 2. The second column presents the number of alleles differing from the called alleles with $m$=2. The third column shows the percentage of alleles that were called differently for a given $m$ relative to $m$ = 2, divided by the total number of called alleles.

| **Number of** | **Number of allele** | **Percentage of differences** |
|---|---|---|

| **mismatches (*m*)** | **differences** | **(%)** |
|---|---|---|
| 0 (exact mismatches) | 186 | 5.54% |
| 1 | 92 | 2.74% |
| 3 | 10 | 0.298% |
| 4 | 20 | 0.595% |
| 5 | 190 | 5.65% |
| 6 | 645 | 19.20% |

### 3.2.2 Estimation Of Genotyping Error Rates

Genotyping error rate is the proportion of wrongly called genotypes in the whole set of called genotypes [62]. Even moderate genotyping error rates can cause severe problems in subsequent analyses: for example, genotyping errors could lead to misassignment of kinship and parentage relationships or reduced likelihood of detecting linkage in genetic linkage analysis [60, 62]. Several methods can be used to estimate genotyping error rates, including repeat genotyping, comparison of error-prone genotypes with high-quality reference samples, calculating concordance on resampled individuals, and pedigree-based approaches [63, 64]. The repeat-genotyping method calculates mismatches between genotypes called from duplicately sampled individuals. The pedigree-based method calculates the genotyping errors by identifying the mismatches between scored alleles with known parent-offspring pairs based on some rules of genetic inheritance of alleles [61, 65]. Known parent-offspring pairs should share an allele at one or more loci; any genotypes that violate rules of genetic inheritance will be detected as genotyping errors.

In our experiment, we used the repeat-genotyping and pedigree-based methods to evaluate the accuracy of MEGASAT-scored microsatellite genotypes. In repeat-genotyping approach, 37 individuals were randomly resampled for tissue (scales). DNA extractions, PCR, sequencing and genotyping using MEGASAT were carried out independently, except that some repeat-genotyped individuals were sequenced in the same sequencing run as the original samples. In the pedigree-based approach, 71 guppies from known, lab-reared crosses were genotyped, and parent-parent-offspring triads were examined for genotypes that would violate the rules of inheritance (i.e., alleles present in an offspring that are absent from both parents). In both approaches, genotyping error was evaluated for MEGASAT-scored genotypes both with and without additional manual editing of genotypes.

Rates of genotyping error for MEGASAT-scored genotypes differed between these two methods among all the 43 loci (Figure 12). For MEGASAT-scored genotypes, the mean estimated error rate per allele was 0.021 for the pedigree-based method. Most of the genotyping errors detected using this method were concentrated within a few loci: three loci had error rates exceeding 0.1 (0.109-0.129), whereas 16 loci had no detected errors (error rate < 0.007), and 10 additional loci had a single error (error rate ≈ 0.007).

Estimates of genotyping error obtained with the repeat-genotyping method were lower (mean genotyping error = 0.012). This is because for pedigree-based method, more individuals were compared and a wider range of errors such as allelic dropout, null alleles and mis-genotyped alleles could be identifiable with the pedigree information [46, 48]. However, the repeat genotyping method cannot detect those genotyping errors when errors exist in both duplicates. For the repeat genotyping approach, the three most error-prone loci had estimated genotyping error rates of 0.040-0.050 (around half of the estimated error rates on the basis of pedigree). Among the other loci, 18 loci had no detected genotyping errors, and eight loci had a single error (error rate ≈ 0.007).

Using the histograms of sequence length variation produced for each locus genotype, an expert curator performed manual updates which reduced the genotyping error rates, particularly for those few loci that had the highest error rates in the automated genotype calls. Manual editing reduced the mean error rate from 0.021 to 0.010 in the pedigree-based estimates, and from 0.012 to 0.007 in the repeat genotyping-based estimates. In the pedigree-based estimates, manual editing substantially reduced genotyping error at the eight most error-prone loci; mean genotyping error rates for these eight loci were 0.087

and 0.023 before and after manual editing, respectively. By contrast, manual editing produced no gains in accuracy for 28 loci, either because no errors were detectable in the automatically scored loci, or because the rare errors that did occur were scored the same way by a person and by MEGASAT. Results were similar with repeat genotyping-based error estimates, except that overall error rates, and the gains realized from manual editing, were smaller. Manual editing reduced the mean error rate from 0.044 to 0.018 for the three most error-prone loci in this analysis, but produced little or no gain in accuracy for 35 loci, for the same reasons as before. These results suggest some important considerations for microsatellite genotyping using MEGASAT. Generally, it took around 12 hours for MEGASAT to generate all the genotyping output on 1024 individuals for 43 loci, or approximately one genotype call per second. Manual checking of MEGASAT results took approximately six hours.

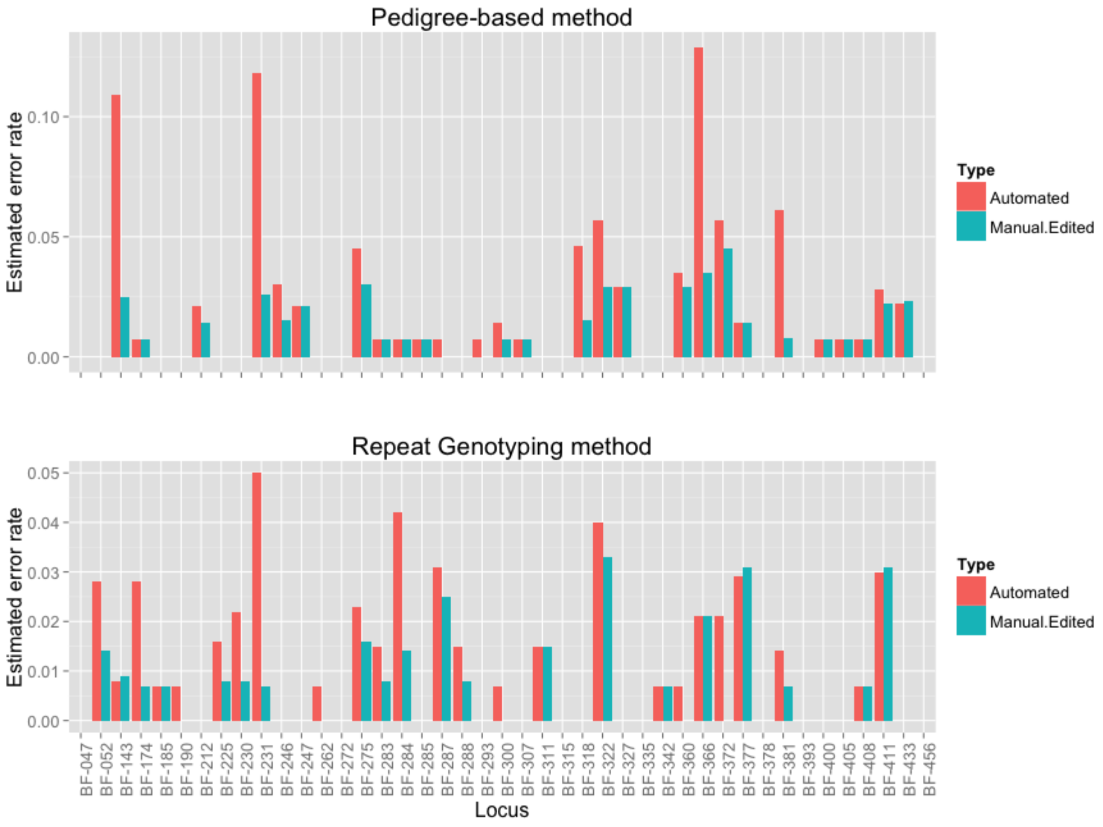


Figure 12 Estimated genotyping error rates of automated genotype calls inferred by MEGASAT with (blue bars) and without (red bars) manual editing for all the 43 microsatellite loci using repeat genotyping method and pedigree-based

method. In the pedigree-based method, 16 loci had no detected genotyping errors. 18 loci had no detected errors in the repeat genotyping approach.

### 3.2.3 Pedigree Construction

Genotype information is often used to infer pedigrees for a set of individuals for which direct information about relationships is not available. This approach is particularly valuable when applied to individuals sampled from different locations or time points. In order to further explore the contributions of MEGASAT to pedigree analysis, more than 600 guppies captured from 12 cohorts between 03/2008 to 03/2009 on the island of Trinidad were automatically genotyped at 43 microsatellite loci using MEGASAT. In order to better present the cohort information in pedigrees, we used integers ranging from 0 to 11 to represent the cohorts captured from 03/2008 to 03/2009. Then the program FRANz [66] was used to conduct a pedigree analysis based on the MEGASAT-scored genotypes and genotypes updated via manual curation. FRANz is a program that uses an error model with Markov Chain Monte Carlo (MCMC) sampling to estimate the statistical confidence of constructing parent-offspring relationships while allows the incorporation of prior information such as age and sex of individuals [66].

In our guppy samples, we used the already known sex and age (defined based on the month that those guppies were captured) as prior information for FRANz. Age was used to order the entire guppy individuals into successive generations. The parameter specifying the number of candidate fathers in FRANz was set to be the number of known males in our data. FRANz generates several output files, including a comma-separated file that lists the likeliest parents of each individual and the corresponding statistical score and posterior probability of each parentage in the pedigree. The Pedigree Viewer [125] software was used to read this file and visualize the FRANz output. Figure 13 is the constructed pedigree made directly with the automatically scored genotypes and Figure 14 shows the pedigree based on the genotypes assigned by MEGASAT with manual editing.

From both pedigrees, it is obvious to find that some individuals in newer cohorts mated with individuals in earlier cohorts. The pedigree made with the manually edited genotypes shows more parentage relationships between new generations especially between the second generation and the third generation. Using the statistical scores calculated from manually edited and automated genotypes, we found differences between these two pedigrees where around 73% of the inferred parentage relationships were larger than those from the automated genotypes (as shown in Figure 15). The Pearson's correlation coefficient on the statistical scores was 0.968 ($p$ value < 2.2e-16). From the comparison of these two pedigrees, it is obvious that manual editing could affect the inferred parentages in the built pedigrees and might give more statistical confidence for the inferred parentages.

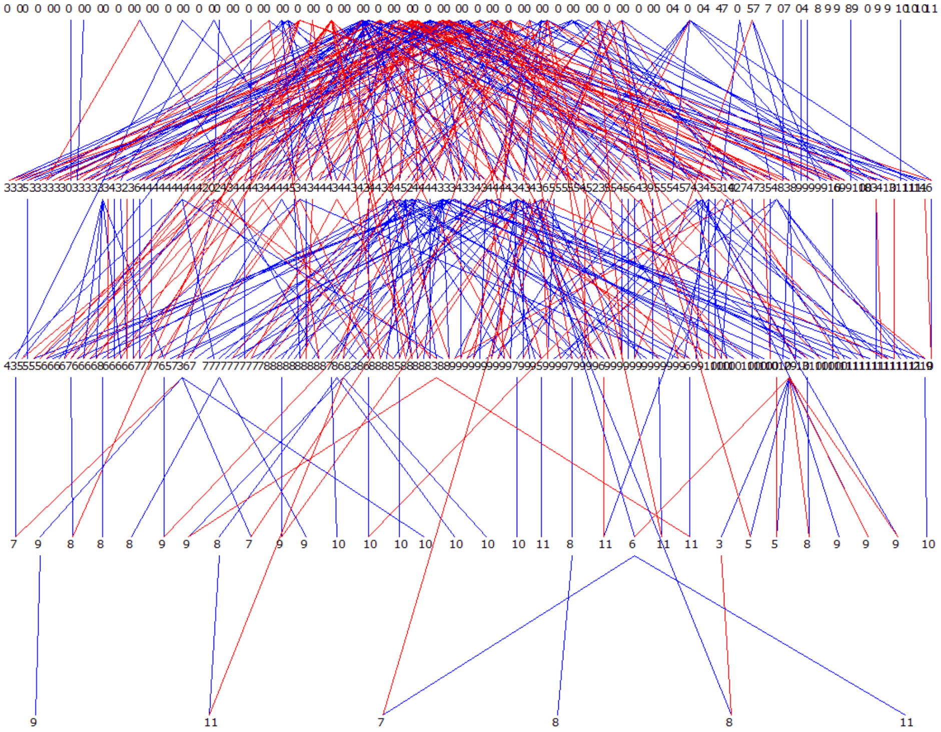


Figure 13 Pedigree built on the MEGASAT automatically scored genotypes. The founder population is arrayed along the first row and new generations are listed on each following row. Numbers represent the cohort number in

which offspring was captured. Blue lines run from paternal parent to offspring and pink lines run from maternal parent to offspring.

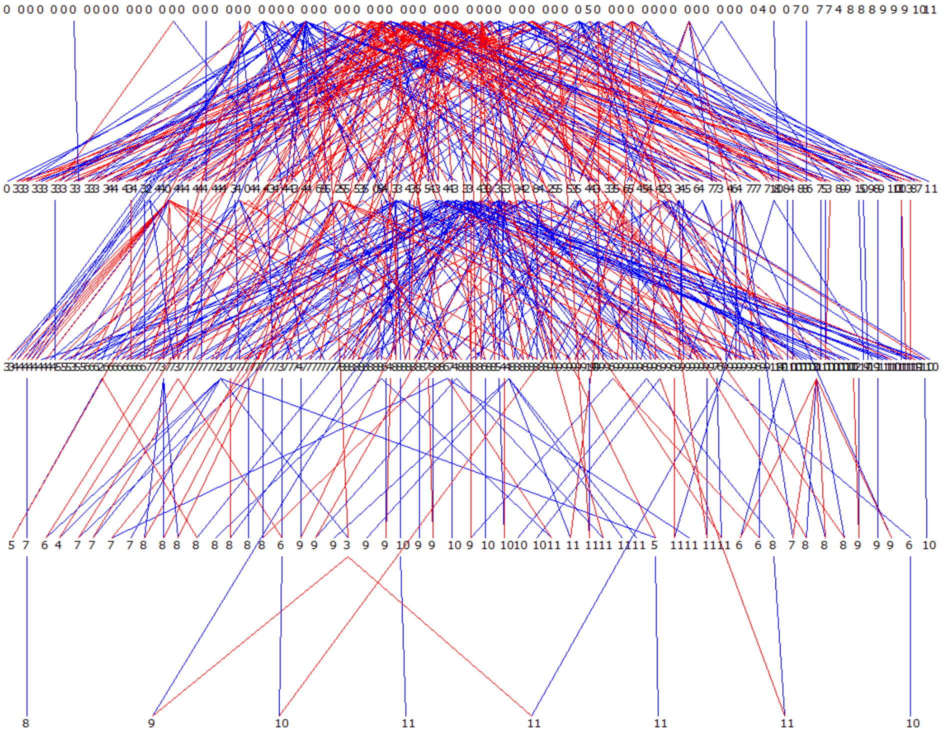


Figure 14 Pedigree built on the MEGASAT-scored genotypes with manual editing.

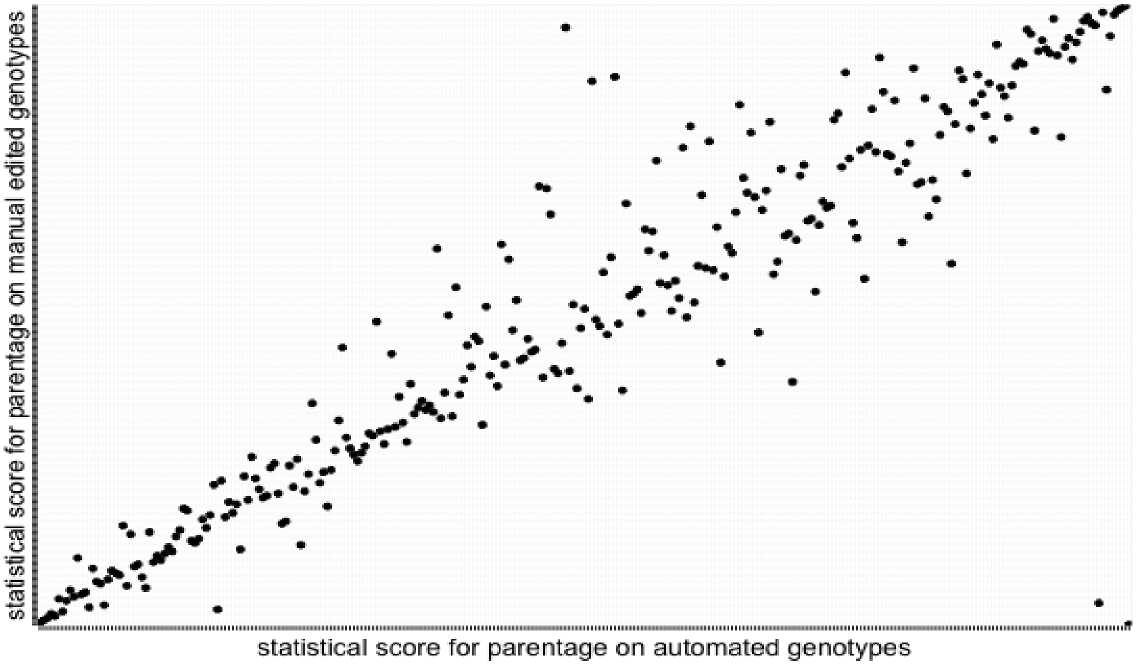


Figure 15 Comparison of statistical score for the same inferred parentage on automated genotypes and genotypes with subsequent manual curation.

### 3.2.4 Comparison With MicNeSs Results

To the extent of our knowledge, MicNeSs is the only other program that infers microsatellite genotypes from NGS data. Suez et al. [22] employed an algorithm that first extracts the microsatellite-containing sequence with the largest number of repeat motifs among all reads of all individuals. Then sequences whose repeat motifs differ from the extracted repeat motif will be referred to as different microsatellites. The following step is to build the observed microsatellite length distribution for each individual using the number of repeat motifs and number of substitutions. After the observed distributions have been built, the next step is to assign a theoretical parametric distribution to each allele in an individual. Then an individual's genotype is inferred by using the optimized parameters of the theoretical distributions via minimizing the squared difference between observed length distribution and the theoretical parametric distributions.

