COMPUTATIONAL METHODS FOR EFFICIENT PROCESSING AND ANALYSIS OF SHORT-READ NEXT-GENERATION DNA SEQUENCING DATA

by

Praveen Nadukkalam Ravindran

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
March 2020

I dedicate this thesis to Suria for all the love and motivation and to my parents and brother for their endless support.

Table of Contents

List of Tables

List of Figures

Abstract

DNA sequencing has transformed the discipline of population genetics, which seeks to assess the level of genetic diversity within species or populations, and infer the geographic and temporal distributions between members of a population. Restriction-site associated DNA sequencing (RADSeq) is a Next-generation sequencing (NGS) technique, which produce data that consists of relatively short (typically 50 to 300 nucleotide) fragments or "reads" of sequenced DNA and enables large-scale analysis of individuals and populations. In this thesis, we describe computational methods, which use graph-based structures to represent these short reads obtained and to capture the relationships among them.

A key challenge in RADSeq analysis is to identify optimal parameter settings for assignment of reads to loci (*singular*: Locus), which correspond to specific regions in the genome. The parameter sweep is computationally intensive, as the entire analysis needs to be run for each parameter set. We propose a graph-based structure (RADProc), which provides persistence and eliminates redundancy to enable parameter sweeps. For 20 green crab samples and 32 different parameter sets, RADProc took only 2.5 hours while the widely used Stacks software took 78 hours.

Another challenge is to identify paralogs, sequences that are highly similar due to recent duplication events, but occur in different regions of the genome and should not to be merged into the same locus. We introduce PMERGE, which identifies paralogs by clustering the catalog locus consensus sequences based on similarity. PMERGE is built on the fact that paralogs may be wrongly merged into a single locus in some but not all samples. PMERGE identified 62%-87% of paralogs in the Atlantic salmon and green crab datasets.

Gene flow is the movement of *alleles*, specific sequence variants at a given locus, between populations and is an important indicator of population mixing that changes genetic diversity within the populations. We use the RADProc graph to infer gene flow among populations using allele frequency differences in exclusively shared alleles in each pair of populations. The method successfully inferred gene flow patterns in simulated datasets and provided insights into reasons for observed hybridization at two locations in a green crab dataset.

List of Abbreviations Used

GBS	Genotype By Sequencing
MID	Molecular Identifier
MYA	Million Years Ago
NGS	Next-generation sequencing
PSV	Paralogous Sequence Variants
RADSeq	Restriction site Associated DNA Sequencing
SNP	Single Nucleotide Polymorphism

Glossary

allele An allele is a variant of a gene or locus, there could be more than one allele for a given gene or locus.

barcode A short unique sequence (typically 6–12bp) used to identify individual samples. Inline barcodes occur on the end of the adapter (short synthetic sequence ligated to a DNA molecule during sequencing) that is immediately adjacent to the genomic DNA fragment after adapter ligation. The barcode is sequenced immediately prior to sequencing of the DNA fragment, and thus the barcode sequence will appear at the beginning of sequence reads

contig A group of overlapping sequence reads assembled to form a longer sequence

depth of coverage The number of sequence reads obtained from a given locus or nucleotide site in a DNA sequencing experiment.

filtering Removing unwanted sequence reads from a dataset due to low sequence quality, low depth of coverage, evidence for paralogy, or other reasons

gene A gene is a sequence of nucleotides in DNA that determines a certain trait.

locus A specific region in the genome

Next-Generation Sequencing	Technologies first emerging around 2005 that sequence millions of DNA molecules simultaneously
paired-end sequences	Illumina sequencing of both ends of each DNA fragment
reduced-representation	DNA library comprised of a subset of loci, rather than the entire genome
single-end sequencing	Illumina sequencing of only one end of each DNA fragment

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. Robert Beiko for all his support, patience, guidance and enthusiasm. My sincere thanks also goes to my co-supervisor Dr. Ian R. Bradbury for his direction and questions, and my committee members Dr. Norbert Zeh and Dr. Nauzer Kalyaniwalla for their valuable inputs and time. Special thanks to Dr. Paul Bentzen for his guidance. I would also like to thank Natural Sciences and Engineering Research Council of Canada for funding. Last but not least, I would like to thank all current and past members of Beiko Lab and Blouin Lab for all the fun activities, laughs and support that made my PhD enjoyable.

Chapter 1

Introduction

DNA sequencing is the process of determining the sequence of nucleotide bases (As, Ts, Cs, and Gs) in DNA fragments. DNA sequencing has many applications in different areas such as forensics, medicine and agriculture. One of its important applications is in the field of population genetics. The evolutionary processes of mutation, migration, genetic drift, and natural selection shape patterns of genetic variation among individuals, populations, and species, and they can do so differentially across genomes. Therefore, in population genetics these mechanisms and their interactions and evolutionary consequences are investigated by building mathematical models, developing statistical methods for inferring parameters of ancestral processes, and testing hypotheses based on the analysis of real data [19]. Population genetics involves studying changes in the frequencies of genetic variation in populations to infer population structure or population subdivision in space and time (e.g. [16]).

Genetic variations are the differences in homologous DNA segments or genes or loci (*singular:* locus) between individuals within and among populations and each variation of a gene is called an allele (Figure 1.1a). Homologous genes or loci are two or more genes that descend from a common ancestry and have high levels of sequence similarity. Genetic variations could be found in orthologous loci and paralogous loci. Orthologous locus represent homologous sequences that descended from the same ancestral sequence and separated by speciation event. Paralogous locus represent homologous sequences that are separated by a duplication event. Duplication can occur in any region of a genome, and the resulting duplicated genes can be retained or lost in the population during evolution. Genetic variations can arise as a result of mutation, random mating, random fertilization or recombination events. Mutations are considered as the original source of genetic variation, which results in permanent alteration of DNA sequences due to different events such as substitution, insertion, deletion and duplication [73]. A genetic variation is usually considered as DNA polymorphism if it appears with a 1% or higher frequency in a population [15, 62] and can be a single base or thousands of bases found throughout the genome and may or may not have phenotypic effects. DNA polymorphism with single nucleotide difference in a DNA sequence (Figure 1.1a) that occurs in a significant proportion in a population is known as Single Nucleotide Polymorphism (SNP).The abundance of SNPs and the ease with which they can be measured make these genetic variations significant.

Reconstructing orthologous loci from the short reads is a necessary step in the inference of SNPs. Orthologous loci can be built from the short reads by aligning (mapping) them to a reference genome that serves as a representative sequence database for a species. Numerous tools and algorithms are available to perform sequence alignments to a reference genome (e.g., Burrows-Wheeler Alignment [67], Bowtie [65]). But not all organisms have high-quality reference genomes readily available, such as non-model organisms that are not usually studied extensively because they are hard to investigate. In such cases, we have to rely on identifying loci and alleles *de novo*. In the *de novo*-based methods, DNA sequences with a minimum threshold level of similarity are identified and grouped as presumed alternative alleles of an orthologous locus [102]. Identification of orthologous loci and true SNPs from short-read sequences remains especially challenging for species with a duplicated genome due to the difficulties in distinguishing between duplicated genes. The single nucleotide differences between duplicated loci in the genome are known as Paralogous Sequence Variants (PSV) (Figure 1.1b) [49]. Since the quality of the identified SNPs can impact the downstream population genetic analysis, it is important to differentiate SNPs and

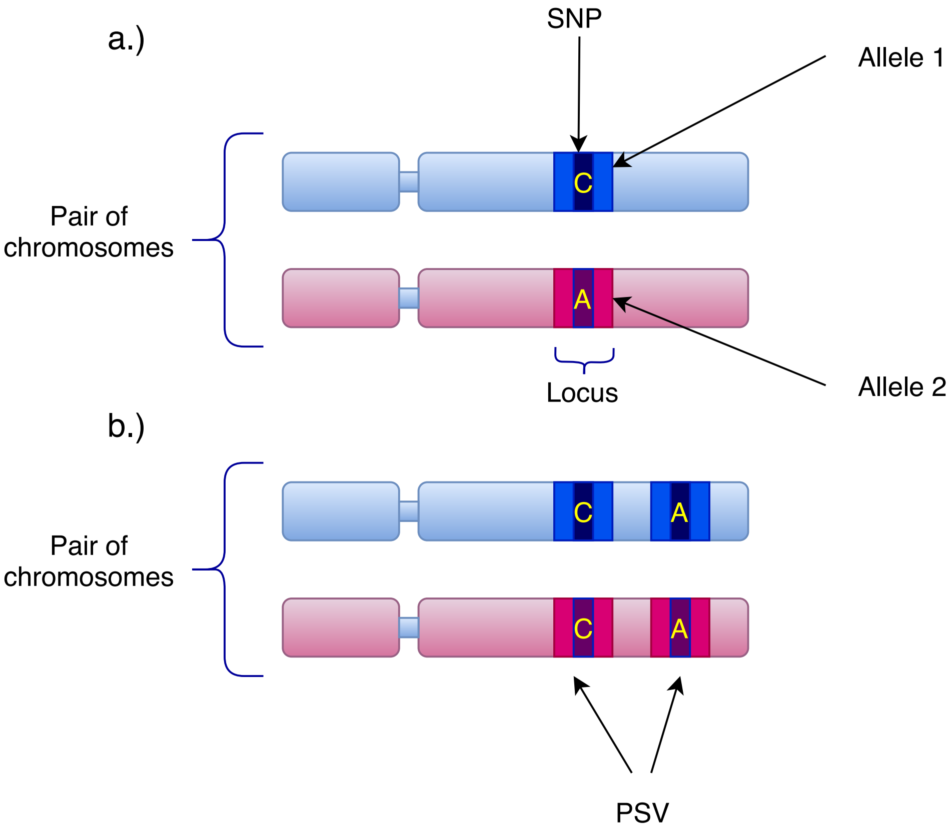

Figure 1.1: SNP versus PSV. a.) A Single Nucleotide Polymorphism (C/A) representing two alleles of locus b.) Two duplications with a variant (C/A). The duplicated loci may be erroneously clustered as a single SNP.

PSVs. However, given the large number of short-read DNA sequences often examined per individual, assembling the sequences into orthologous loci is a crucial and often challenging step in analyzing these genomic datasets.

1.1 Restriction site Associated DNA sequencing (RADSeq)

Population-genomic studies survey hundreds to thousands of genetic variations or genetic markers to describe genome-wide variation and make population-wide inferences [50, 9, 14]. Obtaining whole genome sequences from a large number of individuals is expensive and not necessary in many cases [87]. The development of Next-Generation Sequencing (NGS) [94] technologies has provided an alternative to whole genome sequencing by examining a small percentage of the target genome. NGS technologies have led to an increase in the size and number of population-genomic studies even in non-model organisms for which few or no genomic resources presently exist [16]. There are several short-read next-generation DNA sequencing methods [25], which rely on reduced-representation sequencing approaches (e.g. [5, 54, 93]). Reduced representation sequencing is a cost-effective method for sequencing a large number of genome-wide loci across multiple individuals.

Restriction site Associated DNA sequencing (RADSeq) [7] is a reduced-representation sequencing approach that can be applied to both model and non-model species. RADSeq is a next-generation DNA sequencing-based genotyping method (e.g. [43]), which allows sampling the genomes of multiple individuals in a population and identifying and genotyping SNPs simultaneously. RADSeq involves the cutting of a template genome with a specific enzyme called a restriction endonuclease, followed by mechanical shearing and molecular biological processing steps. The resulting libraries are then sequenced using a next-generation DNA sequencing platform. RADSeq is increasingly being used in evolutionary and quantitative genomic analyses [60] including genome-wide association [89], phylogenetic [105], and landscape genetic studies [3].

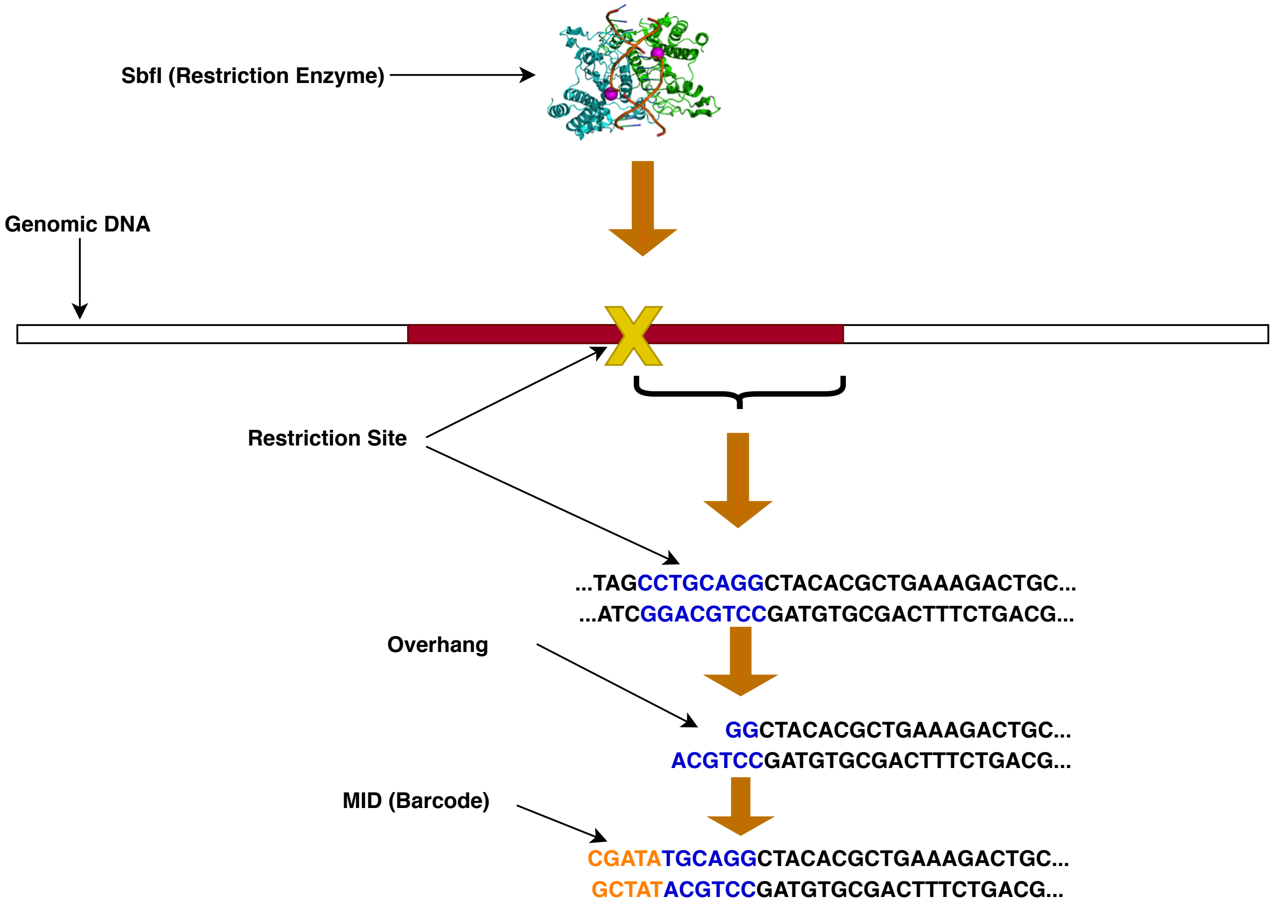
SbfI (Restriction Enzyme)
Genomic DNA
Restriction Site
...TAGCCTGCAGGCTACACGCTGAAAGACTGC...
...ATCGGACGTCCGATGTGCGACTTTCTGACG...
Overhang
GGCTACACGCTGAAAGACTGC...
ACGTCCGATGTGCGACTTTCTGACG...
MID (Barcode)
CGATATGCAGGCTACACGCTGAAAGACTGC...
GCTATACGTCCGATGTGCGACTTTCTGACG...

Figure 1.2: Restriction-Site Associated DNA sequencing (RADSeq) using restriction enzyme SbfI, which has a recognition site (CCTGCAGG). The restriction site is identified and the flanking DNA sequences (Red region) are sheared, which results in cut site overhang.The barcode or molecular identifier (MID; CGATA) is then attached to the overhang, following which the DNA is sequenced using a DNA sequencing platform. Restriction enzyme image accessed from https://daily.jstor.org/ on 2019-10-02.

RADSeq (Figure 1.2) focuses on the flanking DNA sequences around each restriction site known as RAD-tags, to generate a reduced representative library [25]. The technique involves cutting the target genome with a particular restriction enzyme (e.g., SbfI), and as each restriction enzyme has a specific recognition site (e.g., the DNA motif CCTGCAGG for SbfI) the frequency of this site across the genome determines the number of RAD-tags obtained [79]. In practice, the RAD-tags are attached to a molecular identifier (Molecular Identifier (MID)), a short, synthetic DNA sequence frequently referred to as a "barcode", which allows the simultaneous sequencing of different DNA samples [51]. The obtained sequences are then assigned back to their corresponding samples using the same set of barcodes used during sequencing (demultiplexing) (Figure 1.3). RADSeq methods can differ in the details of the experimental procedure used [6] (e.g., 2bRAD [104, 44] ,Genotype By Sequencing (GBS) [36], CRoPS [103], ddRAD [84], ezRAD [101]).