The algorithm implemented by MEGASAT requires the reference data provided by the users to allow the sizing of microsatellite repeats, and employs some ratio functions to disentangle real alleles from amplification and sequence-driven artifacts. By contrast, MicNeSs uses parametric models to simulate the mode of genotypes and then infer genotypes via optimizing models based on the prior information provided by the data. This method does not need the user to define any parameters; instead, the software can automatically find the optimal parameters from data. And if the assumed model is correct for the data, it could achieve highly accurate results with enough data. However, this method suffers from some disadvantages. If the parametric model cannot correctly simulate the mode of data, it could induce bias into the inferred results. Furthermore, parametric modeling is always accompanied with a high computational cost. Several additional important algorithm and functionality differences between MEGASAT and MicNeSs are listed below:

- MicNeSs can deal with sequencing errors, but it cannot overcome artifacts that may occur during the PCR amplification, which could result in some amplification artifacts being miscalled as real alleles (see below).
- MEGASAT can run on files that contain NGS reads encompassing multiple loci. However, the input files of MicNeSs need to be locus-specific files that contain reads for only one microsatellite locus of interest, which means that MicNeSs cannot directly perform on sequencing data encompassing multiple loci. Pre-processing for these complicated data is needed but the code for pre-processing is not available.
- A user graphical interface is available for MEGASAT but MicNeSs can only be run through a command line.
- MEGASAT outputs discarded and trimmed sequences to files for each individual and each locus. This function allows users to ensure MEGASAT is not overly rigorous in throwing out some microsatellite-containing sequence. However, MicNeSs does not offer the opportunity to review sequences that are filtered out by it.

- MEGASAT visualizes the data used to infer genotypes by creating depth vs. size histogram plots, which allows verification and manual editing of automated genotypes. While in MicNeSs, only genotypes are generated so no function is offered to verify the accuracy of called genotypes.

In order to compare the performance of MEGASAT to MicNeSs, we ran MicNeSs on a guppy data set comprising 172 individuals. Then we compared the genotypes scored by MicNeSs with the genotypes inferred by MEGASAT (as shown in Figure 16 (a)). The accuracy of genotypes scored by MEGASAT on this data set was manually validated by a researcher with rich experience in assigning genotypes based on length distributions. The percentage of difference is the number of alleles scored by MicNeSs that differ from the MEGASAT output over the total number of scored alleles (172*43*2). The alleles called by MicNeSs were maximally different from MEGASAT at two microsatellite loci (pink bars in Figure 16 (a)). This difference may be a consequence of incorrect pre-processing of multiplexed FASTQ files using the scripts offered by the developer of MicNeSs, which causes a majority of microsatellite-containing sequences to be discarded. The mean percentage of differences for the 41 remaining loci is 13.17%. Only 18 loci had percentage of differences smaller than 10%. Eight loci showed 20% difference from MEGASAT genotyping output. Figure 17 shows two examples, where MEGASAT correctly filtered out the amplification artifacts and called the real alleles. However, MicNeSs called incorrect genotype “36 56” from Figure 3.10 (a) and genotype “58 58” from Figure 17 (b).

The running time is also a good indicator to evaluate the performance of a program. Therefore, we compared the running time of MicNeSs and MEGASAT on the same individual samples. We found that the computing time of MicNeSs among these 43 loci is highly variable (Figure 16 (b)). This is because the number of alleles and associated variance has huge impacts on the running time of MicNeSs [22] and the number of alleles fluctuates among all the 43 loci. As an illustration, it took around 27 minutes for locus BF-047 to get all the genotyping output but for locus BF-322, around 155 minutes were required (Figure 16 (b)). The running time for the two wrongly processed loci (pink bars) was small since no real alleles existed in the input data. But for MEGASAT, allelic variance does not make a big difference for the running time. The running time for each of the 43 loci on the same dataset is just around 2 minutes. However, MicNeSs needs at least 10 minutes even for the locus with smallest number of alleles. The maximum running time for MicNeSs to assign genotypes at a locus on the same dataset with MEGASAT is around 155 minutes, 77.5 times slower than MEGASAT.

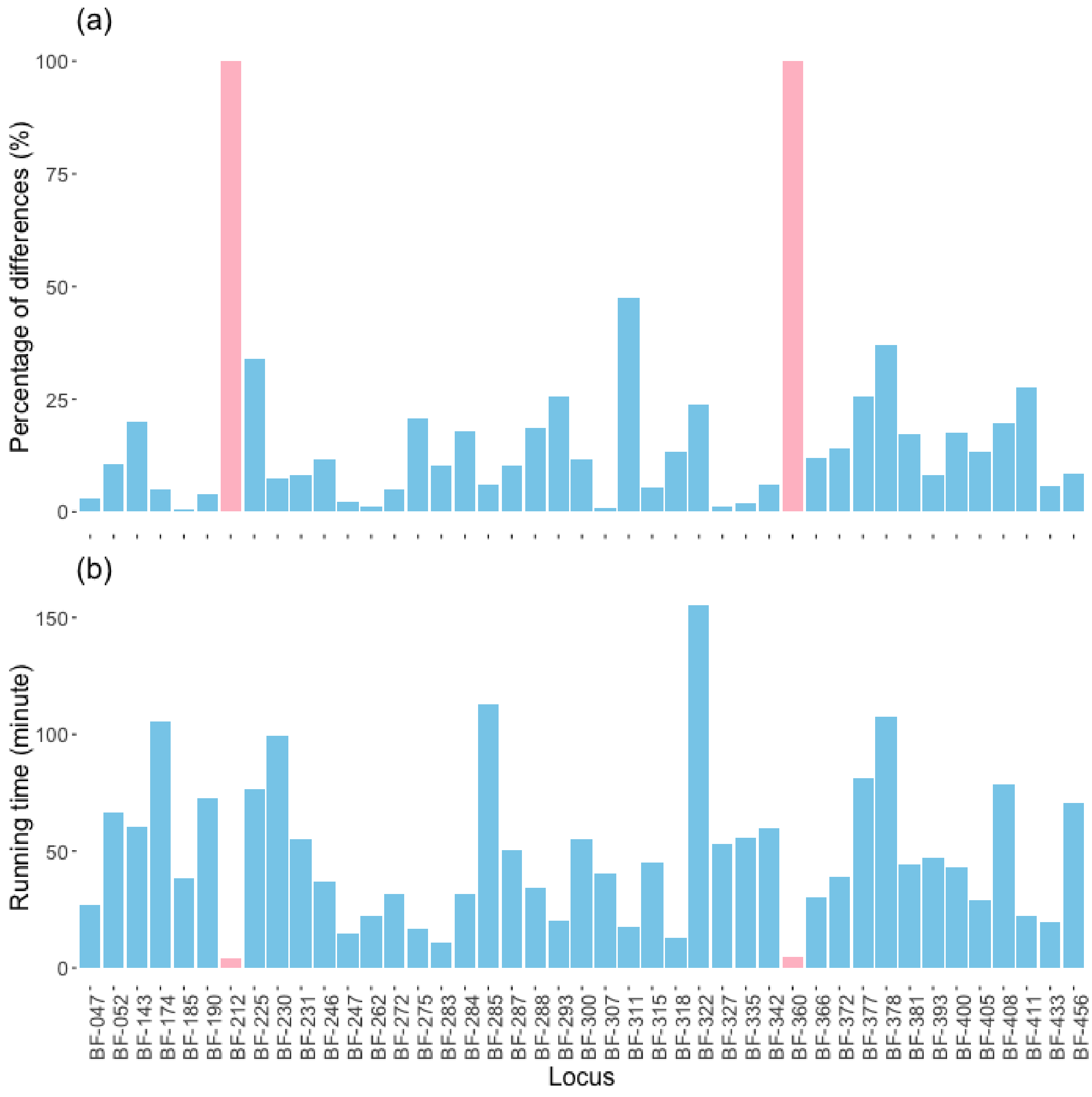


Figure 16 (a) The percentages of genotypes inferred by MicNeSs that differed from those scored by MEGASAT for all the 43 loci. Two loci had percentage of difference equal to 100% (pink bars), which were caused by incorrect pre-processing of input data and have no direct causal correlation with MicNeSs. (b) The running time of MicNeSs on ~ 170 individuals for all the 43 loci. The two wrongly processed loci are colored in pink.

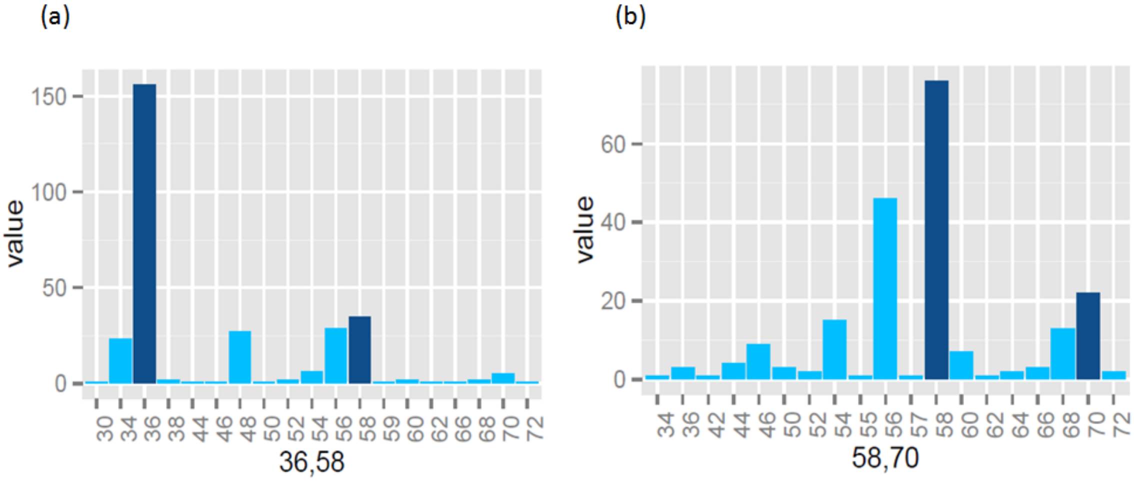


Figure 17 Two examples of MEGASAT histograms that show the correct genotype inference by MEGASAT. However, MicNeSs called the genotype as “36 56” from (a) length distribution. Genotype “58 58” was inferred by MicNeSs from (b) length distribution.

## 3.3 Discussion

Many of the disadvantages associated with microsatellite genotyping stem from its reliance on the capillary electrophoresis and the imperfect inference of genotypes from DNA fragement mobility data. The preparation of samples for each capillary electrophoresis and genotypes inference remain far too expensive and time-consuming. Even though several genotyping softwares such as GeneMapper (Applied Biosystems) or GeneMarker (Softgenetics) have been developed to visualize and filter out some stutter products, most of the high-precision software packages are commercial and therefore the genotyping costs are relatively high. However, MEGASAT addresses these problems by calling microsatellite genotypes from raw sequence data. Furthermore, it provides several contributions to microsatellite genotyping.

First, completely automated genotype prediction is feasible for many loci. In our experiments, automated genotype prediction resulted in mean error rates of 0.003-0.004 for 28 loci (estimated using either method), and no genotyping errors detected for 16-18 loci. Slightly higher mean rates of genotyping error occur with fully automated genotyping of up to 40 loci in our panel. Moreover, our panel of 43 loci was selected for their ease of amplification, high degree of polymorphism, and suitable allele size ranges, but they were not rigorously screened for their tendency to produce easily interpretable genotypes. A clear implication is that further screening of candidate microsatellite loci could have produced more loci that met all desired criteria, including amplification products amenable to highly accurate, fully automated scoring. We enjoyed the luxury of having thousands of candidate microsatellite loci to choose from; however, NGS-based approaches to identifying novel microsatellite loci make the identification of large numbers of microsatellite loci feasible and relatively inexpensive in any species (e.g. [17] [83]).

Second, such fully automated genotype prediction brings great advantages in genotyping throughput (and associated labour costs), low genotyping error rates, and ease of data standardization across experiments and laboratories. A single researcher can obtain data for ~41,000 single-locus genotypes per week with fully automated scoring, and ~44,000 genotypes per week with some manual editing. As noted, genotyping error rates are low, and comparable to those obtained with carefully selected loci using conventional electrophoretic methods in other studies [57] [61] [64]. Since genotypes are based on direct counts of DNA sequence lengths rather than indirect inference from electrophoretic data, data standardization between platforms and laboratories is not a concern.

Third, there will be situations where manual editing is desirable. For example, it may be advantageous to include somewhat more difficult-to-score loci to enable comparisons with older data sets, or comparisons across species, or because particular loci have particular merits, such as being linked to genes or traits of interest. The data visualization feature in MEGASAT enables easy manual checking and editing of genotypes, and our results suggest that rapid manual editing can improve genotyping accuracy at loci that might otherwise be of marginal utility. Conversely, although the default values of variables MEGASAT uses to guide the decision making process for identifying true alleles among amplification or sequencing artifacts work well for a wide variety of di- and trinucleotide microsatellite loci, locus-specific adjustments of some of these user-definable variables may improve the automated scoring accuracy of some problematic loci.

# CHAPTER 4 DETECTING CLIMATE-ASSOCIATED ADAPTATION USING SPATIALLY CONTROLLED RANDOM FORESTS

The main objective of this chapter is to present a novel approach that can identify adaptive loci and also environmental factors associated with those loci, with a focus on a previously decribed sea scallop dataset [137]. This dataset spans a large latitudinal range, and we hypothesized that environment-driven adaptation might play an important role in shaping the genetic structure of sea scallop. However, the large numbers of environmental factors and their multicollinearity complicate our analysis. In order to identify adaptive signals, we proposed a random-forest (RF) approach, which differs from other approaches that are commonly used in environmental association studies. The non-linear RF models used all the environmental factors as predictors to predict individual genotypes at each SNP in order to identify candidate adaptive loci. The RF model could automatically select important environmental features, even those that are highly correlated. Finally, by combining our RF approach with a matrix transformation to control for spatial autocorrelations, we were able to generate classifications with a likely decrease in the false-positive rates of identified associations.

## 4.1 INTRODUCTION

### 4.1.1 Associating Genotype With Environment

Adaptive genetic variants play an important role in the maintenance of species viability in a heterogenous environment. In order to identify loci under selection, a common approach is to associate genetic data (commonly SNPs) with environmental variables; and identify gene variants that show strong correlations with the environment. However, large amounts of genetic data, often corresponding to thousands or tens of thousands of

SNPs, and highly correlated environmental factors complicate the association analysis. Furthermore, any inference method needs to consider non-adaptive effects such as spatial autocorrelation caused by geographical distance. Spatial autocorrelation is quite common in ecological data. Therefore, if the model does not take these effects into consideration, the detected important loci might not be adaptive and have no relationships with selective forces.

However, exploring the relationships between many SNPs and correlated environmental variables is challenging. A common approach is to consider only a subset of features in detail. For example, many studies identify ”outlier” loci that deviate from a null hypothesis of no environmental association using Bayesian statistics or $F_{ST}$-based methods. Ordination techniques such as PCA build metavariables that are linear combinations of correlated environmental variables; these metavariables can be used directly, or replaced by representative environmental variables with which they are highly correlated. In the final step of building an association model, multivariate constrained ordination methods such as canonical correspondence analysis (CCA) or redundancy analysis (RDA) are widely used to reveal the relationships between the reduced set of SNPs and environmental variables. Partial ordination techniques can also be used to control for neutral effects such as geographical effects. Non-outlier-based regression models have been developed to identify adaptive loci from large sets of markers: for example, in some studies linear regressions while controlling for spatial effects using Moran’s eigenvector maps (MEM) have been used to identify adaptive loci and selective forces [76, 119]. Another common used type of model is mixed effect model (e.g. Latent factor mixed model (LFMM) [120]) that uses random effects to control for non-adaptive forces when testing for genotype-environment associations.

Here we apply a supervised learning approach, random forests, to build association models between genotype and environment. In order to uncover the signals of selection, random forest models used all the environmental factors as predictors to classify the genotypes at each SNP independently. SNPs with high prediction performance were selected as potential adaptive SNPs. Then, multivariate analysis was performed to discover how some important environmental factors structure the genetic data at identified adaptive loci. However, several challenges need to be addressed before building the multivariate model. First, how to deal with high multicollinearity in the environmental data is a concern. Second, strong correlations identified by multivariate analysis may suffer from a high false positive rate due to the impacts of spatial effects on the genotype and environmental data. Therefore, adjustments are required to account for the spatial effects in order to build a set of good predictors that are not highly correlated with geographical distance.

### 4.1.2 Random Forest Classifier

Decision trees are the building blocks of random forests. Decision trees consist of terminal nodes, which represent a set of target outcomes, internal nodes, which define decision criteria, and branches that represent decision paths. Decision trees are built by choosing important discriminating variables as internal (decision) nodes and splitting the data set based on the observation values on those variables until the tree growing is finished. A node can be described in terms of its *purity*, which expresses the extent to which the node represents members of a single class, whereas an impure node does not effectively distinguish members of different classes. The choice of important variables for splitting is based on the calculation of node impurity decrement, i.e., to what extent a given variable can distinguish two or more classes. After splitting on each decision node, node impurity is calculated for parent and child nodes. The more the node impurity decreases from splitting, the more important the variable is.

Decision trees are popular due to their easy interpretation, automatic feature selection and their elimination of the need for data normalization. However, decision trees often have higher error rates in comparison with other supervised learning models due to their relative simplicity, and have a high risk of overfitting as the decision criteria can be too specific to the training data. Tree complexity is inversely correlated with the prediction accuracy. Decision trees may have poor performance on test dataset when the built tree is very complex [104, 136].

*Bootstrap aggregation* or *bagging* is a powerful algorithm that was proposed by Breiman in 1994 [105] to reduce the variance of a statistical model [104]. The reduction of variance is achieved by assembling multiple separate models and averaging the predictions on those models. Therefore, bagging can be used in decision trees to solve the overfitting problem.