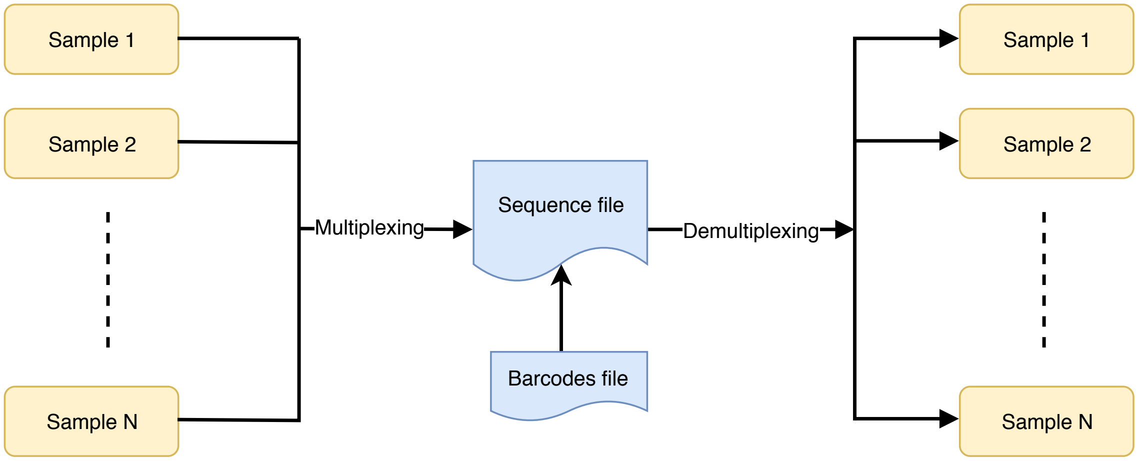

Figure 1.3: Representation of multiplexing and demultiplexing samples using barcodes.

1.2 RADSeq analysis tools/pipelines

Currently there exist many software programs to process RADSeq data such as Stacks [17], pyRAD [33] and AftrRAD [97]. Stacks is the most widely used software package for RADSeq analysis and consists of modules to perform all tasks from quality filtering, *de novo* or reference-aligned locus identification, genotyping and generating population-genetic statistics. pyRAD is designed particularly for phylogenetic applications. Like Stacks, pyRAD and AftrRAD also can perform quality filtering, *de novo* locus formation and genotyping and can also handle insertion and deletion variations among alleles. In all these programs, the basic idea is to identify SNPs by determining unique stacks of reads with a minimum depth of coverage and aligned as allelic pairs of a candidate locus (*de novo* locus formation) if they satisfy a predefined percentage identity or maximum nucleotide distance. The method used to perform pairwise comparisons among the unique stacks to determine potential allelic pairs differs in each program. For the purpose of describing the general process of obtaining population genetics statistics from the raw RADSeq data (Figure 1.4) we are using the Stacks pipeline.

1.3 *De novo* locus formation

De novo locus formation is the reconstruction of genomic regions from the raw reads obtained from each sample/individual. The process involves identifying the reads that are from a specific genomic region based on certain criteria. The two important criteria are the minimum coverage depth and the maximum allowed nucleotide distance, which control the reconstruction of loci from the short reads. Exactly matching reads with minimum coverage depth m or greater form unique stacks; these stacks can be thought of as representing alleles, although some of them will represent errors (Figure 1.5). Any stack that does not meet this depth threshold is kept aside for

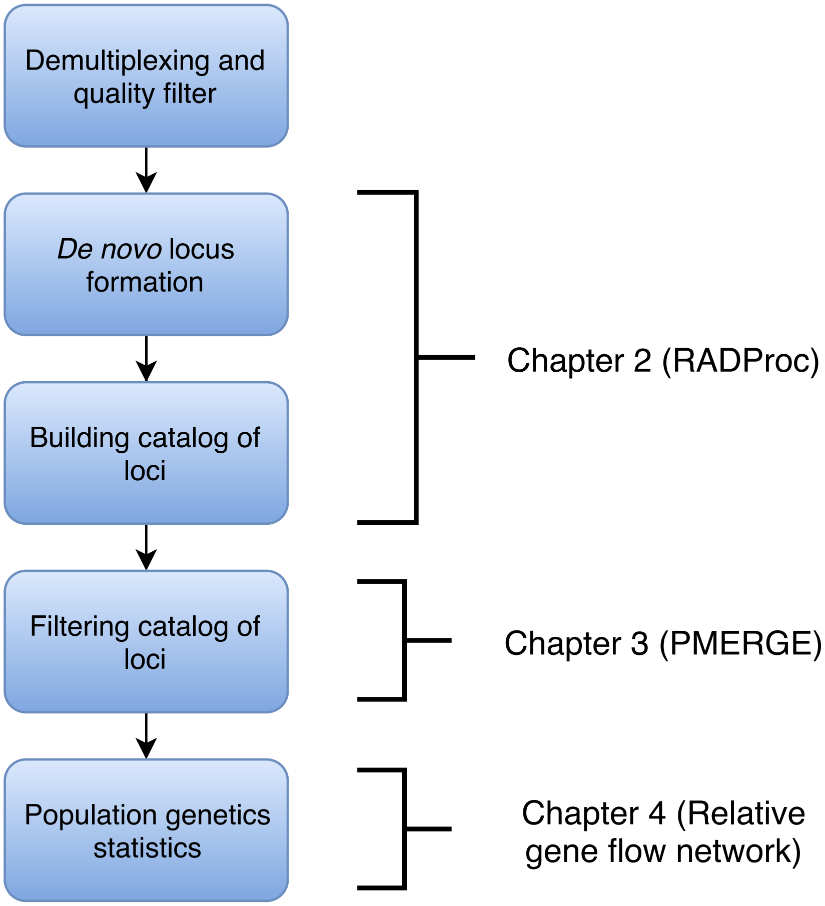
Demultiplexing and quality filter
De novo locus formation
Building catalog of loci
Filtering catalog of loci
Population genetics statistics
Chapter 2 (RADProc)
Chapter 3 (PMERGE)
Chapter 4 (Relative gene flow network)

Figure 1.4: Typical steps involved in processing RADSeq data, from raw reads to generate population genetics statistics and the corresponding position of thesis chapters in the workflow.

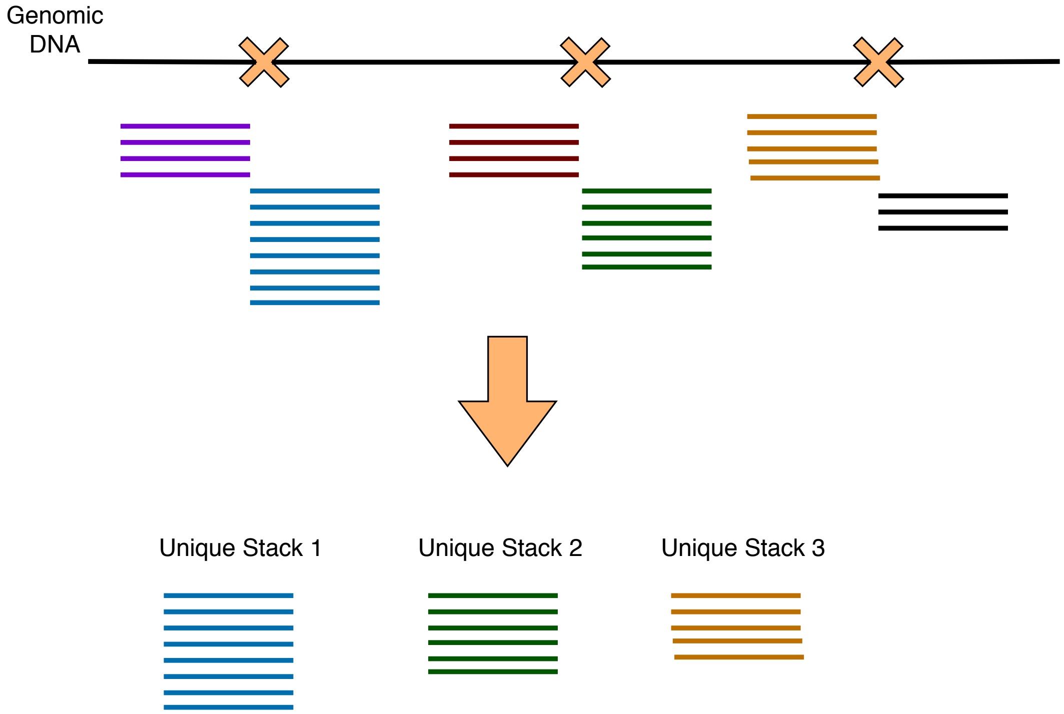

Figure 1.5: Illustration of reads sequenced at a cut site and identification of unique stacks from the raw reads obtained from sample 1 for minimum coverage depth m=5.

adding to a locus later in the process and an allele is not formed. The reads that form a unique stack are termed as primary reads and the ones that are kept aside are known as secondary reads. If m is set very low, small numbers of reads with identical sequencing errors may be erroneously labeled as stacks. If m is set very high, then true alleles may be dropped from forming stacks, as they do not occur frequently enough in the sequenced DNA. After forming the unique stacks, the next stage is to merge putative alleles into an orthologous locus.

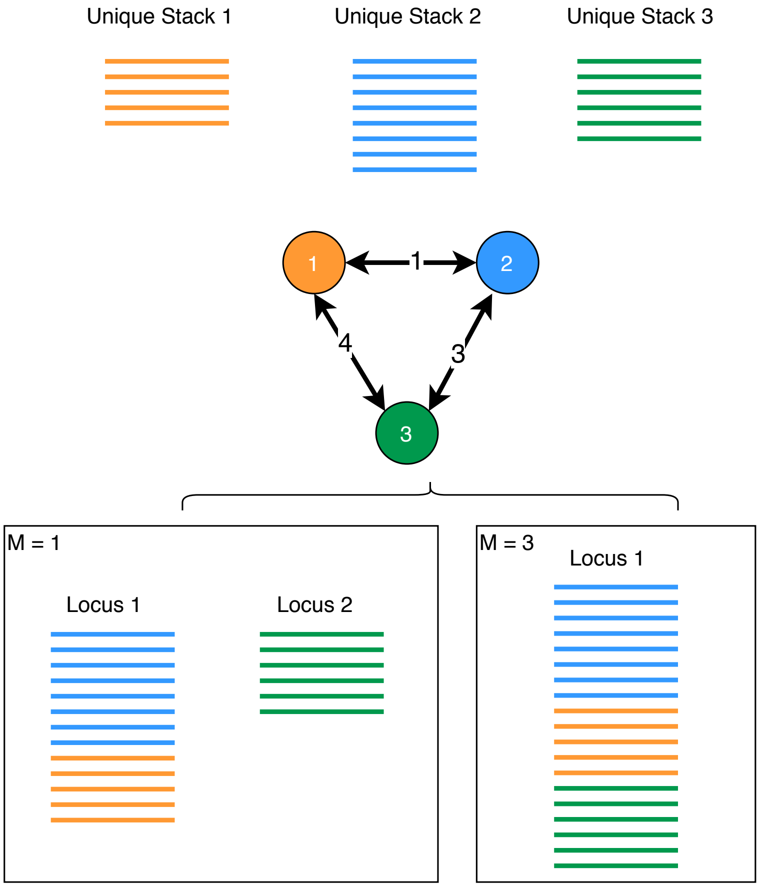

Figure 1.6: Determining the unique stacks within a given maximum allowed nucleotide distance to merge into locus. When M =1, unique stack 1 and unique stack 2 are merged into locus 1, where as unique stack 3 forms locus 2 with only one allele. But, when M = 3, all three unique stacks are merged into one locus.

The unique stacks within an allowed nucleotide distance (M) are then merged to form the *de novo* locus (Figure 1.6). The secondary reads can be merged with any locus that is within a nucleotide distance higher than M, usually $M+2$. One important thing to notice is that the nucleotide distance may be due to polymorphisms or may be due to sequencing error. When M is set very low, relatively few stacks will be merged and fewer SNPs will be identified (Figure 1.7a), on the other hand, a very high value will merge repetitive stacks into large loci [45] (Figure 1.7b). *De novo* locus formation is followed by identifying SNPs either using maximum likelihood [51] or Bayesian approaches [81, 39].

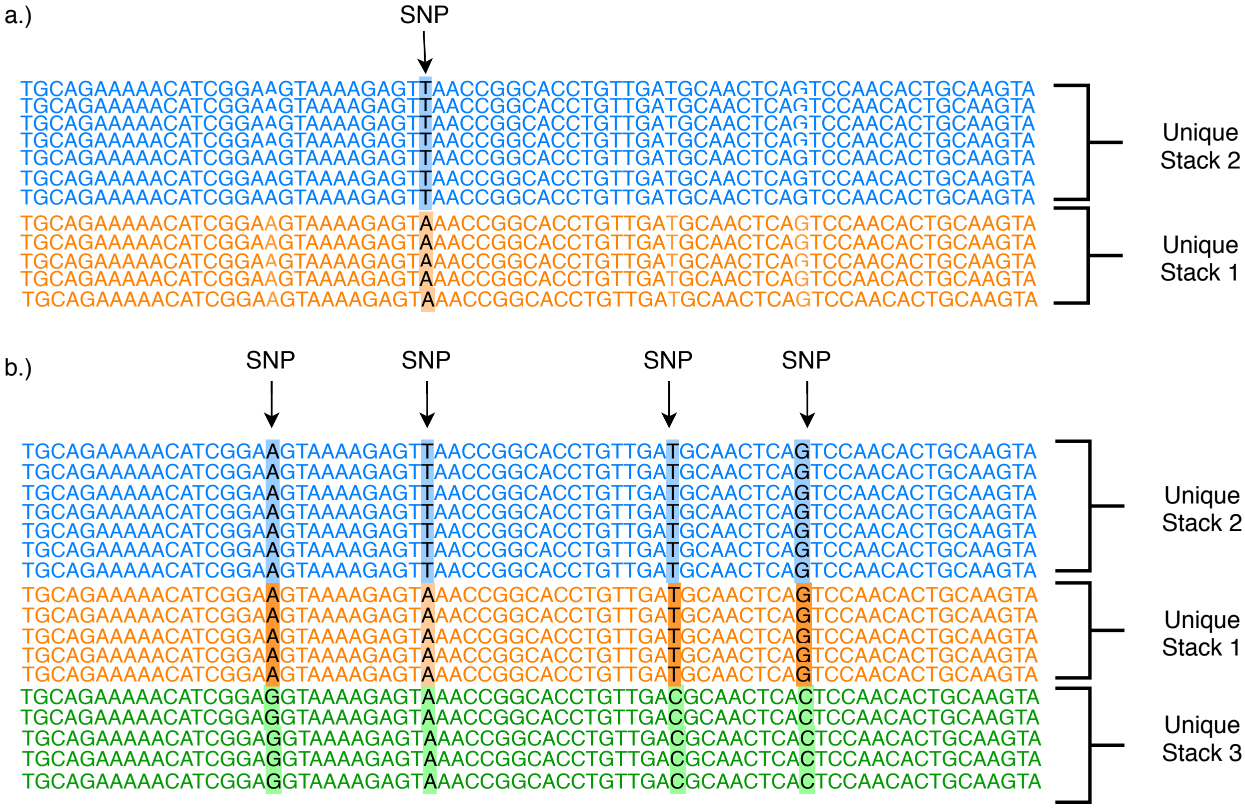
a.)
SNP
TGCAGAAAAACATCGGAAGTAAAAGAGTTTAACCGGCACCTGTTGATGCAACTCAGTCCAACACTGCAAGTA
TGCAGAAAAACATCGGAAGTAAAAGAGTAAACCGGCACCTGTTGATGCAACTCAGTCCAACACTGCAAGTA
Unique Stack 2
Unique Stack 1
b.)
SNP
SNP
SNP
SNP
TGCAGAAAAACATCGGAAGTAAAAGAGTTTAACCGGCACCTGTTGATGCAACTCAGTCCAACACTGCAAGTA
TGCAGAAAAACATCGGAAGTAAAAGAGTAAACCGGCACCTGTTGATGCAACTCAGTCCAACACTGCAAGTA
TGCAGAAAAACATCGGAGGTAAAAGAGTAAACCGGCACCTGTTGACGCAACTCACTCCAACACTGCAAGTA
Unique Stack 2
Unique Stack 1
Unique Stack 3

Figure 1.7: a.) *De novo* locus formed from unique stack 1 and unique stack 2. One single nucleotide polymorphism (SNP) is identified with alleles T and A. b.) *De novo* locus formed from unique stack 1, unique stack 2 and unique stack 3. Four SNPs are identified with alleles A/G, T/A, T/C, and G/C.