Consider a training data set *S,* and a predictor *P* (*S,* $x$), which uses *S* to predict the value at input point $x$. In order to aggregate multiple predictors, the training data set *S* is resampled with replacement (i.e., bootstrapped) to generate *N* different training data sets $S^1, S^2, \dots, S^N$. Then we train the predictor *P* on each bootstrapped training data set and calculate

$$P_{\text{avg}}\,(\text{x}) = \frac{1}{N}\sum_{i=1}^{N} P\ (S^i, x) \tag{4.1}$$

This equation is for predicting quantitive outcome. If the outcome is qualitative, then

$$P_{\text{avg}}\,(\text{x}) = majority\ vote\ \{P\ (S^i, x)\}_{i\,=\,1}^{N} \tag{4.2}$$

The majority vote is to assign the most frequently present class to the prediction class of $x$.

Since their introduction in 2001 [103], random forest (RF) classifiers have gained widespread use in many domains of machine learning. RF is an ensemble approach that benefits from growing a large group of decision trees to improve its overall performance. RFs refine the concept of bagged trees by using a minority of predictors on splitting in order to build a number of independent decision trees on bootstrap samples [104, 105]. This refinement substantially reduces the variance of bagged trees. Suppose if most of the predictors are considered as candidates for splitting, the majority of the built bagged trees will choose the most important predictors for splitting, which will result in overfitting since all the bagged trees will be similar [104]. In random forests, decision trees are built using the following steps:

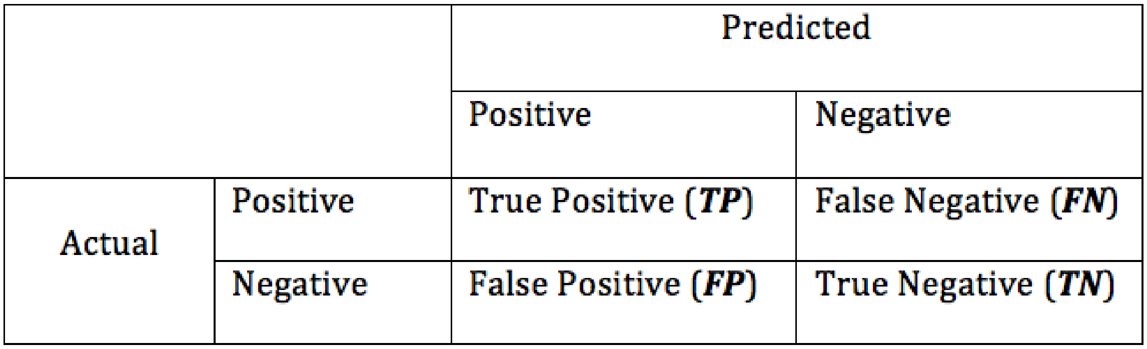
- Randomly select *m* predictors from the full set of *p* predictors.
- Choose the top splitting predictor from *m* predictors.
- Split the nodes
- Repeat the first three steps until the tree growing is finished

The above four steps will be repeated N times (determined by the tunning parameter: number of trees) and the final prediction of random forests will be calculated the same as bagging algorithm.

The two key attributes of RF are the out-of-bag (OOB) error estimation and variable importance measurement. OOB data refers to those observations that do not contribute to the construction of bagged trees. Therefore, the OOB error is a good quantitative measurement for the generalization error of an RF. Another attractive feature is the automatic computation of variable importance. Usually two types of importance are offered: Gini importance (RF classification) and permutation importance. Gini importance calculates the total node impurities decrease that results from splitting on each variable and averages over all trees [135]. Permutation importance counts the average prediction error increase on OOB data after permuting the values of each variable. If the variable is an important predictor for the RF model, the permutation on the values of this variable will produce more prediction errors. Therefore, large permutation importance values indicate important variables. These two types of importance are both critical for feature selection.

### 4.1.3 Evaluation And Verification Method

Many evaluation methods are available for assessing the performance of a machine-learning model. A simple measure of accuracy is the percentage of predictions that are correct. However, in imbalanced data, the accuracy is not a good quantitative measure of model performance. Imbalanced data refers to data whose classes are not present equally, and in some cases the number of instances in a class is much less than others. As an illustration, in a classification task for predicting the existence of a rare disease in a patient, the majority of the training set may be non-diseased individuals. The overall prediction accuracy will be high even though no individuals will be correctly predicted as diseased samples. In order to better illustrate the performance of a model, the confusion matrix (as shown in Figure 18) is used to give more details of correct classification and misclassification for each class.



| | | Predicted | |
|---|---|---|---|
| | | Positive | Negative |
| Actual | Positive | True Positive (***TP***) | False Negative (***FN***) |
| | Negative | False Positive (***FP***) | True Negative (***TN***) |

Figure 18 Confusion matrix for a two-class classification problem.

Another commonly used evaluation measure is the Receiver Operating Characteristic (ROC) curve, which was recognized as a good evaluator for model performance by Metz [106]. A ROC curve (as shown in Figure 19) is a visualization tool that illustrates how the true positive rate (TP/(TP+FN), also referred to as sensitivity) fluctuates with the false positive rate (FP/(TN+FP)). A perfect classifer will have a 100% true positive rate and a 0% rate of false positives, which means that the curve will have a point in the top left corner of the plot. The area under the ROC curve (AUC) is used to assess the prediction performance across a range of decision thresholds. For a two-class classification task, an AUC value of 1.0 represents a perfect model, while a value of 0.5 corresponds to random predictions, as increases in sensitivity are directly correlated with decreases in specificity.

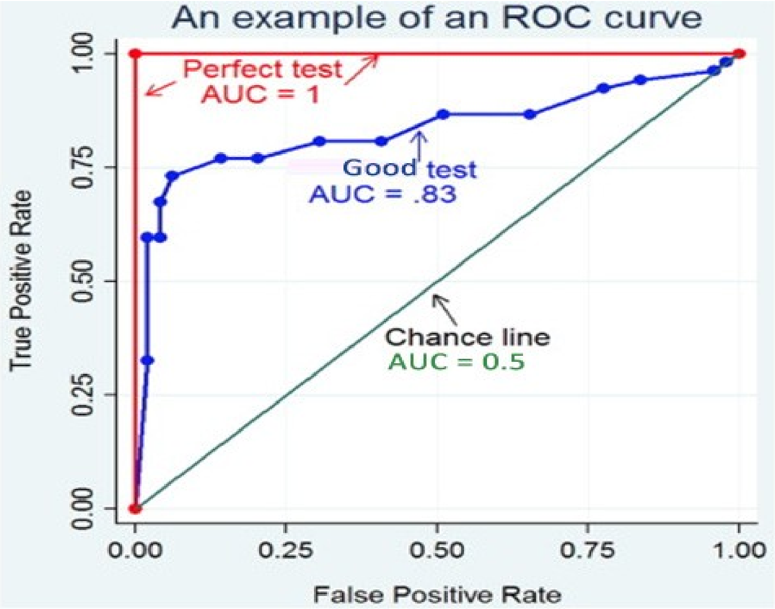


Figure 19 An example of Receiver Operating Characteristic (ROC) curve retrieved from [126]. This example presents three kinds of ROC curves: a perfect ROC curve with AUC values equal to 1, a relatively good ROC curve with AUC values = 0.85 and a chance line with AUC values = 0.5.

In order to evaluate the predictive performance of machine learning model on the test data, the *k*-fold cross validation has been widely used in machine learning evaluation stage. It first divides the whole data set into *k* subsets. Subsequently, $k - 1$ subsets are utilized as the training set to train the model, and the remaining subset is used as the test set for test error estimation. The above process is repeated *k* times to get an averaged test error.

### 4.1.4 Population Of Interest

The Atlantic sea scallop (*Placopecten magellanicus*) is one of the most commercially important marine species in North America. In 2012, the overall economic value of this species was around $113.5 million. This species is common along the shore of the Gulf of St. Lawrence and the Bay of Fundy off Digby, and the coasts of the Mid-Atlantic Bight as far south as North Carolina [84] [85] [87]. Even though several management measures have been made to increase their sustainability to excessive fishing, long-term effective management is still in need owing to the increasing marketing demand of the sea scallop. Furthermore, ongoing environmental changes, such as global warming, would lead to the increment of sea surface temperature, which might shape the future distributions or the geographic ranges of species including scallops. These induced changes would have negative impacts on sea scallop habitats. Therefore, research exploring climate-associated adaptation in sea scallop populations is required to provide critical insights for the understanding of how climatic factors drive distributions of marine species, and subsequently implicate the environmental and management policies.

Many markers such as random amplified polymorphic DNA (RAPD), RFLPs and microsatellites have been developed for population genetics studies in sea scallop [97]. A large panel of SNPs assessed using NGS data can present a complete genomic picture across all regions of the genome, providing the opportunity to identify both neutral genes and adaptive genes that have strong relationships with environmental gradients. This enables the detection of adaptive loci through environmental association analysis. Normally the genotypes at a bi-allelic SNP can only have three classes (homozygous for either of the two alleles, or heterozygous), but large numbers of SNPs provide high-resolution differentiation of individuals. Individuals with similar genotypes, which indicate that they are closely related, can be clustered into a population. Individuals may also have admixed genotypes which indicate similarity to two or more populations. Therefore, SNP genotypes enable the inference of the composition of a population (population structure).

### 4.1.5 Analysis Workflow

Our objective in this work was to pair a machine-learning approach with statistical models to identify adaptive loci and the environmental factors that are associated with these adaptive variants. The analysis proceeded in several steps (Figure 20). First we inferred the number of distinct populations assignments of individuals to each of these populations using the *STRUCTURE* software [100]. We then trained RF classifiers using environmental variables as input and individual SNPs as output, with well-predicted

SNPs considered as candidate adaptive loci. Next, to correct for spatial autocorrelations in the genetic and environmental data, we used statistical models to partition the variance of these data into spatially correlated and uncorrelated components. Another random forest analysis was then performed to select important SNPs whose effects are at least partially independent of spatial relationships.

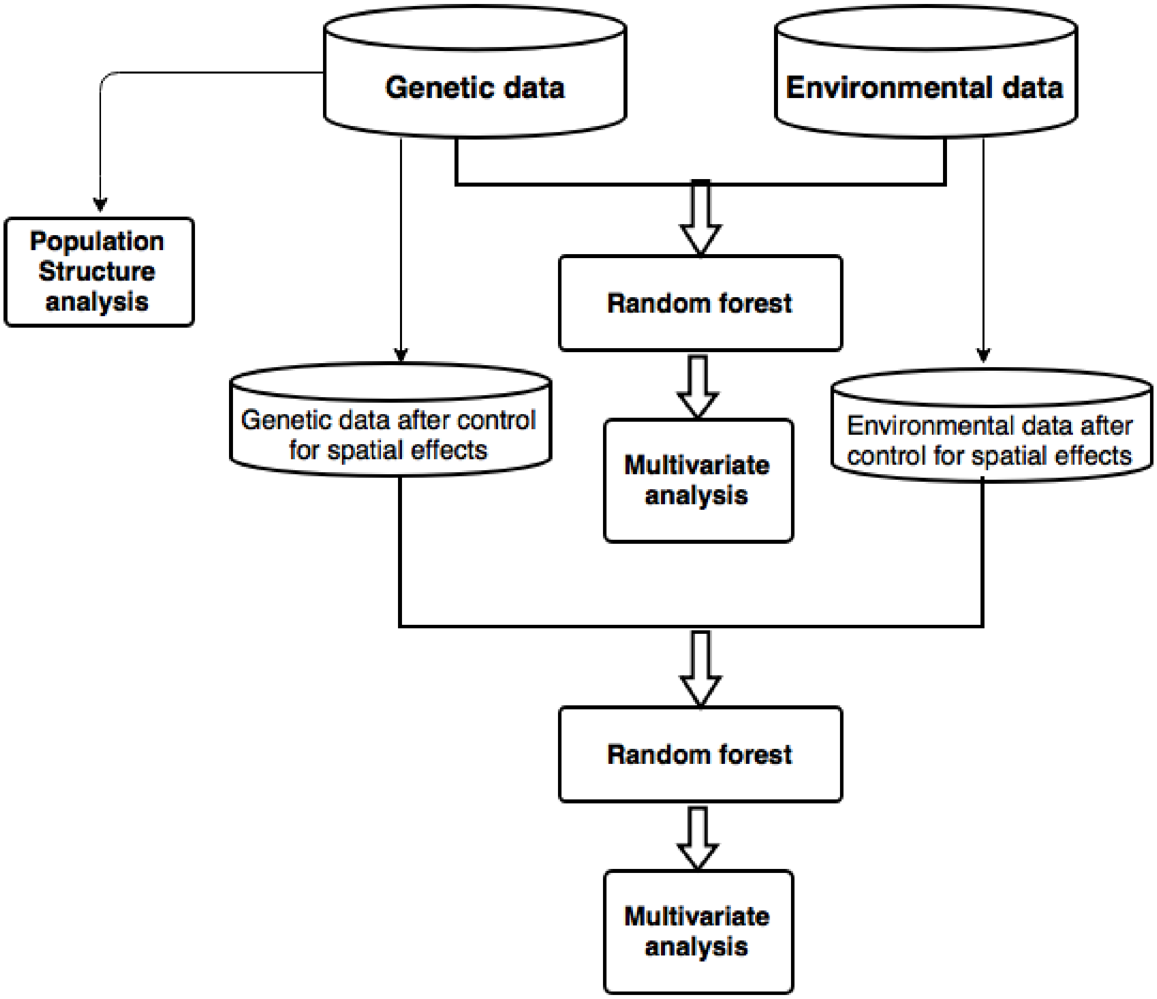

Genetic data
Environmental data
Population Structure analysis
Random forest
Genetic data after control for spatial effects
Multivariate analysis
Environmental data after control for spatial effects
Random forest
Multivariate analysis

Figure 20 Workflow of proposed environmental association analysis. Genetic data are to infer population structure. Two random forest analyses are performed on the environmental and genetic data with and without spatial effects controlled. Important genotypes selected by both models can then be run into multivariate analysis to investigate the underlying associations.

## 4.2 Data Preparation

### 4.2.1 Genetic Data

A total of 252 adult scallops were collected by hand or bottom trawl from a total of 12 locations across the entire range of the species between 2011 and 2013 (Table 3, Figure 21) with a minimum of 12 scallops per population (mean value of 20.4 ± 2.8 scallops). The details of DNA extraction, library preparation and sequencing can be found in Appendix C.1.

SNPs were detected using the *de novo* pipeline in STACKS v.0.9999 [90]. We tested several variations of the STACKS parameters with repeated runs to help ensure appropriate choice and examine sensitivity of SNP calling. The final dataset was filtered using PLINK v.1.07 [91] [92] to include only SNPs that were present in 75% of individuals in SNP discovery and calling. Therefore, all SNPs included in the analysis were present in 75% of individuals with a minor allele frequency greater than 5%. Furthermore, we excluded individuals with more than 20% missing loci from the analysis. Hardy-Weinberg Equilibrium (HWE) describes the situation where allele frequencies and genotype frequencies are in balance; in diploid organisms with alleles "A" and "a" with respective frequencies $p$ and $q$, the expected genotype frequencies are $p^2$, $q^2$, and $2pq$ for the homozygous "pp", homozygous "qq", and heterozygous "pq" genotypes, respectively. Devations from HWE at a locus can indicate unusual patterns of inheritance and key assumptions of population structure. Loci were filtered for HWE using the program GENEPOP v.4 [93], excluding loci out of equilibrium in six or more populations from the analysis (<0.7% of all loci). The final genetic data contains genotypes of 245 samples at 7163 SNP loci.

Before running the following environmental association analysis, missing genotypes in the genetic SNP data need to be addressed since many statistical models cannot handle missing values. Removing missing values is not a good approach because it will cause the loss of too much important information. Therefore, inference or *imputation* of missing genotypes was performed. A method based on weighted k-nearest neighbours (KNN) called KNNcatImpute [95] was used to impute the missing genotypes in our genetic SNP data. The imputation was performed using the *scrime* package in R [99].

### 4.2.2 Environmental Data

Environmental data was collected from several databases including the Department of Fisheries and Oceans Canada, BioChem (DFO 2014) (years 2009-2014), and AZMP (DFO 2015), and from the National Aeronautics and Space Administration in the United States of America (NASA, years 1990-2010), and the MODIS satellite database (NASA Goddard Space Flight Center Ocean Ecology Laboratory 2014) (years 2002-2013). Data was averaged over the range of years available to remove the signatures of short-term variation in the marine environment. Collected variables included water temperature, salinity, sigmaT that is a measure of water density which is a product of the temperature and salinity, and chlorophyll A that can be used to assess the levels of phytoplankton in a system, and nutrient concentrations including silicic acid ($SiO_4$), nitrite ($NO_2$), nitrate ($NO_3$) and phosphate ($PO_4$). Some important environmental variables that will be used in the following analysis and their corresponding abbreviations are listed in Table 4.

Data validation and preparation were completed using R [96]. The details of how we normalized data and removed some outliers were described in Appendix C.2. The final environmental data set contained 90 variables spanning all variable data types.

Table 3 Information of sampling locations and the number of individuals at each sampling site.

| **Location Name** | **Abrreviations** | **Number of individuals collected at this site** |
|---|---|---|
| Sunnyside, NL | SUN | 20 |
| Little Bay, NL | LTB | 21 |
| Magdalen Islands | MGD | 21 |
| Northumberland Strait | NTS | 22 |

| Location Name | Abrreviations | Number of individuals collected at this site |
|---|---|---|
| Passamaquoddy Bay | PSB | 12 |
| Bay of Fundy | BOF | 22 |
| Scotian Shelf – Middle | SSM | 19 |
| Gulf of Maine Inshore | GMI | 20 |
| Browns Bank | SSB | 22 |
| Gulf of Maine offshore | GMO | 22 |
| George's Bank | GEO | 22 |
| Mid-Atlantic Bight | MDA | 22 |

Table 4 Important environmental variables and corresponding abbreviations.

| Environmental variables | Abrreviations |
|---|---|
| Surface average winter temperature | SAWT |
| Depth average summer temperature | DAST |
| Depth max temperature | DMT |
| Depth average autumn salinity | DAAS |
| Surface average winter chla | SAWCA |
| Depth average summer $SiO_4$ | DASSiO4 |
| Surface average winter $SiO_4$ | SAWSiO4 |
| Depth min PO4 | DMIPO4 |
| Surface max chlorophyll A | SMCA |
| Surface average spring chlorophyll A | SASCA |
| Surface average autumn salinity | SAAS |
| Depth average spring temperature | DAST |
| Depth average summer temperature | DASUT |
| Surface average spring $SiO_4$ | SASSiO4 |
| Surface average autumn $SiO_4$ | SAASiO4 |

| Environmental variables | Abrreviations |
|---|---|
| Depth average spring $NO_2$ $NO_3$ | DASNN |
| Depth min $NO_2$ $NO_3$ | DMNN |
| Depth max $NO_3$ | DMN3 |
| Depth average autumn $PO_4$ | DAAPO4 |
| Depth max $PO_4$ | DMPO4 |

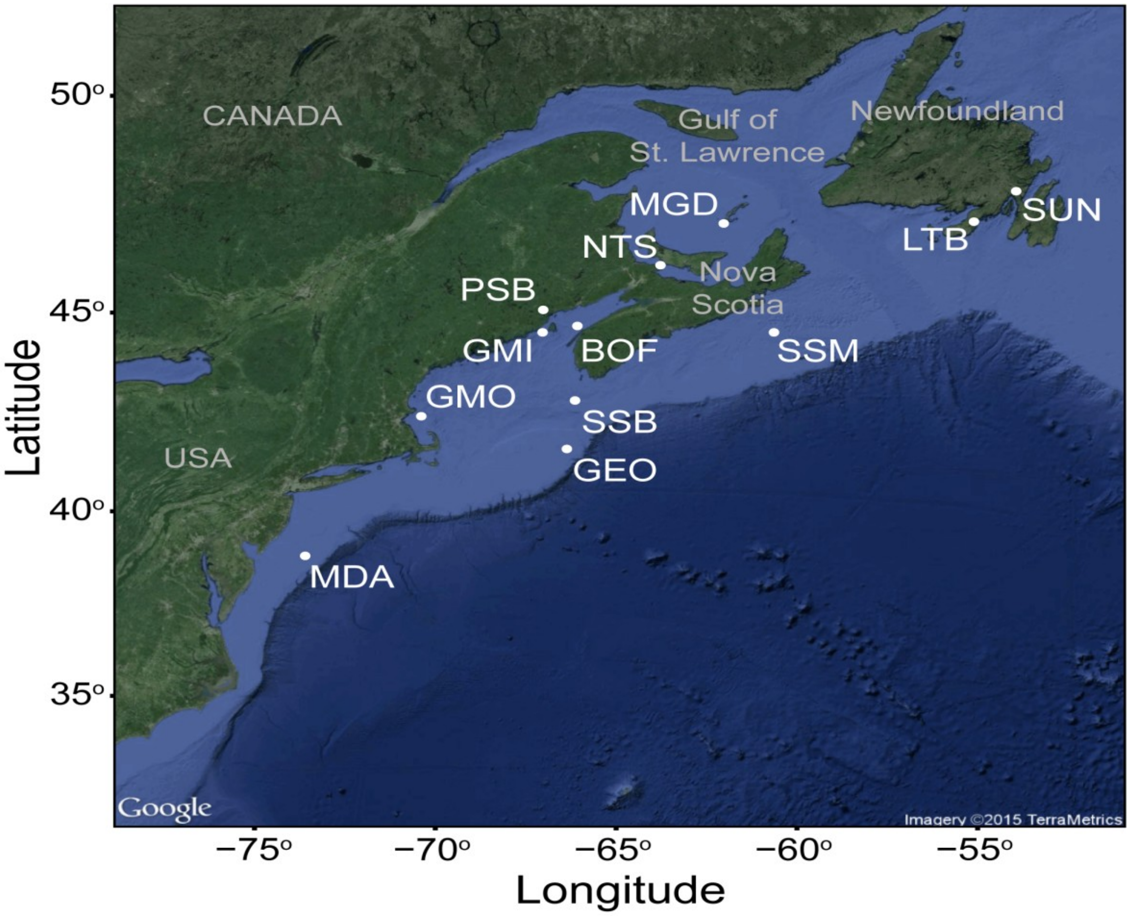


Figure 21 Map of the 12 sea scallop collection sites in the Northwest Atlantic Ocean. Populations are marked by abbreviations that are the same as Table 3.

## 4.3 Population Structure And Environmental Association Analysis

### 4.3.1 Population Structure

A Bayesian model-based program, *STRUCTURE* [100], was used to estimate the number of distinct populations in the sea scallop, as well as the admixture proportions for each individual. *STRUCTURE* uses Markov chain Monte Carlo (MCMC) algorithm to differentiate populations from multilocus genotype data. In our sea scallop data, 200,000 MCMC interations were run in *STRUCTURE* after an initial discarded burn-in phase of 50,000 runs. The tested number K of genetic clusters was allowed to vary between 1 and 15, and for each value of K, *STRUCTURE* runs were replicated 3 times in order to get the most stable likelihood values. The number of correct clusters was estimated based on the log probability value L(P) and as well as the rate of change in the log probability with respect to K values (also called as ΔK method) [101]. The *STRUCTURE* results were summarized using the Structure Harvester [102] and visualized using software *CLUMPP* [133] and *distruct* [134].

Plot of L(P) values (Figure 22(a)) showed that L(P) values reached a plateau at around K=2. From the deltaK plot, we can see a large peak at K=2 and a much smaller peak at K=4 (Figure 22 (b)), which indicated the number of inferred clusters best fitting our data was 2. Each individual in the bar plot was represented by a vertical line and each line was filled by different colors that represent the proportions of membership in the inferred clusters [100]. Bar plots of individuals for K=2 and K=4 revealed a substantial degree of admixture, as seen by the large number of individuals composed of bars of >1 color (Figure 22 (c)). For K=2, SUN, LTB, MGD and NTS were identified as a group and all the other populations were clustered together. A larger degree of admixture was seen in the sampling sites SUN, MGD, NTS, PSB, SSM and GMO. This degree of admixture might affect the detection of environmentally adaptive SNPs since allele frequencies at those adaptive SNPs might decrease or disappear when admixed individuals contain alleles from more than one population. The influence of admixture on the identification of adaptive loci was explored in the following section. When K=4, no clear evidence of population structure was observed in the majority of sites, however, in this case SUN, MGD and NTS were more similar to one another, while LTB was more distinct from other three sites.

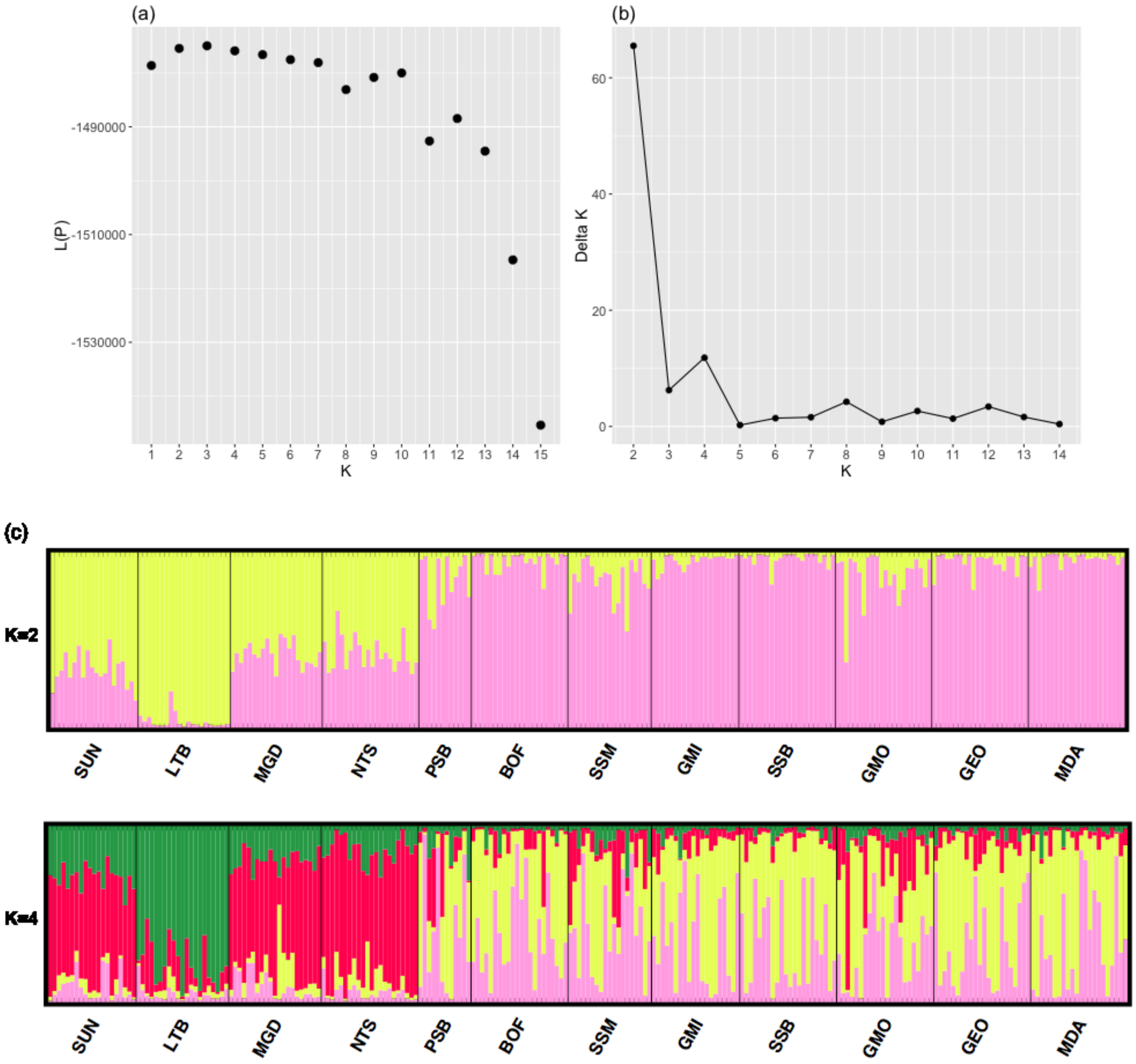


Figure 22 *STRUCTURE* results showing the spatial structure of the 12 sea scallop populations. (a) Plot showing the estimated log probability L(P) for each K value (b) The deltaK plot for detecting the optimal number of populations. (c) Bar plot of inferred ancestry of individuals obtained with the *STRUCTURE* analysis. In the bar plot, each vertical line represents an individual and is filled with colors whose lengths are proportional to the estimated memberships in the inferred clusters. The abbreviations for all the populations are listed in Table 3.

### 4.3.2 Random Forest Classification Of Sea Scallop Data

The 7163 SNPs in our study are bi-allelic genetic markers, which indicates that only two alleles (denoted as A and B) can be present at each site, and all individuals will have a genotype of AA, AB or BB. To encode the SNP data, we transformed the genotypes to three categories and performed three-class RF classification. All the environmental variables were used as predictors of SNP genotypes, with a total of 7163 predictive models built, one per SNP. Classification was performed using the *randomForest* package [107] in R. Ten-fold cross validation was used to evaluate test-set accuracy, and we used AUC as an evaluation measure for the predictions of all the 7163 random forest models. Normally the AUC calculation is only applicable to a binary classification problem. Therefore, in our case, a different AUC calculation method proposed by Hand & Till in 2001 [108] was used for our three-class classification tasks. This method calculates AUC for each pair of classes and averages the AUC for all pairs of classes [108]. The multi-class AUC measure was performed using the package *pROC* in R [109].

Based on the value of AUC, we obtained a ranked list of SNPs (Figure 23 (a)) and selected the 14 top-ranked SNPs whose AUC value was larger than 0.7 for the following analysis. In the top 14 SNPs, SNP "4668_81" showed the best classification accuracy with AUC around 0.96. The majority of SNPs have AUC values close to 0.5, while 85 SNPs (around 1%) had an AUC larger than 0.65. The AUC values for the top ranked 200 SNPs were also illustrated in Figure 23 (b). The accuracies at these top 14 SNPs were also calculated by sampling site. At SNPs "17055_73" and "25627_51", the accuracies at sampling site SSB were larger than other sites owing to the smaller degree of admixture in the SSB site (as shown in Figure 22 (c)). Poorer prediction performance was seen in the sites of NTS and MGD sites at some SNPs. Since all genotype predictions for a given site must be identical (since the environmental predictors are the same), it is not surprising that sites with greater admixture are likely to exhibit lower classification accuracies.

The 14 trained RF models for the top 14 SNPs were used to rank environmental variables using the variable importance function built in random forests. The permutation importance was used to estimate the variable importance. Since the importance values of some variables might be negative, which indicate that those variables are not important, the exponential values of the variable importance array were calculated. Then these values were averaged over the total importance sum of all environmental variables to calculate an importance percentage for each environmental variable. In the ten-fold cross validation, 10 RF models were built and importance percentage array was calculated for all the environmental variables by each built RF model, to ensure the estimated variable importance values are reliable. Subsequently we averaged the 10 importance percentage arrays to get average importance percentages. Therefore, in each of the 14 trained RF models for the top 14 SNPs, one average importance percentage array was generated. In order to explore which environmental variables contribute the most to the prediction of genotypes at all the 14 SNPs, we averaged again on all the 14 average importance percentage arrays to get the final variable importance ranking list (as shown in Figure 24). Among these environmental variables, depth average summer salinity, surface average spring sigmaT, and depth min salinity were the three most important predictors for the classifiers. In addition to these salinity variables, depth average winter temperature, surface average autumn temperature and depth min temperature were also ranked as important variables. The importance percentages of the most important predictors are low. This is because the importance percentages of the same environmental variables may differ for different SNP outputs, and relatively large importance percentages might decrease due to the averaging on all the 14 importance percentage arrays.

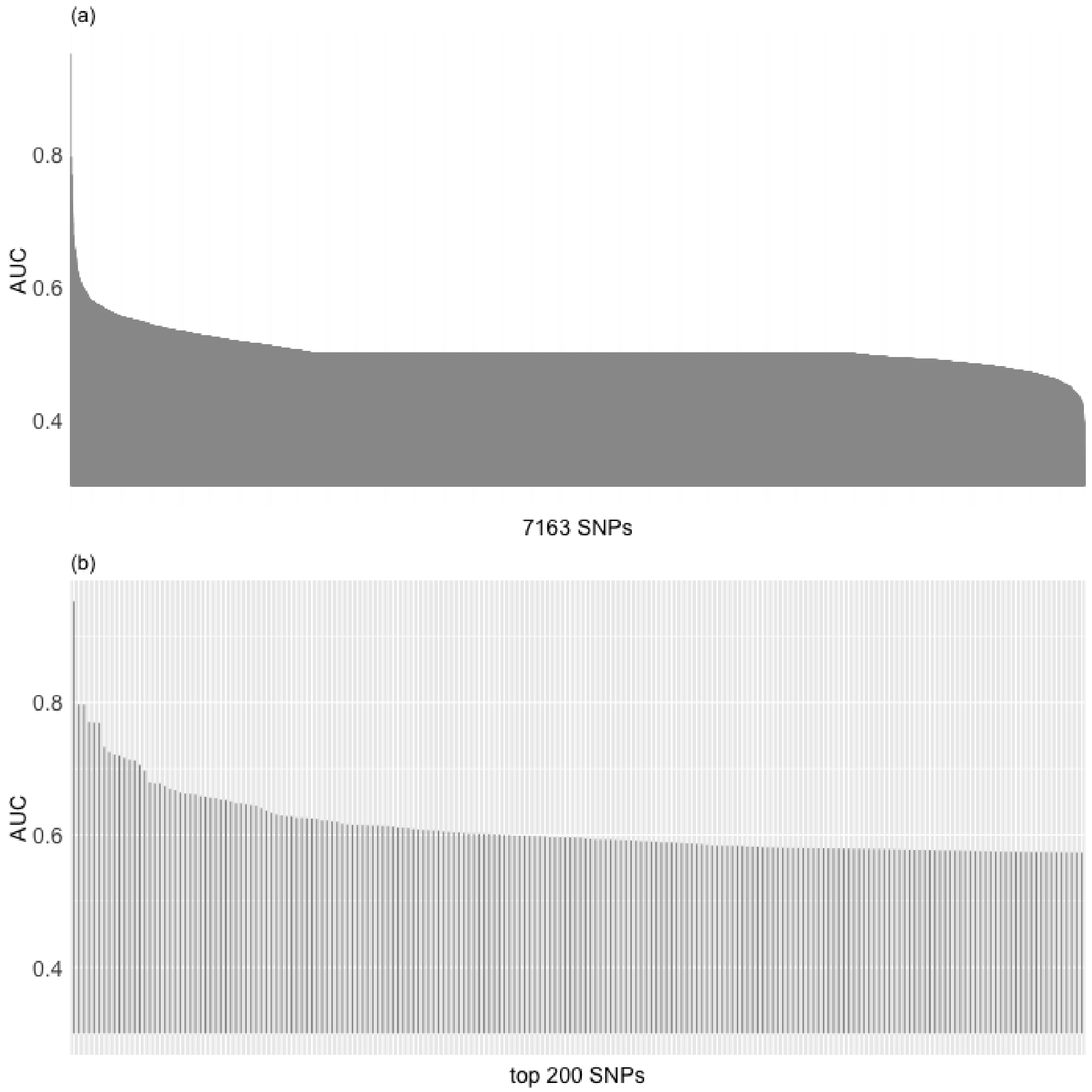


Figure 23 (a) AUC values for all the 7163 SNPs. (b) AUC values for the top ranked 200 SNPs.