1.4 Building catalog of loci

Once the *de novo* locus formation is complete and SNPs are identified, loci from all the samples are added to a catalog along with the SNPs and alleles detected. The catalog contains all the loci and alleles identified from all the samples in the dataset (Figure 1.8). During catalog formation, the loci that are from different samples and within a specified nucleotide distance threshold (n) are merged into one locus. This is to consider differently fixed versions of the same locus as alleles rather than separate loci. Two loci are merged into one catalog locus if they have one or more alleles within the specified nucleotide distance threshold. The process can become complex and time consuming as the number of samples increases, since the catalog keeps growing as loci from new samples are processed and added to the catalog. The catalog of loci is then filtered based on different criteria, for example, a locus may be required to be present in at least a specified percentage of individuals in each population or in a minimum number of populations or the frequency of the minor allele (the second-most-frequent allele at a given locus) must be higher than a given threshold. From the retained catalog loci, population genetics statistics like the F_{ST} [107], F_{IS}, nucleotide diversity etc., are generated.

1.5 Need for efficient methods to process RADSeq data

One limitation with the existing programs is the uncertainty in the minimum coverage depth (m) and the maximum nucleotide distance (M) to be satisfied to determine potential allelic pairs, and the maximum catalog nucleotide distance (n). The number of loci formed in each sample, the quality and quantity of the SNPs identified, and the catalog built all depend on these parameters. Many recent publications [56] [75, 66, 91] have emphasized how the parameter settings during *de novo* assembly and identifying SNPs can significantly affect the number of loci obtained, genotyping error rate and

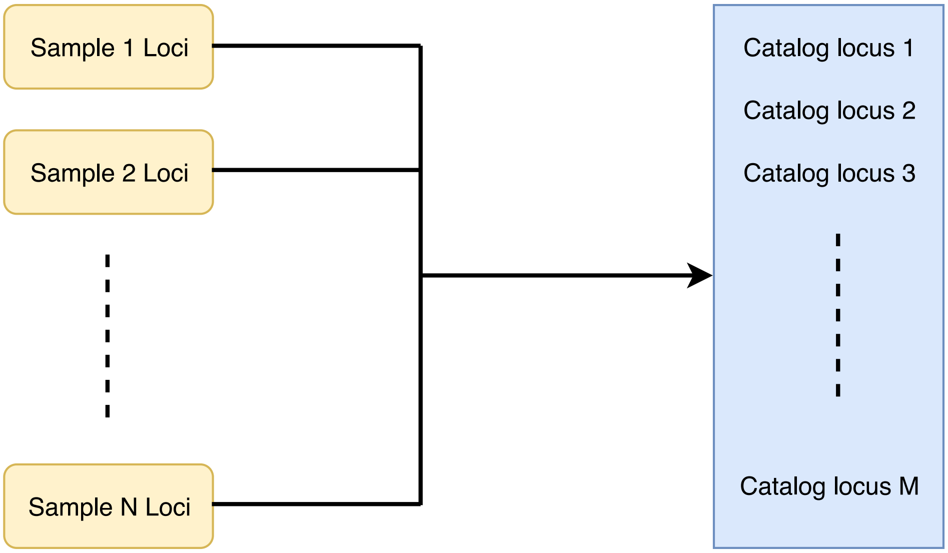

Figure 1.8: Building catalog of loci from all the loci identified from the samples 1 to *N*. The resulting catalog has *M* number of loci.

population genetic inferences. Comparing the results for different sets of parameter values requires repeating the *de novo* locus formation and catalog building. RADSeq studies often require processing hundreds of samples collected from many different locations [30, 114] and as such it is often a computationally demanding task to explore different parameter sets. In order to enable parameter sweep on the *de novo* locus formation and catalog building processes and compare their downstream effects, we need faster and efficient RADSeq data processing methods.

1.6 Need to identify paralogs

Alignment methods can fail if too-stringent cut-offs for *M* exclude pairs of orthologs with high levels of divergence, whereas paralogs with lower levels of divergence might be mistaken for alleles of the same locus, leading to high rates of false positives [21]. Paralogs can be erroneously merged into a single locus, leading to the conflation of allelic variation with differences among closely related gene family members [31]. Assembling paralogs as single loci increases false heterozygous genotype calls and can also confound genetic differentiation among individuals and populations, complicating genomic studies [26, 1]. The identification of variants due to genome duplications complicates population genomic inferences and remains an ongoing challenge in species lacking reference genomes or a reference database of DNA sequences representing the genes of a species. Extensive exploration of assembly parameters and downstream analysis and pruning of putative paralogous loci is a necessary quality control measure in RADSeq studies.

1.7 Directional relative gene flow among populations

Gene flow (migration) is the movement of alleles from one population to another. Gene flow from source population to recipient population introduces new alleles in to the recipient population and changes the corresponding allele frequencies (proportion of individuals in which the allele is present) in the populations (Figure 1.9). One of the primary applications of RADSeq in population studies is to understand population structure [91] using measures of genetic differentiation. The level of genetic differentiation can be estimated using measures like F_{ST} [112, 52] and Nei's G_{ST} [80] using allele frequency data. The pairwise genetic differentiation between populations can be estimated using the allele frequency data and used as a distance metric to cluster the populations and infer the population structure.These genetic differentiation measures assume gene flow among populations to be symmetric (same migration rates in both directions), which is not always the case. If the gene flow is asymmetric, the rate of gene flow is not same in both directions and may lead to skewed genetic differentiation estimates among the populations [29]. Hence it is important to understand the gene flow patterns and the processes leading to genetic structuring of populations (e.g., [86]).

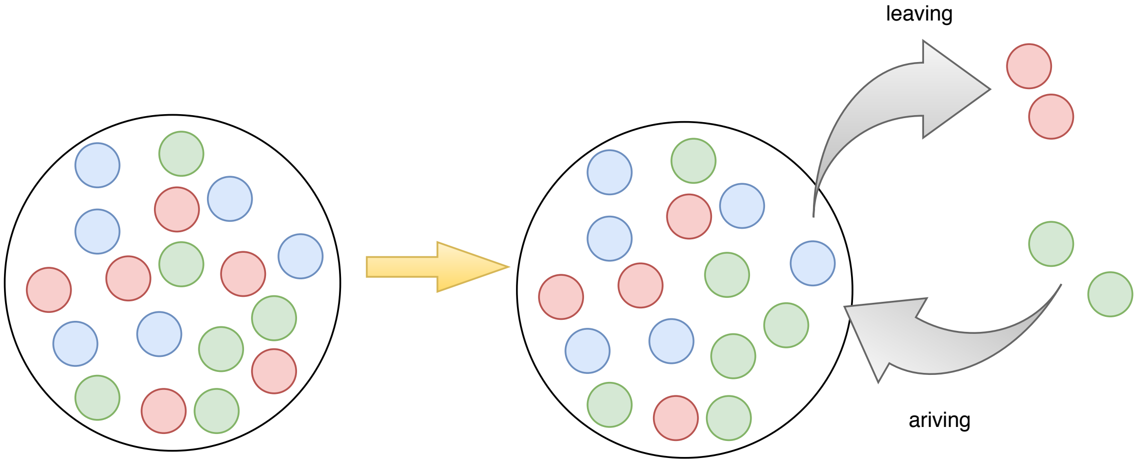

Figure 1.9: Alleles arriving and leaving in a population as a result of gene flow.

1.8 Computational methods developed

Graph theory and graph-based structures have a wide range of applications in the field of computational biology, such as genome assembly, genome alignment, population genetics, landscape genetics, biomedical informatics etc. A graph G consists of a set of vertices V and a set of edges E. Two vertices are linked if there exists an edge connecting them. A graph can be directed, which means the edges are directed from one vertex to another, or undirected where there is no direction associated with the edge. The ease of representing biological data into graph-based structures, to study them and make new inferences makes graph theory more applicable in this discipline. For example, Population Graphs [32] uses a graph theoretic framework to estimate population genetic statistics. In whole-genome shotgun NGS sequencing, a De Bruijn graph representing the connections among the reads is commonly used for *de novo* assembly of reads into longer continuous contig and scaffolds (e.g., Edena [48], Velvet [115], ABySS [95], ALLPATHS-LG [42]).

In chapter 2, we introduce RADProc, a software package that uses a graph data structure to represent all sequence reads and their similarity relationships. RADProc builds a graph from the unique stacks identified from all the samples in the dataset (Figure 1.10). Each vertex represents a unique stack and all relevant information about the unique stack required for *de novo* locus formation and catalog building. This includes coverage depth, the samples/individuals containing the unique stack, the population in which the unique stack is present. This is not a fully connected graph because there is a maximum nucleotide distance threshold used to build the graph. The graph built is similar to the one built during *de novo* locus formation by Stacks, except that in Stacks the graph is built from unique stacks identified from each sample separately, whereas in RADProc the unique stacks from all the samples are used to build the graph. While Stacks builds the graph at individual/sample level, RADProc builds the graph at the entire dataset level.