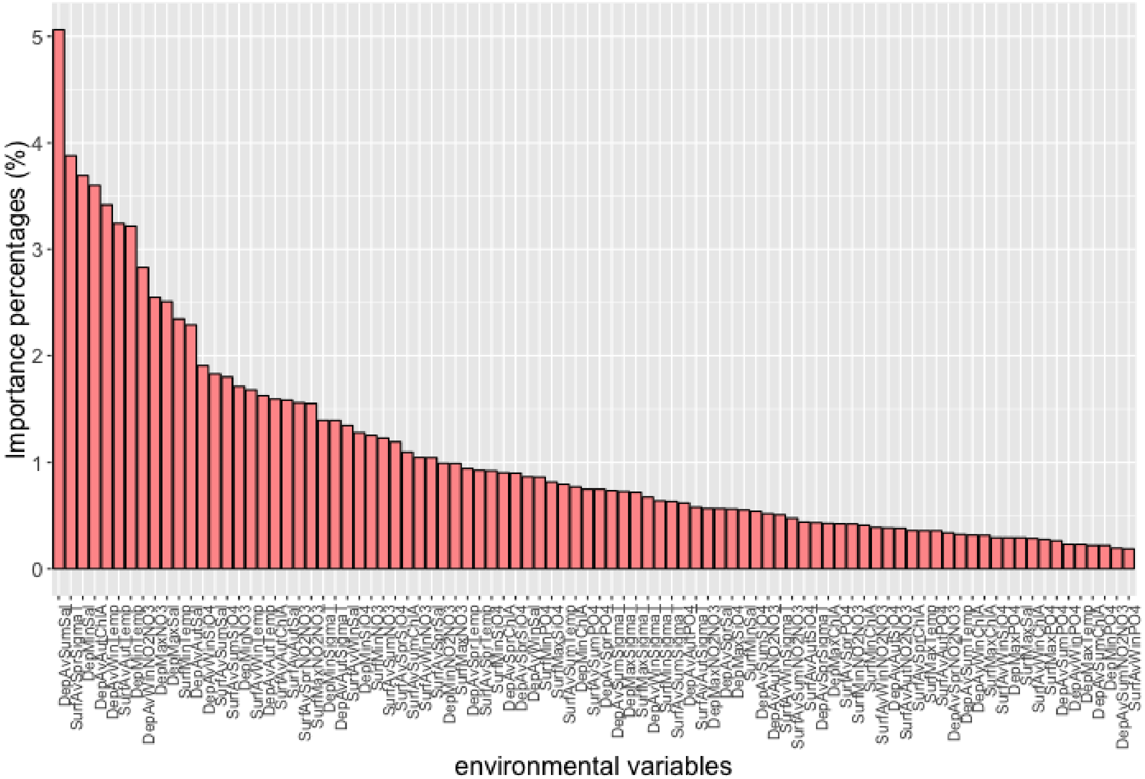


Figure 24 Averaged variable importance percentages for all the environmental variables.

### 4.3.3 Association With Environmental Data Using Redundancy Analysis

In order to validate the performance of RF model in the selection of environment-associated loci and further explore the underlying associations with environmental factors, redundancy analysis (RDA) was conducted on environmental data and the selected SNP data (Figure 25). RDA is a statistical model that is similar to PCA. However, where PCA is an unconstrained ordination analysis that extracts the maximum variance of one set of variables, RDA is a constrained ordination analysis that seeks dimensions that capture the maximum variance in the response data, where the response data are linear combinations of the explanatory variables [110].

Assume response data is a matrix $Y$ ($n, p$) and explanatory variables are in a matrix $X$ ($n, q$), which can be displayed as follows:

$$Y = \begin{bmatrix} y_{11} & y_{12} & \dots y_{1p} \\ y_{21} & y_{22} & \dots y_{2p} \\ & & \dots \\ y_{n1} & y_{n2} & \dots y_{np} \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} & x_{12} & \dots x_{1q} \\ x_{21} & x_{22} & \dots x_{2q} \\ & & \dots \\ x_{n1} & x_{n2} & \dots x_{nq} \end{bmatrix}$$

Where *n* indicates the number of samples, *p* and *q* refer to the number of reponse variables and explanatory variables. The RDA model is built by firstly performing a multiple linear regression (MLR) for each feature in *Y* on matrix *X.* Then PCA is conducted on the fitted values of MLR to generate canonical eigenvectors *E* and eigenvalues, where the matrix *E* is used subsequently to calculate ordination scores for response and explanatory variables respectively [110].

The top 14 SNPs were selected from the SNP ranking list based on the AUC values. Owing to the multicolinearity in the environmental variables, we did a PCA on all the variables and extracted the first five PCA axes that explained over 90% of the variance of the environmental data set. Subsequently, we selected the two variables that were most highly correlated with each of these first five PCA axes. The extracted 10 important environmental variables accounted for 6.76% of the total variance in the environmental data set. The RDA model was performed using these 10 environmental variables and the allele frequencies at the 14 SNPs.

Part of the variance in genetic data can be explained by geographical distance due to more-frequent interbreeding and migration between proximal sample sites. In order to control for these effects, a partial RDA aiming to remove the contributions of conditional variables to the variance of response data was performed. The variance explained by the environmental variables in RDA was 31.71%, with the first 2 RDA axes accounting for 21.90% and 5.05% of the variance, respectively (Figure 25(a)). A permutation test of significance yielded a *p*-value smaller than 0.001 and *F* value of 10.864. Another permutation test was used to calculate the amount and significance of variation that could be attributed to each explanatory variable. The results showed that DMT (test statistic *F* = 6.2827, *p*-value <0.001) and DMIPO4 (test statistic *F* =5.7803, *p*-value <0.001) presented a strong correlation with allele frequencies. DAST (test statistic *F* = 2.2066, *p*-value 0.039) and DASSiO4 (test statistic *F* = 2.2508, *p*-value 0.032) had associated *p*-

values that were larger but still less than 0.05, which means they were still ‘significant’ correlations. In the partial RDA model, environmental factors contributed to 12.94% of the variance and geographic distance explained 19.34% of the total variance (Figure 25 (b)).

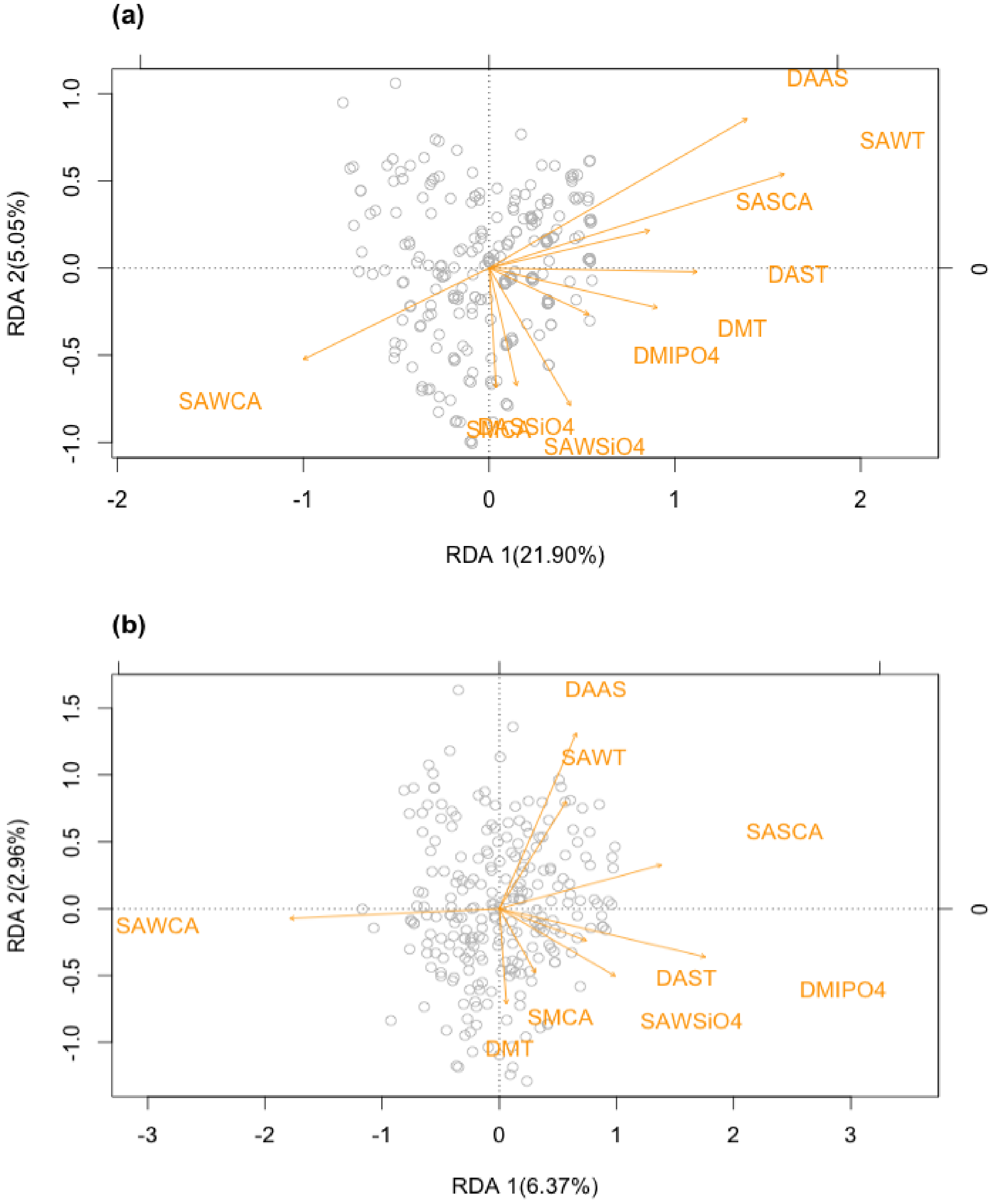


Figure 25 (a) RDA biplot for associations between the allele frequencies at the top14 SNPs and 10 selected environmental variables. (b) Partial RDA biplot with the removal of geographical effects. Environmental variables are plotted as arrows and objects plotted as grey points. The explanations of these abbreviations are listed in Table 4.

### 4.3.4 RF Classification With Control Of Geographical Effects

From the results of partial RDA, we found that geographical distance accounted for larger proportions of the variance in our selected genetic data than did environmental factors. In order to remove the effects of spatial distance on environmental factors and genetic data, we used an inverse Cholesky transformation method on these two data sets. Inverse Cholesky transformation [127] is a method to decorrelate highly correlated variables. Suppose the environmental data is *E,* which are the combined effects of spatial independently environmental variables (*E'*) and the geographical distance (*D*). Then using the inverse Cholesky matrix of *D* times the environmental data *E* will generate the transformed data that are not correlated with *D* (which is *E'*). In order to validate the performance of inverse Cholesky transformation method on controlling spatial effects, first we did a Moran's I test to measure the spatial autocorrelation in the original environmental variables. Figure 26 shows the p-value for the measured spatial autocorrelations on all environmental variables. The p-value represents the probability that no spatial autocorrelation is detected at these environmental variables. In Figure 26, the p-values are log transformed and then absolute values are used as the y-axis values to better present those spatial correlated variables in the top of the plot. In our 90 environmental variables, spatial autocorrelation was found in 45 variables. Then we did another Moran's I test on the environmental data after inverse Cholesky transformation, the number of environmental variables that showed spatial autocorrelation decreased to eight, which demonstrate that spatial effects were controlled to some extent.

Another RF regression model was performed on the corrected genetic and environmental data. Since the values of corrected genetic data are continuous, we adopted the mean squared error (MSE), which is the average of the square of diffirences between predicted value and actual value, as an evaluation measure to assess the performance. A smaller MSE value indicates better performance of the RF regression model. The estimated MSE for the 280 SNPs with MSE smaller than 0.75 can be found in Figure 27. In order to compare the performance of model with spatial effects controlled the previous uncorrected model, the top 14 SNPs were also selected from the MSE ranking list. Only two SNPs, "4668_81" and "15645_89", were selected by both the corrected and uncorrected RF models. PCA was also used on the transformed environmental variables and 10 important environmental factors were picked using the same method. The extracted 10 environmental variables accounted for around 6.12% of the total variability in the environmental data set. RDA was once again conducted on these two data sets (as shown in Figure 28).

The variance explained by environmental variables in RDA was 44.21%, where the first 2 RDA axes accounted for 39.99% and 2.27%. Then a permutation test on the RDA model illustrates that this model was significant with a *p*-value smaller than 0.001 and *F* value is equal to 18.545. Another permutation test was used to calculate the significance of variation explained by each explanatory variable. The results showed that all the 10 environmental variables presented a strong correlation with allele frequencies (*p* value smaller than 0.001).

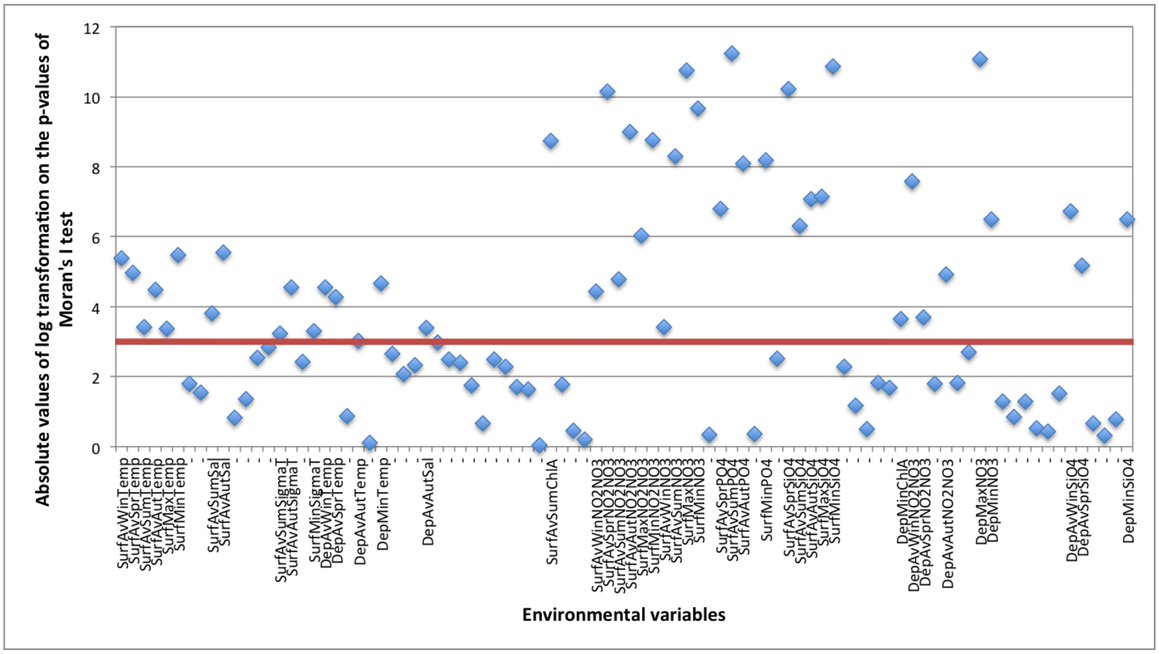


Figure 26 The absolute values of log transformation on the p-values of Moran's I test for spatial autocorrelation on all environmental varibales. The red line specifyies the absolute values of log transformation on the p-value of 0.05 (which is 2.995732). Any points above the red line means that those environmental factors are correlated with spatial distance; only those variables found to be significant are labelled on the x-axis.

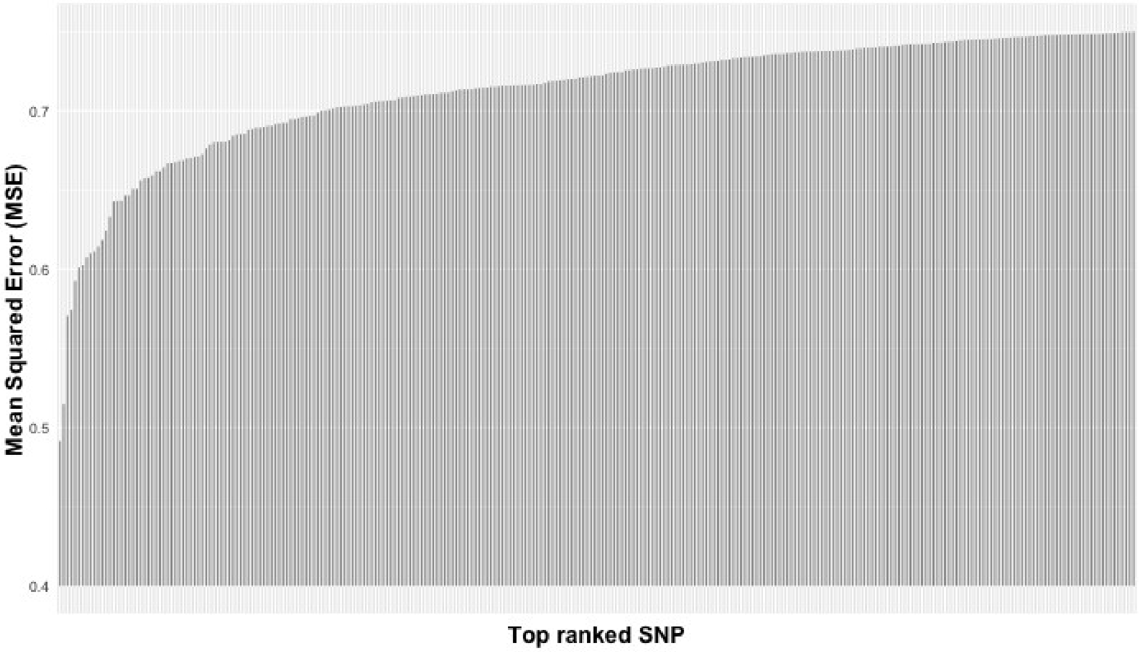


Figure 27	MSE values for the top 280 SNPs with MSE smaller than 0.75.