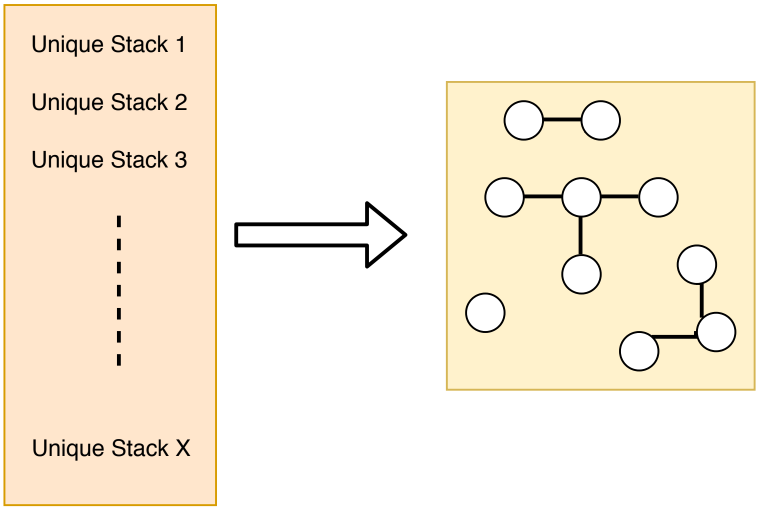

Figure 1.10: Unique stacks transformed into a network with stacks connected based on maximum nucleotide distance threshold. The nodes represent the unique stack (putative allele) sequences and the edge weights represent the nucleotide distance between the unique stacks.

For *de novo* locus formation from RADSeq data, we use a graph-based structure built from unique stacks of reads and the similarity among them as edge weights to merge them into putative loci with each unique stack representing a putative allele in the locus. Storing sequence-comparison results in a graph eliminates unnecessary and redundant sequence-similarity calculations. *De novo* locus formation for a given parameter set can be performed on the pre-computed graph, making parameter sweeps far more efficient. RADProc also uses a clustering based approach for faster nucleotide-distance calculation. Since the RADProc graph structure contains population level information of the unique stack similarity relationships, it can also accelerate the catalog building process. The catalog locus is nothing but a subgraph, which consist of all the unique stacks representing the locus alleles that are at least one edge within a specified nucleotide distance threshold.

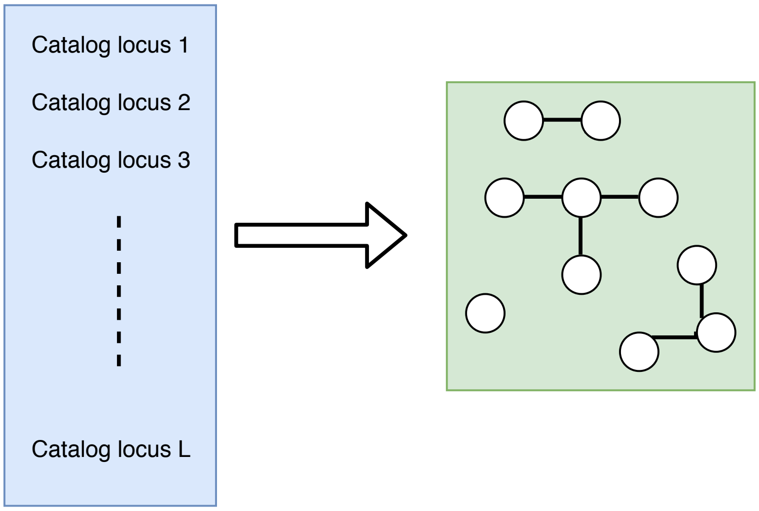

Figure 1.11: Catalog loci transformed into a network with loci connected based on similarity threshold. The nodes represent the catalog loci sequences and the edge weights equal to the sequence similarity.

In chapter 3, we describe PMERGE, a novel method that identifies candidate paralogs or duplicated loci in the catalog of loci built from the RADSeq data. PMERGE works by building networks of catalog loci that share high levels of consensus sequence similarity and flagging highly similar sequences as potential paralogs. To identify paralogs, PMERGE builds a network of catalog loci based on a nucleotide distance threshold value (Figure 1.11). The network obtained is not fully connected since two catalog loci are connected by an edge only if they satisfy the nucleotide distance threshold value. Unlike the RADProc network, in which the network nodes are the unique stacks, in PMERGE the network nodes are the catalog loci. By embedding PMERGE in the analysis pipeline of the widely used Stacks software [17], it is straightforward to apply it as an additional filter in population-genomic studies using RADSeq data and can also be used in addition to other existing approaches.

In chapter 4, we propose an approach to determine relative gene flow among populations using allele frequency data. The method identifies exclusively shared alleles between each pair of populations and uses the distribution of differences in the allele frequencies of such exclusively shared alleles to determine the direction and relative migration between pairs of populations. The proposed method is an extension to the RADProc graph structure application. The information such as the alleles that are exclusively shared between a pair of populations and their allele frequency data for any given set of M, m and n can be easily extracted from the RADProc graph. Based on the extracted information, we calculate a proposed measure W, which represents the relative gene flow between each pair of populations. Once we obtain the W values between each pair of populations, then we can build a network of populations with W being the edge weights to visualize the gene flow patterns (Figure 1.12). The network represents the relative gene flow in both the directions between all pairs of populations.

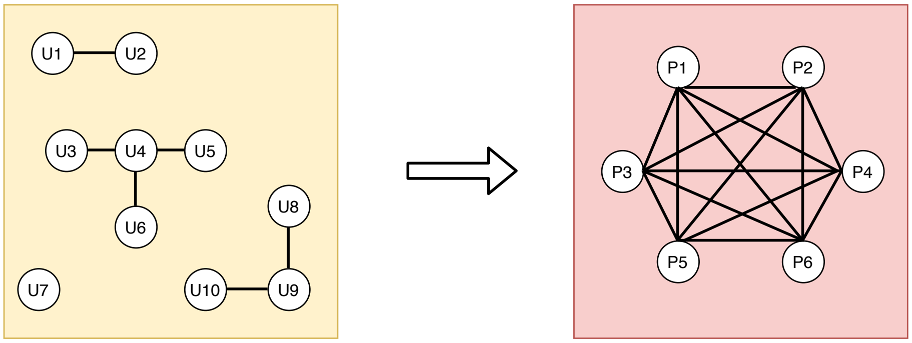

Figure 1.12: Populations network built from RADProc graph based on the proposed measure W to visualize relative gene flow among the populations P1 - P6.

1.9 Publications

1.9.1 Chapter 2

Title: RADProc: A computationally efficient *de novo* locus assembler for population studies using RADseq data.

Publication: Molecular Ecology Resources

Authors: Praveen Nadukkalam Ravindran, Paul Bentzen, Ian R. Bradbury and Robert G. Beiko

Year: 2019

Volume: 19

Pages: 272-282

DOI: https://doi.org/10.1111/1755-0998.12954

Contribution

Aided in the conception of the work. Implemented the software and conducted analysis. Contributed to writing, revising and approving the final draft of the manuscript.

1.9.2 Chapter 3

Title: PMERGE: Computational filtering of paralogous sequences from RAD-seq data.

Publication: Ecology and Evolution

Authors: Praveen Nadukkalam Ravindran, Paul Bentzen, Ian R. Bradbury and Robert G. Beiko

Year: 2018

Volume: 8

Pages: 7002-7013

DOI: https://doi.org/10.1002/ece3.4219

Contribution

Aided in the conception of the work. Implemented the software and conducted analysis. Contributed to writing, revising and approving the final draft of the manuscript.

Chapter 2

RADProc: A computationally efficient *de novo* locus assembler for population studies using RADSeq data

2.1 Introduction

RADProc is an algorithm and software package that streamlines and accelerates *de novo* locus formation and catalog building from RADSeq data by eliminating redundancies and using a highly efficient method for nucleotide distance calculation. RADProc can efficiently sweep through different sets of parameters for *de novo* locus formation and catalog building. Although RADProc is an alternative method for *de novo* locus formation and catalog building from RADSeq data, the output files generated by RADProc are completely compatible with the Stacks pipeline. We use the same parameters defined in Stacks for *de novo* locus formation and catalog building. In Stacks, the parameter m represents the minimum sequence depth criterion and the parameter M determines the maximum nucleotide distance allowed between stacks that are to be merged into a locus. The unique stacks with coverage depth less than m (secondary stacks) are merged with the already formed locus if they uniquely match to a locus with nucleotide distance less than or equal to $M+2$. The secondary stacks that match multiple primary stacks are discarded from the analysis. Once the *de novo* locus formation from RADSeq data for a given sample is completed, the single nucleotide polymorphisms (SNPs) in each locus are identified, with a maximum-likelihood approach to distinguish true variants from probable sequencing errors. During the catalog-building step, loci from different individuals that are within a specified nucleotide distance n are merged into a single catalog locus.

RADProc can be used in two modes; one is to sweep through the different parameter values for m, M, and n to compare and identify optimal parameter settings and the other is to directly use one set of known parameter values for m, M, and n. Although there is no strict definition of optimal parameters, as suggested in [82] the combination of M and m parameter values that give the highest number of polymorphic loci in 80% of individuals in each population could be considered as an optimal parameter set. The idea is to use a subset of population samples to identify the optimal parameters and then apply them to the larger dataset. Since RADProc is faster in *de novo* locus formation and catalog building, the user can also try the parameter sweep on larger datasets. We demonstrate and discuss the RADProc modes using small and large datasets in the sections below.

2.2 Methods

2.2.1 RADProc graph data structure

The core of RADProc is the graph structure (Figure 2.1) used to store all unique stacks with a coverage depth of at least two reads, omitting singleton reads that would be eliminated by any reasonable value of m. The RADProc graph structure encapsulates the principle of connecting unique stacks within a given nucleotide distance (i.e., parameter M in Stacks), but is tailored to avoid redundancy in two important ways. First, all samples are processed and stored in a single graph data structure that incrementally adds unique stacks from all samples. Since most stacks will be present in more than one sample, this approach eliminates a great deal of redundancy in the storing of data and removes the need for repeated comparisons of the same unique stacks during locus construction. Second, the graph connects stacks that differ up to a relatively large nucleotide-distance threshold M_G, which allows *de novo* locus assembly for all $M \epsilon 1, \ldots, M_G$ - 2 directly from the graph, since we need to allow a more lenient nucleotide distance of $M + 2$ for unique stacks with coverage depth less than m to be merged with the already formed loci if they uniquely match to a locus. So, if we want try M values up to 6, we will have to set $M_G = M + 2 = 6 + 2 = 8$. For example, if the graph is built with $M_G = 6$, then *de novo* assembly can be performed efficiently for all values of M from 1 to 4 and M+2 from 3 to 6. In the absence of the graph data structure, the program must perform sequence comparisons each time a new value of M is used in a parameter sweep, whereas the graph structure performs sequence comparisons only once, with locus construction based directly on distance values stored in the graph. Since the minimum coverage depth m required forming a node in the graph structure is two (i.e., a unique stacks), we can also try different values of m without the need to identify unique stacks each time.

Formally, RADProc is a graph structure $G = (N, E)$, where N and E are node and edge set, respectively, defined by

$N = u_1, u_2, u_3, u_4, u_5, \ . \ . \ . \ , u_n$ where u_i represents a unique stack

E = (u, v) $\in$ N $\times$ N: u $\neq$ v, τ (u, v) $\leq$ M_G where τ (u, v) is the nucleotide distance between the unique stack sequences.

Each node in N contains information about a unique stack, and edges in E connect all pairs of nodes where the nucleotide distance between the corresponding sequences is less than or equal to the threshold value M_G. The unique stack represents a putative allele of a putative locus; since RADSeq is typically applied to samples from the same species, many unique stacks are likely to be present in most or all of the samples. Minimal overlap among samples will generate an impractically large graph for datasets with hundreds of samples. To simplify graphs constructed from large datasets, we use a filter to remove unique stacks that are not present in a minimum

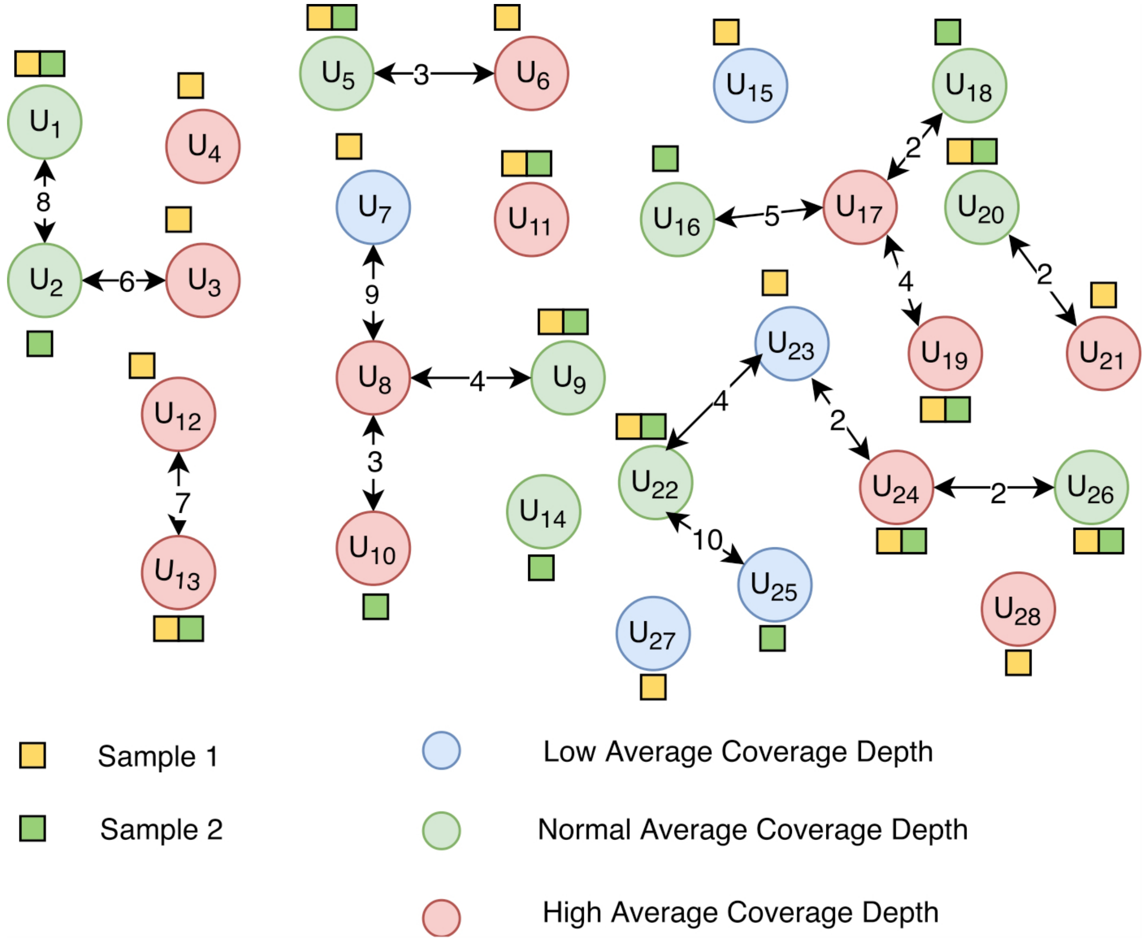

Figure 2.1: Sample RADProc graph structure. The graph represents connected unique stacks in two different samples. The nodes are connected only if they are within the nucleotide distance threshold M_G; in this example, M_G=10.

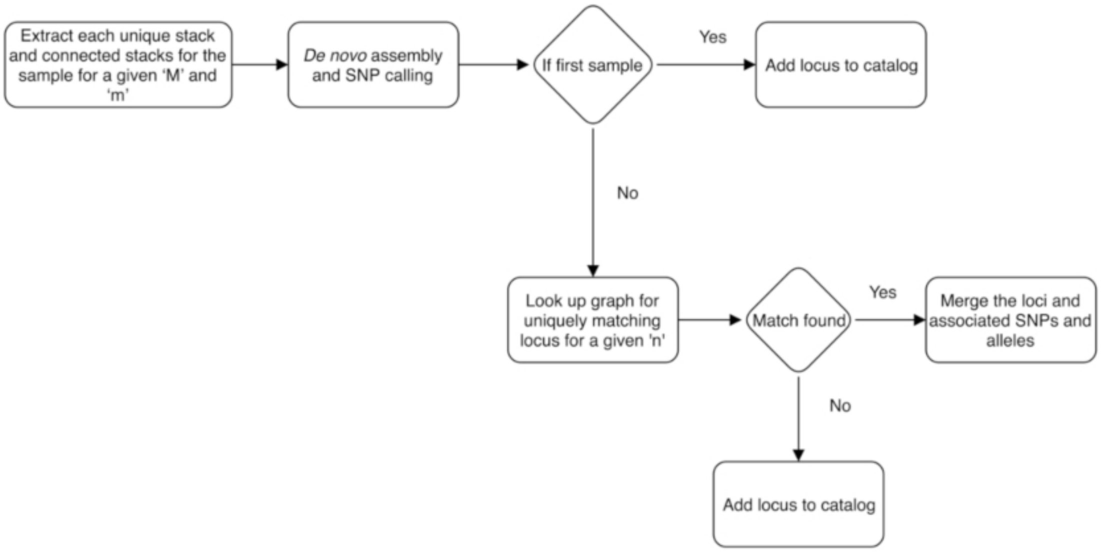
Extract each unique stack and connected stacks for the sample for a given 'M' and 'm'
De novo assembly and SNP calling
If first sample
Yes
Add locus to catalog
No
Look up graph for uniquely matching locus for a given 'n'
Match found
Yes
Merge the loci and associated SNPs and alleles
No
Add locus to catalog

Figure 2.2: De novo assembly and catalog building. The proposed graph structure built from the samples is processed, *de novo* loci for each individual are assembled, and the catalog is constructed from the graph.

number of samples S and the minimum average coverage depth across samples D. This approach ensures that unique stacks are present in at least $S\%$ of the samples, while still retaining the potential to keep "private" alleles that are restricted to a single population. Once the graph has been constructed, *de novo* locus formation requires only lookup operations and extraction of the unique stack nodes that need to be merged; since this process requires no sequence comparisons, it can be performed quickly for all desired parameter combinations (Figure 2.2).