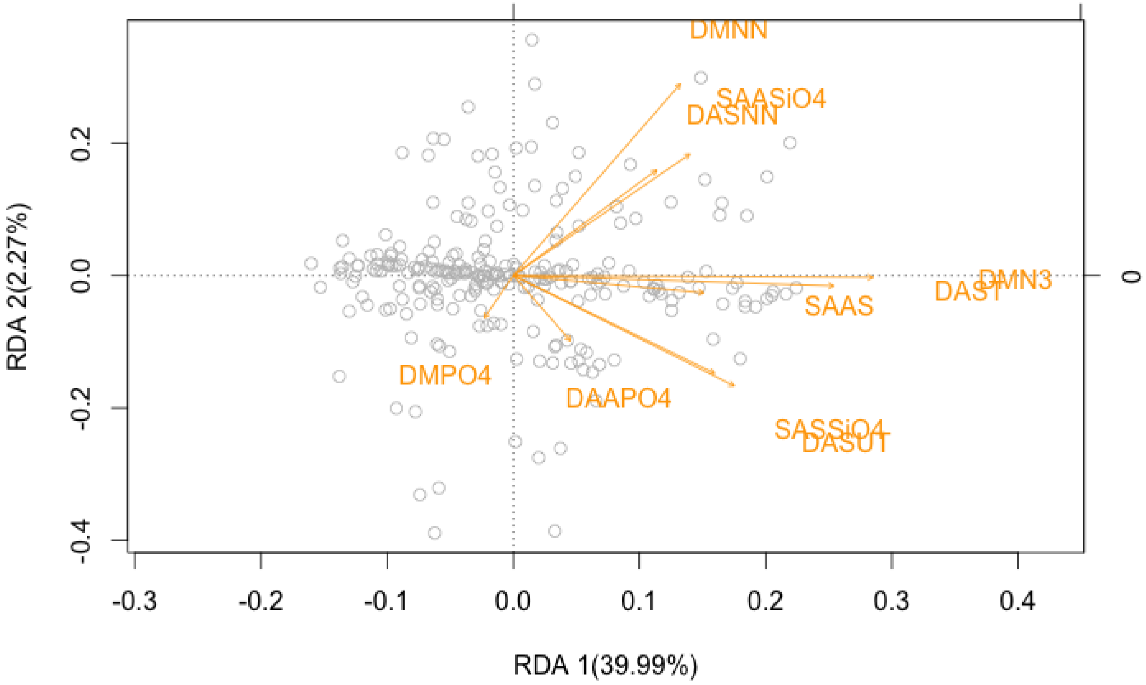


Figure 28 RDA biplot for associations between the allele frequencies at the top 14 SNPs and ten selected environmental variables with the removal of spatial effects. The explanations of these abbreviations are listed in Table 4.

## 4.4 Conclusions

RF offers a non-linear model to explore the associations between genetic data and environmental data, and provides an automatic calculation of variable importance that helps to find which environmental variables are the primary drivers for the spatial patterns of identified adaptive variants. The RF models differ from other most regression models in the identification of adaptive loci. Most regression models regress allele frequencies on a range of environmental variables to test correlations between allele frequencies and environmental gradients while using other variables to account for neutral effects. By contrast, our RF models treat those loci with high prediction performance as signs of selection. However, the correlations identified by this naïve model could suffer from high false positive rates owing to the spatial autocorrelations in genetic and environmental data. The RF regression model on the corrected was constructed to address this problem. Moran's I test on uncorrected and corrected data demonstrated that the geographic effects were well controlled. However, removing the geographic signal might also remove some adaptive signal as well since our environmental variables showed strong correlations with geographic distance.

The variable importance percentages calculated by RF models indicated that salinity (depth average salinity and depth min salinity), depth average winter temperature, surface average spring sigmaT and depth minimum temperature as the most important environmental variables. The RDA analysis on the selected environmental variables and SNPs indicated that depth max temperature and depth minimum $PO_4$ had strong correlations with spatial patterns in genetic data. The identified important variables differed between these two methods since only 10 environmental factors with low multicollinearity were kept in the RDA analysis.

# CHAPTER 5 CONCLUSION AND FUTURE WORK

In this thesis, we have developed and validated approaches to infer, visualize and curate microsatellite data, and developed a new hybrid statistical / machine-learning approach to infer SNP loci that covary with environmental variables. MEGASAT addresses the disadvantages associated with electrophoresis-based methods by inferring microsatellite genotypes from NGS data, which could achieve higher genotyping accuracy, lower consumable costs and some other potential benefits. It uses different functions to deal with sequencing errors and designs algorithms to filter out amplification artifacts. Furthermore, it offers a visualization function to present the microsatellite data used to call genotypes, which enables reviewers to validate the scored genotypes and manually edit the mis-scored genotypes. The estimated genotyping error rates using repeat-genotyping method and pedigree-based method demonstrated that MEGASAT could achieve high genotyping accuracy. The comparison of performance of MEGASAT with MicNeSs [22] further demonstrated the better performance of MEGASAT in genotyping accuracy and running time.

Subsequently, we presented a machine-learning model, random forest, to conduct environmental association analysis with statistical models. We developed an approach that used RF classification to infer environmentally associated gene variants, and performed an RF regression to extract gene variants that are only correlated with environmental gradients. This regression model was built on spatially independent genetic and environmental data sets via using inverse Cholesky transformation to partition raw data into spatially correlated and uncorrelated components. After extraction of adaptive SNPs, the ordination technique RDA was used to futher investigate the relationships between identified adaptive genetic variabilities and environmental gradients.

These new methods have already generated new insights into datasets collected for conservation biology purposes, but further refinements are possible that would further increase their utility.

## 5.1 SNP Detection In Identical-length Alleles

NGS provides the raw sequence reads for microsatellite genotyping, which gives the opportunity to detect SNPs in microsatellite-containing sequences in addition to the length variants currently identified by MEGASAT. If sequence variations can be found in identical-length alleles, those assigned alleles will give higher-resolution information about more closely related individuals. The pedigrees shown in Figure 3.6 and 3.7 still have a large amount of statistical uncertainty; if reliable SNPs can be used to further differentiate alleles, then statistical support would be expected to increase. This is because the high mutation rate of microsatellites, which results in individuals with the same allele lengths that are not identical by descent but instead the product of convergence [130, 131]. However, DNA sequences are less likely to converge due to the large number of nucleotides in a typical microsatellite array.

We developed a preliminary new version of MEGASAT that can use SNP information to distinguish alleles that would not have been detectable by the version described in Chapter 2. This function can differentiate alleles of the same length based on their nucleotide sequences. These refinements required changes to the MEGASAT implementation including using a hash table to store the length and counts of alleles, and genotype assignment from the sequence records in the hash table. Applying this version to the guppy dataset, we calculated the number of additional real variations we found for different alleles (as shown in Table 5). For some alleles at a locus, few variations were detected but some other alleles show large amounts of variation.

Table 5 Number of additional variations found by the SNP-version script for some loci on a sample data set (around 400 guppy individuals). The second and third column shows the number of alleles detected by original script and SNP-version script. The last column presents the number of variations found at different alleles.

| Locus name | Number of alleles detected by | Number of alleles detected by the new | Alleles detected by original script (number of variations found at |
|---|---|---|---|

|  | original script | version | this allele) |
|---|---|---|---|
| BF-047 | 6 | 71 | 30(18) 32(3) 34(20) 36(3) 38(23) 40(4) |
| BF-052 | 11 | 36 | 24(3) 27(3) 30(6) 33(1) 39(4) 42(1) 45(8) 48(6) 51(7) 54(2) 63(1) |
| BF-143 | 8 | 26 | 47(5) 49(2) 51(11) 53(2) 55(1) 57(2) 77(2) 79(1) |
| BF-174 | 11 | 223 | 36(37) 42(47) 44(1) 46(49) 48(15) 50(11) 54(11) 56(2) 58(40) 60(1) 62(9) |
| BF-185 | 11 | 159 | 38(10) 41(3) 44(7) 47(18) 50(12) 53(5) 56(13) 59(26) 62(32) 65(26) 68(7) |
| BF-190 | 9 | 129 | 45(9) 47(3) 49(18) 51(8) 57(36) 59(42) 61(1) 63(6) 69(6) |
| BF-225 | 7 | 148 | 61(6) 63(11) 65(16) 67(45) 69(22) 71(45) 73(3) |
| BF-230 | 16 | 300 | 42(13) 48(2) 60(23) 63(13) 66(51) 69(14) 72(16) 75(32) 78(34) 81(18) 84(21) 87(15) 90(18) 93(18) 96(7) 99(5) |
| BF-231 | 10 | 81 | 52(3) 56(13) 62(3) 66(3) 68(26) 70(6) 72(11) 74(12) 80(2) 82(2) |
| BF-247 | 11 | 72 | 48(1) 51(3) 54(1) 57(4) 60(3) 63(18) 66(16) 69(12) 72(8) 75(4) 84(2) |
| BF-262 | 11 | 45 | 64(3) 70(7) 73(1) 76(2) 79(6) 82(4) 85(7) 88(8) 91(2) 94(3) 97(3) |

However, in the called allele variants, there still exist a proportion of variants that are caued by sequencing errors. In order to address this problem, several methods could be used to distinguish real SNPs from sequencing errors. One way is to utilize models proposed by some [e.g. 128], which uses the quality scores provided by FASTQ files and compare the sequence reads across a large population of individuals using a probabilistic model to detect the SNPs. SNP detection models built in some software pipelines (e.g. Stacks [129]) could also be used to filter out those variants containing sequencing errors. Once genotyping errors were filtered out, the genotyping output of this SNP-version script could also be used into FRANz to build pedigrees. However, if there are enough alleles identified by the previous script, which give enough genetic resolution for inferring parentage information, adding more variants might not make any difference unless we need to build pedigrees from large numbers of individuals.

## 5.2 FUTURE WORK

Although MEGASAT enables microsatellite genotyping in a much faster and more accurate way by using NGS data, there still exist some improvements that could be done for this program. In some microsatellite loci, the flanking regions might not be identical and can have length variations even for the same microsatellite locus. In order to identify those real variations in the flanking region, several changes could be made in the microsatellite identification algorithm to enable MEGASAT to recognize different length variants in flanking regions. Another future refinement would be interactive curation. As we stated in the section 2.5, MEGASAT provides visual representations of allele calls to allow users to verify the inferred genotypes and manually curate few wrongly scored genotypes. Though few genotypes need to be edited, manual editing the genotyping text file is inconvenient. A better way could be achieved by generating interactive histograms that connect the histograms to the text file. Any updates made in the histograms could be incorporated into the text file as well.

The proposed random-forest based workflow worked well to identify adaptive loci and environmental gradients associated with adaptive loci. In this method, we used PCA to extract important environmental variables that are not highly correlated, which aimed to run into the following RDA analysis since RDA assumes explanatory variables are independent. However, the extracted environmental variables just accounted for around 6% of the total variance, which might cause the loss of environmental information. This could result in some important environmental variables failing to be explored with adaptive gene variants in the RDA analysis. A refinement of this approach could use the variable importance inferred by RF to extract important variables from the whole data set. PCA can be performed on the extracted variables and PCA axes can be used into RDA analysis to find which PCA axes are most important. Important environmental factors can be identified via the contributions of each environmental variable on the important PCA axes. Another way could be using LASSO or ridge regressions [104] to do feature selection on environmental data since these two regression methods perform well on collinear data. Then we can apply forward selection in RDA analysis using these selected variables to further determine which subset of variables can be explanatory variables for RDA.

In our method, we used the corrected genetic and environmental data after inverse Cholesky transformation into RF regression to control for spatial autocorrelations. However, we did not implement any changes in the RF model. In some other traditional statistical models for environmental association studies, latent factors or covariates [118, 120] are used in regression models to account for non-adaptive effects, which indicate the probability of incorporating covariates or any latent factors into RF models to account for some neutral effects.

# BIBLIOGRAPHY

[1] Jean-Pierre Feral. How useful are the genetic markers in attempts to understand and manage marine biodiversity. *Journal of Experimental Marine Biology and Ecology,* 268(2):121-145, 2002.

[2] Francisco Rodriguez-Valera. Approaches to prokaryotic biodiversity: a population genetics perspective. *Envrionmental Microbiology,* 4(11):628-633, 2002.

[3] Don C. Delong. Defining biodiversity. *Wildlife Society Bulletin,* 24 (4):738-749, 1996.

[4] Peter J. Edwards and Cyrus Abivardi. The value of biodiversity: where ecology and economy blend. *Biological Conservation.* 83(3):239-246, 1998.

[5] Craig Bullock. The economic and social aspects of biodiversity. 2008.

[6] How many species are we losing? Retrieved from http://wwf.panda.org/about_our_earth/biodiversity/biodiversity/. 2015.

[7] Carol Kearns. Conservation of Biodiversity. *Nature Education Knowledge,* 3(10):7, 2010.

[8] Brian Ford-Lloyd and Kevin Painting. Measuring genetic variation using molecular markers. *International Plant Genetic Resources Institute,* 1996.

[9] Yoseph Beyene, Anna-Maria Botha and Alexander A. Myburg. A comparative study of molecular and morphological methods of describing genetic relationships in traditional Ethiopian highland maize. *African Journal of Biotechnology,* 4(7):586-595, 2005.

[10] Linda Mondini, Arshiya Noorani and Mario A. Pagnotta. Assessing plant genetic diversity by molecular tools. *Diversity,* 1:19-35, 2009.

[11] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology,* 26(10):1135-1145, 2008.

[12] Michael L. Metzker. Sequencing technologies – the next generation. *Nature Reviews Genetics,* 11(1):31-46, 2010.

[13] P. Kumar, V.K. Gupta, A.K. Misra, D. R. Modi and B. K. Pandey. Potential of molecular markers in plant biotechnology. *Plant Omics Journal,* 2(4):141-162, 2009.

[14] John S. Gray. Marine biodiversity: patterns, threats and conservation needs. *Biodiversity and Conservation,* 6(1):153-175, 1997.

[15] Stephanie Manel and Rolf Holderegger. Ten years of landscape genetics. *Trends in Ecology & Evolution,* 28(10):614-621, 2013.

[16] Hans Ellegren. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet Nature Reviews Genetics*, 5(6):435-445, 2004.

[17] E. Guichoux, L. Lagache, S. Wagner, P. Chaumeil, P. Léger, O. Lepais, C. Lepoittevin, T. Malausa, E. Revardel, F. Salin, and R.J. Petit. Current trends in microsatellite genotyping. *Molecular Ecology Resources,* 11(4):591-611, 2011.

[18] Kimberly A. Selkoe and Robert J. Toonen. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters,* 9(5):615-629, 2006.

[19] P.M.Abdul-Muneer. Application of microsatellite marker in conservation genetics and fisheries management: recent advances in population structure analysis and conservation strategies. *Genetics Research International,* 2014:1-11, 2014.

[20] Daniel G. Hert, Christopher P. Fredlake and Annelise E. Barron. Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis,* 29(23):4618-4626, 2008.

[21] A J Sabat, A Budimir, D Nashev, R Sa-Leao, J M van Dijl, F Laurent, H Grundmann, A W Friedrich, on behalf of the ESCMID Study Group of Epidemiological Markers (ESGEM). Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Euro Surveill,* 18(4):p20380, 2013.

[22] Marie Suez, Abdelkader Behdenna, Sophie Brouillet, Paula Graca, Dominique Higuet and Guillaume Achaz. MicNeSs: genotyping microsatellite loci from a collection of (NGS) reads. *Molecular Ecology Resources,* 16(2):524-533, 2015.

[23] Moran P, Teel DJ, LaHood ES, Drake J and Kalinowski S. Standardising multi-laboratory microsatellite data in Pacific salmon: an historical view of the future. *Ecology of Freshwater Fish,* 15(4):597-605, 2006.

[24] Iria Fernandez-Silva, Jonathan Whitney, Benjamin Wainwright, Kimberly R. Andrews, Heather Ylitalo-Ward, Brian W. Bowen, Robert J. Toonen, Erica Goetze and Stephen A. Karl. Microsatellites for next-generation ecologists: a post-sequencing bioinformatics pipeline. *PLOS ONE,* 8(2):e55990, 2013.

[25] Emese Meglecz, Nicolas Pech, Andre Gilles, Vincent Dubut, Pascal Hingamp, Aurelie Trilles, Remi Grenier and Jean-Francois Martin. QDD version 3.1: a user-friendly computer program for microsatellite selection and primer design revisited: experimental validation of variables determining genotyping success rate. *Molecular Ecology Resources,* 14(6):1302-1313, 2014.

[26] John W. Fondon III, Andy Martin, Stephen Richards, Richard A. Gibbs and David Mittelman. Analysis of microsatellite variation in *Drosophila melanogaster* with population-scale genome sequencing. *PLoS ONE*, 7(3):e33036, 2012.

[27] Lucia R. Weinman, Joseph W. Solomon and Dustin R. Rubenstein. A comparison of single nucleotide polymorphism and microsatellite markers for analysis of parentage and kinship in a cooperatively breeding bird. *Molecular Ecology Resources,* 15(3):502-511, 2015.

[28] Christiaan Labuschagne, Lisa Nupen, Antoinette Kotze, Paul J. Grobler and Desire L. Dalton. Assessment of microsatellite and SNP markers for parentage assignment in ex situ African Penguin (*Spheniscus demersus*) populations. *Ecology and Evolution,* 5(19):4389-4399, 2015.

[29] Barkur S. Shastry. SNP alleles in human disease and evolution. *J Hum Genet,* 47(11):561-566, 2002.

[30] Martin Kreitman. Human genome variation: analysis, management and application of SNP data. *Pacific Symposium on Biocomputing,* 5:633-635, 2000.

[31] Phillip A. Morin, Gordon Luikart, Robert K. Wayne and the SNP workshop group. SNPs in ecology, evolution and conservation. *TRENDS in Ecology and Evolution,* 19(4):208-216, 2004.

[32] David Lopez Herraez, Holger Schafer, Jorn Mosner, Hans-Rudolf Fries and Michael Wink. Comparison of microsatellite and single nucleotide polymorphism markers for the genetic analysis of a Galloway cattle population. *BIOCIENCES,* 60(7-8):637-643, 2005.

[33] Tanya Y. Berger-Wolf, Saad I. Sheikh, Bhaskar DasGupta, Mary V. Ashley, Isabel C. Caballero, Wanpracha Chaovalitwongse and S. Lahari Putrevu. Reconstructing sibling relationships in wild populations. *Bioinformatics,* 23(13):i49-i56, 2007.

[34] C. Garke, F. Ytournel, B. Bed'hom, I. Gut, M. Lathrop, S. Weigend and H. Simianer. Comparison of SNPs and microsatellites for assessing the genetic structure of chicken populations. *Animal Genetics,* 43(4):419-428, 2011.

[35] Eric C. Anderson and John Carlos Garza. The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics,* 172(4):2567-2582, 2006.

[36] Francesco Emanuelli, Silvia Lorenzi, Lukasz Grzeskowiak, Valentina Catalano, Macro Stefanini, Michela Troggio, Sean Myles, Jose M Martinez-Zapater, Eva Zyprian, Flavia M Moreira and M Stella Grando. Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. *BMC Plant Biology,* 13:p39, 2013.

[37] Melony J Sellars, Leanne Dierens, Sean McWilliam, Bryce Little, Brian Murphy, Greg J Coman, William Barendse and John Henshall. Comparison of microsatellite and SNP DNA markers for pedigree assignment in Black Tiger shrimp, *Penaeus monodon. Aquaculture Research,* 45(3):417-426, 2014.

[38] J. E. Hess, A. P. Matala and S. R. Narum. 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, 2011.

[39] Jacquelin Defaveri, Heidi Viitaniemi, Erica Leder and Juha Merila. Characterizing genetic and nongenic molecular markers: comparsion of microsatellites and SNPs. *Molecular Ecology Resources,* 13(3):377-392, 2013.

[40] Nathan R. Campbell, Stephanie A. Harmon and Shawn R. Narum. Genotyping-in-Thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources,* 15(4):855-867, 2015.

[41] James E. Seeb, Carita E. Pascal, Ramesh Ramakrishnan and Lisa W. Seeb. SNP genotyping by the 5'-Nuclease Reaction: advances in high-throughput genotyping with nonmodel organisms. *Methods in Molecular Biology,* 578:277-292, 2009.

[42] Wolfgang Forstmeier, Holger Schielzeth, Jakob C. Mueller, Hans Ellegren and Bart Kempenaers. Heterozygosity-fitness correlations in zebra finches: microsatellites markers can be better than their reputation. *Molecular Ecology*, 21(13):3237-3249, 2012.

[43] X.Y. Hauge and M. Litt. A study of the origin of 'shadow bands' when typing dinucleotide repeat polymorphisms by the PCR. *Human Molecular Genetics,* 2(4):411-415, 1993.

[44] Vincent Murray, Chutima Monchawin and Phillip R. England. The determination of the sequences present in the shadow bands of a dinucleotide repeat PCR. *Nucleic Acids Research,* 21(10):2395-2398, 1993.

[45] J. Squirrell, P. M. Hollingsworth, M. Woodhead, J. Russell, A. J. Lowe, M. Gibby and W. Powell. How much effort is required to isolate nuclear microsatellites from plants? *Molecular Ecology,* 12(6):1339-1348, 2003.

[46] P. Sean Walsh, Nicola J. Fildes and Rebecca Reynolds. Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. *Nucleic Acids Research,* 24(14):2807-2812, 1996.

[47] Edwin Meijerink, Branko Kozulic, Gerald Stranzinger and Stefan Neuenschwander. Microsatellite allele sizing: difference between automated capillary electrophoresis and manual technique. *BioTechniques,* 31(4):810-818, 2001.

[48] Dennis L. Deemer and C. Dana Nelson. Standardized SSR allele naming and bining among projects. *BioTechniques,* 49(5):835-836, 2010.

[49] R Ekblom and J Galindo. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity,* 107(1):1-15, 2011.

[50] Juan E. Zalapa, Hugo Cuevas, Huayu Zhu, Shawn Steffan, Douglas Senalik, Eric Zeldin, Brent Mccown, Rebecca Harbut and Philipp Simon. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *American Journal of Botany,* 99(2):193-208, 2012.

[51] John W. Davey, Paul Hohenlohe, Paul D. Etter, Jason Q. Boone, Julian M. Catchen and Mark L. Blaxter. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Genetics,* 12(7):499-510, 2011.

[52] Christophe Van Neste, Filip Van Nieuwerburgh, David Van Hoofstat and Dieter Deforce. Forensic STR analysis using massive parallel sequencing. *Forensic Science International: Genetics,* 6(6):810-818, 2012.

[53] Monika Zavodna, Andrew Bagshaw, Rudiger Brauning and Neil J. Gemmell. The accuracy, feasibility and challenges of sequencing short tandem repeats using next-generation sequencing platforms. *PLoS ONE,* 9(12):e113862, 2014.

[54] Kimberly Robasky, Nathan E. Lewis and George M. Church. The role of replicates for error mitigation in next-generation sequencing. *NIH Public Access,* 15(1):56-62, 2014.

[55] M Ferrandiz-Rovira, T Bigot, D Allaine, M-P Callait-Cardinal and A Cohas. Large-scale genotyping of highly polymorphic loci by next-generation sequencing: how to overcome the challenges to reliably genotype individuals? *Heredity,* 114:485-493, 2015.

[56] Danielle Canceil, Enrique Viguera and S. Dusko Ehrlich. Replication slippage of different DNA polymerases is inversely related to their strand displacement efficiency. *THE JOURNAL OF BIOLOGICAL CHEMISTRY,* 274(39):27481-27490, 1999.

[57] J. I. Hoffman and W. Amos. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, 14(2):599-612, 2005.

[58] Gareth Highnam, Christopher Franck, Andy Matrtin, Calvin Stephens, Ashwin Puthige and David Mittleman. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Research,* 41(1):e32, 2013.

[59] Brant C. Faircloth. MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources,* 8(1):92-94, 2008.

[60] Kelly R. Ewen, Melanie Bahlo, Susan A. Treloar, Douglas F. Levinson, Bryan Mowry, John W. Barlow and Simon J. Foote. Identification and analysis of error types in high-throughput genotyping. *American journal of human genetics,* 67(3):727-736, 2000.

[61] Maureen A. Hess, James G. Rhydderch, Larry L. Leclair, Raymond M. Buckley, Mitsuhiro Kawase and Lorenz Hauser. Estimation of genotyping error rate from repeat genotyping, unintentional recaptures and know parent-offspring comparisons in 16 microsatellite loci for brown rockfish (*Sebastes auriculatus*). *Molecular Ecology Resources,* 12(6):1114-1123, 2012.

[62] Ke Hao, Cheng Li, Carsten Rosenow and Wing Hung Wong. Estimation of genotype error rate using samples with pedigree information-an application on the GeneChip Mapping 10K array. *Genomics,* 84(4):623-630, 2004.

[63] Paul C.D. Johnson and Daniel T. Haydon. Software for quantifying and simulating microsatellite genotyping error. *Bioinformatics and Biology Insights,* 1:71-75, 2007.

[64] Francois Pompanon, Aurelie Bonin, Eva Bellemain and Pierre Taberlet. Genotyping errors: causes, consequences and solutions. *Nature Reviews. Genetics,* 6(11):847-859, 2005.

[65] Yumeng Gao, Dabing Lu, Huan Ding and Poppy H. L. Lamberton. Detecting genotyping errors at *Schistosoma japonicum* microsatellites with pedigree information. *Parasites & Vectors,* 8(1):452, 2015.

[66] Markus Riester, Peter F. Stadler and Konstantin Klemm. FRANz: reconstruction of wild multi-generation pedigrees. *Genetics and population analysis,* 25(16):2134-2139, 2009.

[67] Stephanie Manel, Michael K. Schwartz, Gordon Luikart and Pierre Taberlet. Landscape genetics: combining landscape ecology and population genetics. *TRENDS in Ecology and Evolution,* 18(4):189-197, 2003.

[68] A Storfer, MA Murphy, JS Evans, CS Goldberg, S Robinson, SF Spear, R Dezzani, E Delmelle, L Vierling and LP Waits. Putting the ‘landscape’ in landscape genetics. *Heredity,* 98(3):128-142, 2007.

[69] Ian R. Bradbury, Lorraine C. Hamilton, Martha J. Robertson, Chuck E. Bourgeois, Atef Mansour and J. Brian Dempson. Lanscape structure and climatic variation determine Atlantic salmon genetic connectivity in the northwest Atlantic. *Canadian Journal of Fisheries and Aquatic Sciences,* 71(2):246-258, 2013.

[70] Gernot Segelbacher, Samuel A. Cushman, Bryan K. Epperson, Marie-Josee Fortin, Olivier Francois, Olivier J. Hardy, Rolf Holderegger, Pierre Taberlet, Lisette P. Waits and Stephanie Manel. Applications of landscape genetics in conservation biology: concepts and challenges. *Conservation Genetics,* 11(2):375-385, 2010.

[71] Ian R. Bradbury, Sophie Hubert, Brent Higgins, Tudor Borza, Sharen Bowman, Ian G. Paterson, Paul V. R. Snelgrove, Corey J. Morris, Robert S. Gregory, David C. Hardie, Jeffrey A. Hutchings, Daniel E. Ruzzante, Chris T. Taggart and Paul Bentzen. Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society B,* 277(1701):3725-3734, 2010.

[72] Rolf Holderegger, Urs Kamm and Felix Gugerli. Adaptive vs. neutral genetic diversity: implications for landscape genetics. *Landscape Ecology,* 21(6):797-807, 2006.

[73] Ana R. Amaral, Luciano B. Beheregaray, Kerstin Bilgmann, Dmitri Boutov, Luis Freitas, Kelly M. Robertson, Marina Sequeira, Karen A. Stockin, M. Manuela Coelho and Luciana M. Moller. Seascape genetics of a gloabally distributed, highly mobile marine mammal: the short-beaked common dolphin (*Genus Delphinus*). *PLoS ONE,* 7(2):e31482, 2012.

[74] Kimberly A. Selkoe, James R. Watson, Crow White, Tal Ben Horin, Matthew Iacchei, Satoshi Mitarai, David A. Siegel, Steven D. Gainess and Robert J. Toonen. Taking the chaos out of genetic patchiness: seascape genetics reveals ecological and oceanographic drivers of genetic patterns in three temperate reef species. *Molecular Ecology,* 19(17):3708-3726, 2010.

[75] Graham Coop, David Witonsky, Anna Di Rienzo and Jonathan K. Pritchard. Using environmental correlations to identify loci underlying local adaptation. *Genetics,* 185(4):1411-1423, 2010.

[76] Stephanie Manel, Felix Gugerli, Wilfried Thuiller, Nadir Alvarez, Pierre Legendre, Rolf Holderegger, Ludovic Gielly, Pierre Taberlet and Intra Bio Div Consortium. Broad-scale adaptive genetic variation in alpine plants is driven by temperature and precipitation. *Molecular Ecology,* 21(15):3729-3738, 2012.

[77] Christian Rellstab, Felix Gugerli, Andrew J. Eckert, Angela M. Hancock and Rolf Holderegger. A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology,* 24(17):4348-4370, 2015.

[78] Katie E. Lotterhos and Michael C. Whitlock. The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology,* 24(5):1031-1046, 2015.

[79] Pierre de Villemereuil and Oscar E. Gaggiotti. A new $F_{ST}$-based method to uncover local adaption using environmental variables. *Methods in Ecology and Evolution,* 6(11):1248-1258, 2015.

[80] Peter M. Vallone and John M. Butler. AutoDimer: a screening tool for primer-dimer and hairpin structures. *BioTechniques,* 37(2):226-231, 2004.

[81] Elfath M. Elnifro, Ahmed M. Ashshi, Robert J. Cooper and Paul E. Klapper. Multiplex PCR: optimization and application in diagnostic virology. *Clinical Microbiology Reviews,* 13(4):559, 2000.

[82] Matthew Kearse, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, Alex Cooper, Sidney Markowitz, Chris Duran, Tobias Thierer, Bruce Ashton, Peter Meintjes and Alexei Drummond. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics,* 28(12):1647-1649, 2012.

[83] Michael G. Gardner, Alison J. Fitch, Terry Bertozzi and Andrew J. Lowe. Rise of the machines – recommandations for ecologists when using next generation sequencing for microsatellite development. *Molecular Ecology Resources,* 11(6):1093-1101, 2011.

[84] US Atlantic sea scallop. Retrieved from https://www.msc.org/track-a-fishery/fisheries-in-the-program/certified/north-west-atlantic/us-atlantic-sea-scallop/. 2015.

[85] Atlantic deep-sea scallop, *Placopecten magellanicus.* Retrieved from http://www.geog.mcgill.ca/climatechange/ReportsMap/Placopecten%20magellanicus.pdf.

[86] Deborah R. Hart and Paul J. Rago. Long-term dynamics of U.S. Atlantic sea scallop *Placopecten magellanicus* populations. *North American Journal of Fisheries Management,* 26(2):490-501, 2006.

[87] Wendy Norden. Sea scallop *Placopecten magellanicus.* Seafood Watch, *Monterey Bay Aquarium*. 2012.

[88] Sarah R. Cooley, Jennie E. Rheuban, Deborah R. Hart, Victoria Luu, David M. Glover, Jonathan A. Hare and Scott C. Doney. An integrated assessment model for helping the United States sea scallop fishery plan ahead for ocean acidification and warming. *PLoS ONE,* 10(5):e0124145, 2015.

[89] Paul D. Etter, Jessica L. Preston, Susan Bassham, William A. Cresko and Eric A. Johnson. Local De Novo assembly of RAD paired-end contigs using short sequencing reads. *PLoS ONE,* 6(4):e185, 2011.

[90] Julian M. Catchen, Angel Amores, Paul Hohenlohe, William Cresko and John H. Postlethwait. *Stacks*: building and genotyping loci *de novo* from short-read sequences. *G3,* 1(3):171-182, 2011.

[91] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly and Pak C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics,* 81(3):559-575, 2007.

[92] Manuel A. R. Ferreira and Shaun M. Purcell. A multivariate test of association. *Bioinformatics,* 25(1):132-133, 2009.

[93] Francois Rousset. GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources,* 8(1):103-106, 2008.

[94] Klaus Hechenbichler, Klaus Schliep. Weighted k-nearest-neighbouring techniques and ordinal classification. *Universitätsbibliothek der LMU Muenchen*, 2004.

[95] Holger Schwender. Imputing missing genotypes with weighted k nearest neighbours. *Journal of Toxicology and Environmental Health,* 75(8-10):438-446, 2012.

[96] R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. R Core Team, 2015.

[97] Andy Beaumont. Genetic considerations in transfers and introductions of scallops. *Aquaculture International,* 8(6):493-512, 2000.

[98] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological measurement,* 20(1):27-46, 1960.

[99] Holger Schwender and with a contribution of Arno Fritsch. Scrime: analysis of high-dimensional categorical data such as SNP data. R package version 1.3.3. http://CRAN.R-project.org/package=scrime. 2013.

[100] Jonathan K. Pritchard, Matthew Stephens and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics,* 155(2):945-959, 2000.

[101] G. Evanno, S. Regnaut and J. Goudet. Detecting the number of clusters of individuals using software *STRUCTURE*: a simulation study. *Molecular Ecology,* 14(8):2611-2620, 2005.

[102] Dent A. Earl, Bridgett M. vonHoldt. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources,* 4(2):359-361, 2012.

[103] Leo Breiman. Random forests. *Machine learning,* 45(1):5-32, 2001.

[104] Trevor Hastie, Robert Tibshirani and Jerome Friedman. The elements of statistical leanring data mining, inference, and prediction. *New York: Springer,* 2009.

[105] Leo Breiman. Bagging predictors. *Machine Learning,* 24(2):123-140, 1996.

[106] Gerold Wefer, David Billet, Dierk Hebbeln, Bo Barker Jorgensen, Michael Schluter and Tjeerd C.E. Van Weering. Ocean margin systems. *Hanse Conference on Ocean Margin Systems. Delmenhorst, Germany.* 2000.

[107] Principles of genetic variation. Retrieved from http://garlandscience.com/res/pdf/ggm_ch04.pdf. 2014.

[108] Ulo Vali, Mikael Brandstrom, Malin Johansson and Hans Ellegren. Insertion-deletion polymorphisms (indels) as genetic markers in natural population. *BMC Genetics,* 9:8, 2008.

[109] NJ Schork, D Fallin and JS Lanchbury. Single nucleotide polymorphisms and the future of genetic epidemiology. *Clin Genet*, 58(4):250-264*,* 2000.

[110] Alain Vignal, Denis Milan, Magali SanCristobal and Andre Eggen. A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics, Selection, Evolution,* 34(3):275-305, 2002.

[111] Joel Ira Weller. Genomic selection in animals. *John Wiley & Sons,* 2016.

[112] Sebastian Junemann, Fritz Joachim Sedlazeck, Karola Prior, Andreas Albersmeier, Uwe John, Jorn Kalinowski, Alexander Mellmann, Alexander Goesmann, Arndt von Haeseler, Jens Stoye and Dag Harmsen. Updating benchtop sequencing performance comparison. *Nature Biotechnology,* 31(4):294-296, 2013.

[113] Jun Zhang, Rod Chiodini, Ahmed Badr and Genfa Zhang. The impact of next-generation sequencing on genomes. *J Genet Genomics,* 38(3):95-109, 2011.

[114] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu and Maggie Law. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology,* 2012: 1-11, 2012.

[115] Rana W. El-Sabaawi, Michael C. Marshall, Ronald D. Bassar, Andres Lopez-Sepulcre, Eric P. Palkovacs and Christopher Dalton. Assessing the effects of guppy life history evolution on nutrient recycling: from experiments to the field. *Freshwater Biology,* 60(3):590-601, 2015.

[116] Ronald D. Bassar, Thomas Heatherly Π, Michael C. Marshall, Steven A. Thomas, Alexander S. Flecker and David N. Reznick. Population size structure dependent fitness and ecosystem consequences in Trinidadian guppies. *Journal of Animal Ecology,* 84(4):955-968, 2015.

[117] Monica G. Turner and Robert H. Gardner. Lanscape ecology in theory and practice pattern and process. *Springer eBooks,* 2015.

[118] S. Manel, B. N. Poncet, P. Legendre, F. Gugerlis and R. Holdereggers. Common factors drive adaptive genetic variation at different spatial scales in *Arabis alpine. Molecular Ecology,* 19(17):3824-3835, 2010.

[119] Deborah Zulliger, Elvira Schnyder and Felix Gugerli. Are adaptive loci transferable across genomes of related species? Outlier and environmental association analyses in Alpine Brassicaceae species. *Molecular Ecology,* 22(6):1626-1639, 2013.

[120] Eric Frichot, Sean D. Schoville, Guillaume Bouchard and Olivier Francois. Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution,* 30(7):1687-1699, 2013.

[121] Cooper and Necia Grant. The Human genome project: desciphering the blueprint of heredity. *Mill Valley, Calif. : University Science Books*, 1994.

[122] Thomas Mitchell-Olds, John H. Willis and David B. Goldstein. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics,* 8(11):845-856, 2007.

[123] Sewall Wright. The genetical structure of populations. *Annals of Human Genetics,* 15(4):323-354, 1951.

[124] Hadley Wickham. ggplot2: elegant graphics for data analysis. *Springer-Verlag New York,* 2009.

[125] Brian and Sandy Kinghorm. Pedigree Viewer. Retrieved from http://bkinghor.une.edu.au/pedigree.htm. 2015.

[126] Teri A. Reynolds and David L. Schriger. The conduct and reporting of meta-analysis of studies of diagnostic tests, and a consideration of ROC curves. *Annals of Emergency Medicine,* 55(6):570-577, 2010.

[127] Aravindh Krishnamoorthy and Deepak Menon. Matrix inversion using Cholesky decomposition. *arXiv Preprint arXiv:1111.4144*, 2011.

[128] Vikas Bansal, Olivier Harismendy, Ryan Tewhey, Sarah S. Murray, Nicholas J. Schork, Eric J. Topol and Kelly A. Frazer. Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Research,* 20(4):537-545, 2010.

[129] Julian Catchen, Paul A. Hohenlohe, Susan Bassham and Angel Amores.Stacks: an analysis tool set for population genomics. *Molecular Ecology,* 22(11):3124-3140, 2013.

[130] Allan J. Baker. Molecular methods in ecology. *Oxford; Malden, MA, USA: Blankwell Science,* 2000.

[131] John Carlos Garza and Nelson B. Freimer. Homoplasy for size at microsatellite loci in Humans and Chimpanzees. *Genome Research,* 6(3):211-217, 1996.

[132] Richard Shen, Jianbing Fan, Derek Campbell, Weihua Chang, Jing Chen, Dennis Doucet, Jo Yeakley, Marina Bibikova, Eliza Wickham Garcia, Celeste McBride, Frank Steemers, Francisco Garcia, Bahram G. Kermani, Kevin Gunderson and Arnold Oliphant. High-throughput SNP genotyping on universal bead arrays. *Mutation Research,* 573(1-2):70-82, 2005.

[133] Mattias Jakobsson and Noah A. Rosenberg. CLUMPP: a cluster matching and permutatioin program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics,* 23(14):1801-1806, 2007.

[134] Noah A. Rosenberg. Distruct: a program for the graphical display of population structure. *Molecular Ecology Notes,* 4(1):137-138, 2004.

[135] H. John B. Birks, Andre F. Lotter, Steve Juggins and John P. Smol. Tracking environmental change using lake sediments. *Springer eBook,* 2012.

[136] Oded Maimon and Lior Rokach. Data mining and knowledge discovery handbook. *Springer New York Dordrecht Heidelberg London,* 2010.

[137] Mallory Van Wyngaarden, Paul V. R. Snelgrove, Claudio DiBacco, Lorraine C. Hamilton, Naiara Rodriguez-Ezpeleta, Ryan R. E. Stanley, Ian R. Bradbury. Identifying patterns of dispersal, connectivity, and selection in the sea scallop, *Placopecten magellanicus*, using clines in RAD-seq derived SNPs. In review, 2016.

# APPENDIX A Flowchart Of The Algorithm That MEGASAT Uses to Trim Off Microsatellite Primers

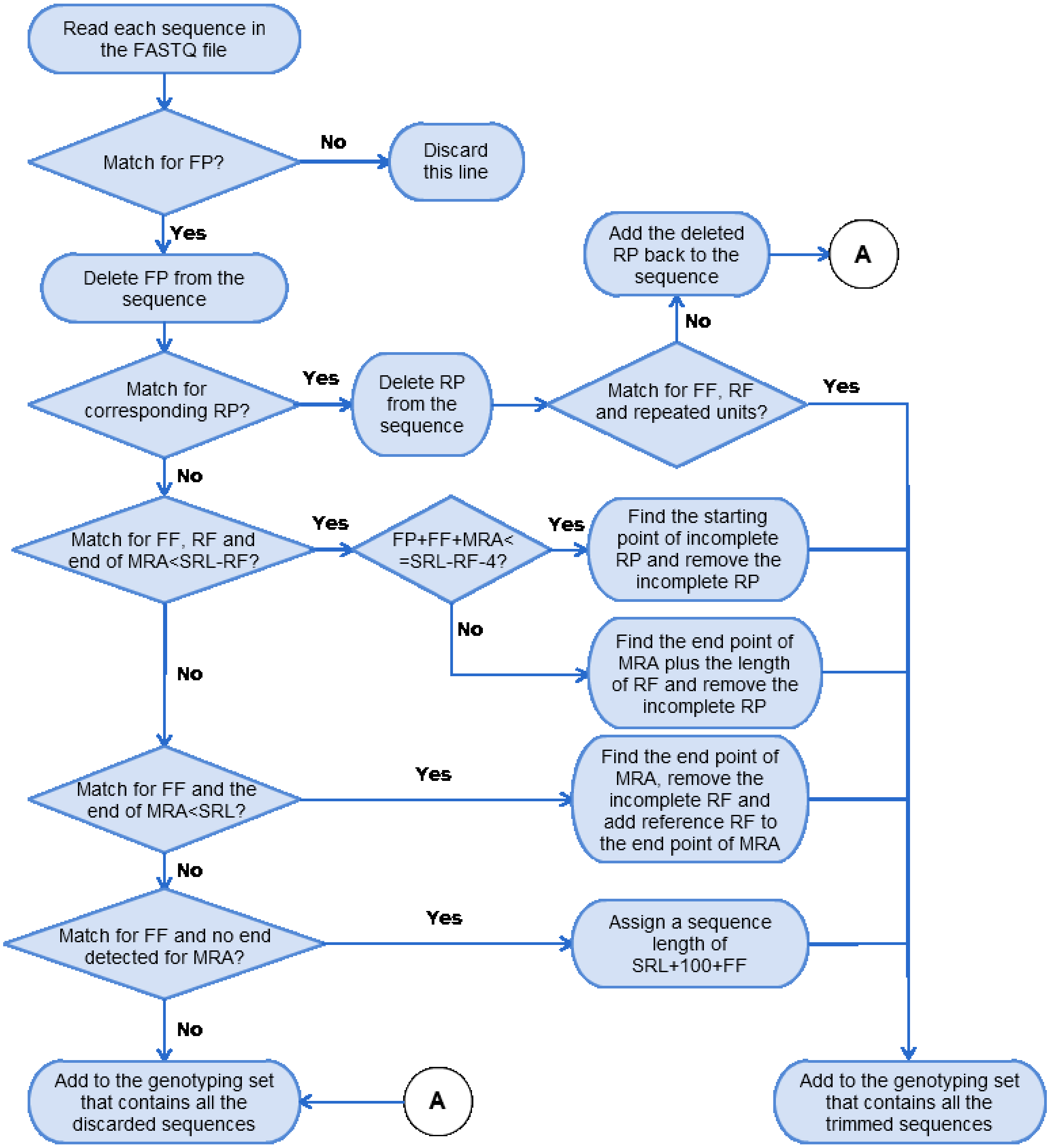

Read each sequence in the FASTQ file
Match for FP?
No
Discard this line
Yes
Delete FP from the sequence
Match for corresponding RP?
Yes
Delete RP from the sequence
Match for FF, RF and repeated units?
No
Add the deleted RP back to the sequence
A
Yes
No
Match for FF, RF and end of MRA<SRL-RF?
Yes
FP+FF+MRA< =SRL-RF-4?
Yes
Find the starting point of incomplete RP and remove the incomplete RP
No
Find the end point of MRA plus the length of RF and remove the incomplete RP
No
Match for FF and the end of MRA<SRL?
Yes
Find the end point of MRA, remove the incomplete RF and add reference RF to the end point of MRA
No
Match for FF and no end detected for MRA?
Yes
Assign a sequence length of SRL+100+FF
No
Add to the genotyping set that contains all the discarded sequences
A
Add to the genotyping set that contains all the trimmed sequences

Figure A.1 Flowchart of the algorithm that MEGASAT uses to trim off microsatellite primers. The abbreviations for microsatellite amplicon components are the same as those given in Figure 5.

# APPENDIX B Laboratory Method For Generating Validation Data Set

The following protocols were carried out in the laboratories of Dr. Paul Bentzen (Department of Biology, Dalhousie University) by members of these labs.

## B.1 First Multiplex PCR

Initial multiplex PCR reactions were set up using 10 loci per multiplex. Loci were untested, thus the only criterion used while creating the mixes was absence of dimers based on the AUTODIMER output. PCRs were performed in 3.5ul volumes using Qiagen (Venlo, Netherlands) Type-IT 2x Mastermix (1.75ul), 0.2uM each oligo (20 oligos per reaction) and 0.7ul genomic DNA estimated to be ~275pg. 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 30s, 57°C, 180s, 72°C 60s, with a final extension at 68°C for 30min. Each multiplex was run versus 12 individuals from the Guanapo River, Trinidad, West Indies. Multiplexed PCRs were then pooled per individual (i.e. all reactions from one guppy were pooled together), resulting in 12 pools. Figure B.1 shows the process of first multiplex PCR.

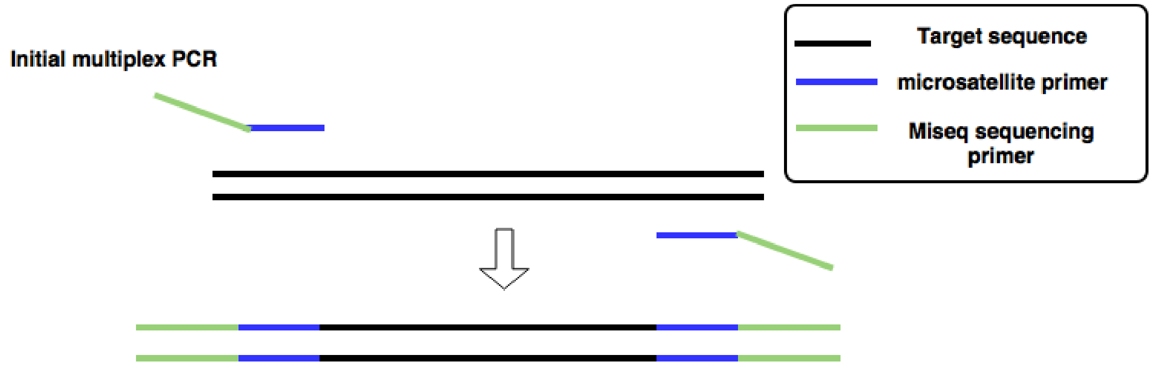


Figure B.1. Overview of the initial multiplex PCR

## B.2 Second Multiplex PCR

Illumina annealing and indexing sequences were added to the PCR products using a second PCR (as shown in Figure B.2). The previous amplicons were diluted 500x and used as template for an indexing-PCR which included oligos composed of the Illumina annealing adapter sequences, a 6bp index (barcode) and the Illumina sequencing primers. By taking advantage of the Illumina dual-indexing capability, we differentiate 1024 individuals in a single sequencing run using a set of 32 Index_1 oligos and a set of 32 Index_2 oligos.  Indexing PCRs were performed in 5ul total volume with 0.25U Taq DNA polymerase (New England Biolabs, Ipswich MA, USA), 0.5ul Thermopol 10x buffer (NEB), 0.2mM each dNTP, 0.2uM Index_1 oligo, 0.2uM Index_2 oligo and 1ul of 500-fold diluted PCR product pooled above. Cycling parameters were: 95°C 2m, followed by 18 cycles of 95°C 20s, 60°C 60s, 72°C 60s with a final extension at 72°C for 10 min.

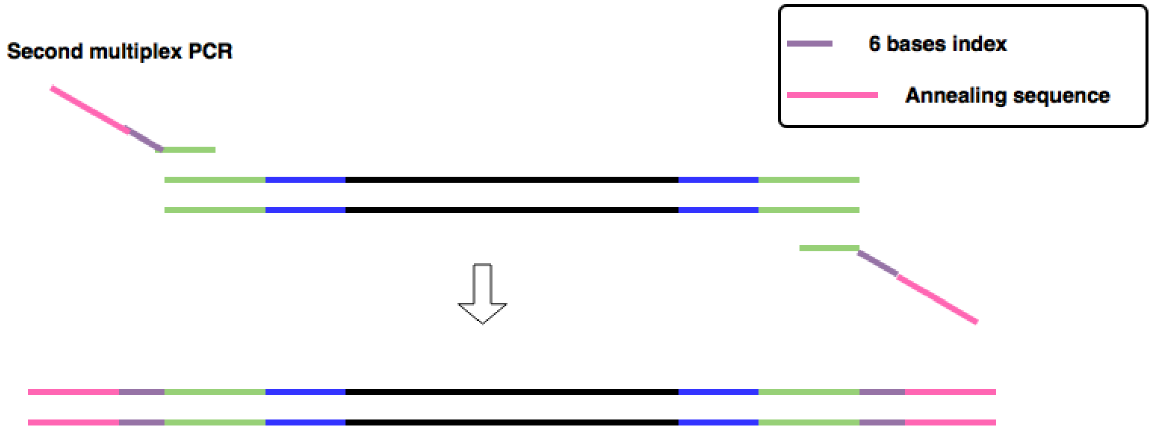


Figure B.2 Overview of the second multiplex PCR.

## B.3 Sequencing

The indexed PCR products were pooled and cleaned using Ampure XP (Beckman Coulter, Pasedena CA, USA) magnetic beads (1.8:1 bead: DNA library ratio). The clean library was quantified using Kapa (Wilmington MA, USA) Library Quantification for

Illumina on a Roche (Basel, Switzerland) LC480 qPCR instrument using manufacturers' protocols. This library was sequenced at 10pM concentration using Miseq v2 chemistry, 150x150 dual index read. Dual-indexed individuals were demultiplexed with the Miseq Sequence Analysis software and then separate FASTQ files were generated for each individual.

Prior to developing MEGASAT, we used GENEIOUS r7 software [82] 'separate by barcode' function to demultiplex loci within an individual and microsatellite genotypes were scored using the depth histograms generated within GENEIOUS. Once MEGASAT was working, we used GENEIOUS to examine reads and verify the performance of MEGASAT. As more data were collected, we refined the laboratory process. From the initial 468 loci tested in 10-plexes, we chose 80 loci and proceeded using 20-plex reactions. As our microsatellite data set grew larger (i.e. as we learned more about each locus), we dropped loci that had low information content, high error rates, evidence of nulls or inability to multiplex well. For long-term data collection, we settled on 43 loci that we multiplex in a single PCR reaction, using the same reaction conditions as our initial PCRs (above). This reaction is diluted 10-fold by adding water directly to the completed PCR plate and 0.3ul of the diluted product is used as template in the indexing reaction.

# APPENDIX C Laboratory Methods For Generating Genetic And Environmental Data

The following protocols were carried out in the laboratories of Dr. Ian Bradbury (Department of Fisheries and Oceans) by members of these labs.

## C.1 Genetic Data

Tissue samples were collected and preserved in AllProtect (Qiagen, Toronto, ON, Canada) or 80% ethanol. DNA extraction and library preparation were performed at the Aquatic Biotechnology Lab at the Bedford Institute of Oceanography in Halifax, Nova Scotia. DNA was isolated from the tissue samples using DNeasy Blood and Tissue kit or DNeasy 96 Blood and Tissue kit (Qiagen) following the manufacturer's protocol, including the optional RNase A treatment. All 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). All samples were normalized to 25ng/µL. The DNA quality for all samples was verified by agarose gel electrophoresis of 100 ng of extracted DNA. DNA was visualized using SYBR Safe (Life Technologies) and documented using a Gel Logic 200 (Kodak).

RAD-seq libraries were prepared as described in [89] with modifications. DNA samples from 22 individuals from the same geographical location comprised each library (with the exception of the library for SUN which consisted of only 20 individuals) with a different in-line barcode in the P1 adapter for each individual sample. Gel size selection performed after sonication and PCR amplification was done on a Pippin Prep (Sage Science, Beverly, MA, USA) using the 2% agarose gel cassette with ethidium bromide (Sage Science) and size selection range of 300-500bp. PCR amplification used Q5 Hot Start Master Mix (NEB, Whitby, ON, Canada) for all libraries. Amplification cycles for all libraries were 98 °C for 30 seconds; *x* cycles of 98 °C for 30 seconds, 65 °C for 30 seconds, 72 °C for 30 seconds; 1 cycle of 72 °C for 5 minutes, where *x* was 18 for all libraries except for SSB, GEO, and SUN where *x* was 13. All libraries were sequenced on a HiSeq 2000 (Illumina) as 100bp paired end sequences with one library per lane. Sequencing was performed at the McGill University and Génome Québec Innovation Centre, Montréal, Canada.

## C.2 Environmental Data

A bounding box of 1 square degree around each sample site was used to collect data for each sampling location and values were averaged within the bounding box to create site-specific averages for each data type. Data from all sources were combined to create the final dataset used in the analysis. Data were separated into surface and depth values based on the collection site depth for each sampling location.

Data validation and preparation were completed using R (R Development Core Team 2012). We removed outliers using z-score analysis. Because of natural season variation in the data, z-scores were calculated for each variable for each sample site per each month and outliers were removed, where necessary. Variables with missing data in more than six sites were removed from subsequent analyses. For the remaining variables with missing data, single imputation using neighbouring sites was used to estimate missing values (sites arranged by latitude, mean of the sites directly north and south of the missing site). Following outlier removal and imputation, we standardized data by subtracting the mean and dividing by the standard deviation. We then calculated site-specific maximum and minimum values as well as seasonal averages for each variable. Winter included January, February, and March, Spring included April, May, and June, Summer included July, August, and September, and Fall included October, November, and December. The final dataset contained 90 variables spanning all available data types.