2.2.2 Nucleotide distance calculation

Nucleotide distance is the number of mismatches between a pair of sequences. In the *de novo* assembly process, calculating the nucleotide distance between DNA sequences is generally the rate-limiting step, depending on the number of unique reads in each sample and the total number of samples. We evaluate the basic pairwise sequence comparison and k-mer counting method used by Stacks and a proposed clustering based nucleotide distance calculation method.

Pairwise sequence comparison

Pairwise sequence comparison or string comparison is the basic method where two sequences are compared character by character and every mismatch increments the nucleotide distance by 1 unit. If we have to identify the pairwise nucleotide distance between 'n' sequences then $n*(n-1)/2$ sequence comparisons need to be performed and the time complexity would be O(n^2). Let 'l' be the length of the sequences, then the time complexity for nucleotide comparisons would be O($n^2 * l$). If we stop the comparison process for a given sequence pair once we reach the maximum mismatches allowed, then the runtime varies based on the maximum mismatches allowed. In the worst case scenario all the l characters need to be compared.

K-mer Counting

The word k-mer refers to all the possible substrings of length k in a string. For a string of length l, there can be $l-k+1$ k-mers (Figure 2.3). The method works by breaking all the sequence strings into k-mers of length k and counting the occurrences of query sequence k-mers in each of the subject sequence. Depending on the k-mer length, the total number of sequences and the maximum nucleotide distance threshold, the method can be computationally expensive and time consuming. There are many efficient algorithms for counting k-mers like Bloom filter [78] based approach, Hashing-based approach (JELLYFISH [72]) and Suffix array based approach (Tallymer [64]). Stacks implements a Hashing-based approach to perform k-mer counting.

Identifying the nucleotide distances between a set of sequences using Hashing-based k-mer counting involves the following steps,

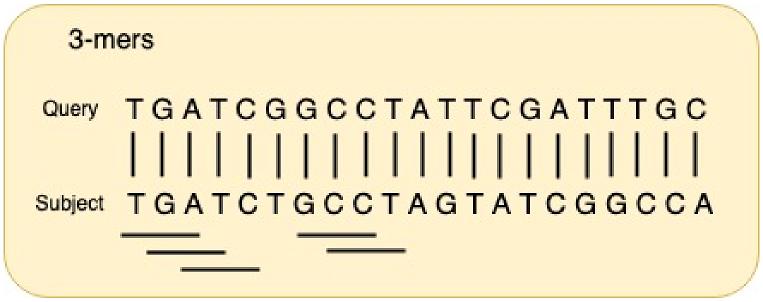

Figure 2.3: Identifying k-mers in query and target sequences. Matching 3-mers (k-mer of length 3) between the query and target sequences are shown.

i. Determine the length of the k-mers to be used in the comparison, based on the maximum nucleotide distance and the sequence length. For a given maximum nucleotide distance, calculate the span from equation 2.1 starting with a default k-mer length and increase the k-mer length until the span value is greater then the sequence length. From equation 2.1, k-mer_length is inversely proportional to maximum_nucleotide_distance, hence increasing maximum_nucleotide_distance would decrease the k-mer_length.

$$span = (k - mer\_length * (maximum\_nucleotide\_distance + 1)) - 1 \tag{2.1}$$

ii. Break all the sequences in the set into k-mers using the value of k determined in the previous step.

iii. Determine the minimum number of matching k-mers in the query and subject sequences using equation 2.2 to consider for a full sequence comparison.

$$minimum\_matches = sequence\_length - span \tag{2.2}$$

iv. All the subject sequences, which have the required minimum number of matching k-mers with the query sequence are considered for full sequence comparison with the query sequence.

If n is the total number of k-mers, then the space complexity would be $O(n)$. Let j be the average number of k-mers generated per sequence, then the time complexity for identifying matching k-mers would be $O(j)$ per sequence (hash table look up is $O(1)$). For i number of sequences the time complexity would be $O(ij)$. Increasing the value of M, reduces the length of the k-mer, as a result the total number of k-mers generated increases and the minimum number of matching k-mers required for full sequence comparison decreases. Consequently, the memory required to store the k-mers and the runtime increase as the value of M increases. In the worst-case scenario, most of the subject sequences would be selected for full sequence comparison; hence, the time taken for k-mer generation and identifying matching k-mers, and the memory occupied by the generated k-mers create overhead without any gain.

RADProc Nucleotide Distance Calculation

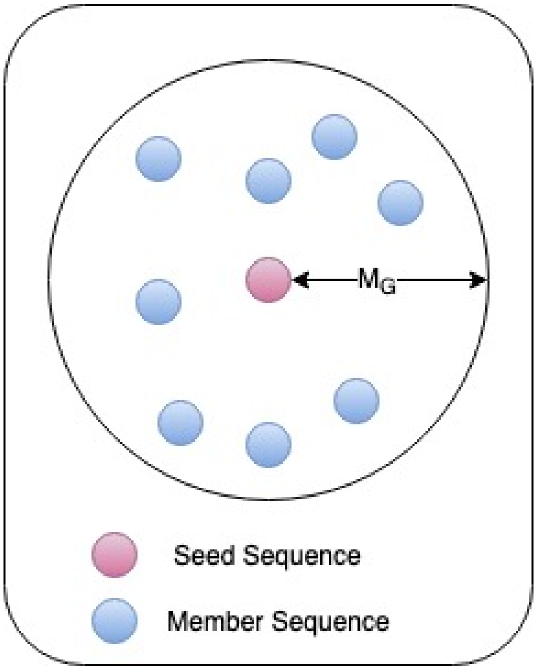

Figure 2.4: Cluster with 8 member sequences and a seed sequence. The variable M_G represents the radius of the cluster or the maximum distance between the seed and the member sequences.

Rather than performing an all-versus-all comparison of sequences, RADProc uses a heuristic approach similar to UCLUST [35] to cluster similar sequences according to M_G and then performs pairwise comparisons only between sequences within these clusters for the purpose of graph construction (Figure 2.4). The clusters are built by iterating through the list of sequences and identifying "seed sequences" that define their corresponding clusters. Each sequence is compared against the existing set of cluster seeds: If the nucleotide distance between the cluster seeds and the new sequence is greater than M_G, then the new sequence becomes a seed sequence of a new cluster. The nucleotide distances between incoming sequence and the seed sequences are calculated by pairwise sequence comparison only if the sequences have a minimum number of matching k-mers. Once the matching seed sequence is identified, the incoming sequence is added to that cluster. As a result, all the sequences are grouped into clusters in which all constituent sequences are within a maximum nucleotide distance from the seed sequence. RADProc also performs pairwise sequence comparisons between sequences from two different clusters if the nucleotide distance between their seed sequences is less than $2 \times M_G$, because the initial clustering of sequences depends highly on the order of incoming sequences, so similar sequences (within M_G) could end up in different clusters (Figure 2.5). Using $3 \times M_G$ as the nucleotide distance threshold between the seed sequences covers all the clusters with sequences less than M_G apart, but increases the run time as more clusters are compared than using $2 \times M_G$. If we have to make sure that all the sequences within the maximum nucleotide distance are identified then using $3 \times M_G$ would be appropriate.

2.2.3 Validation datasets

Evaluation of RADProc was performed using RADSeq data extracted from two different studies. The first dataset comprised green crab (*Carcinus maenas*) samples, which were first used to study their population structure in the Northwest Atlantic [58]. A

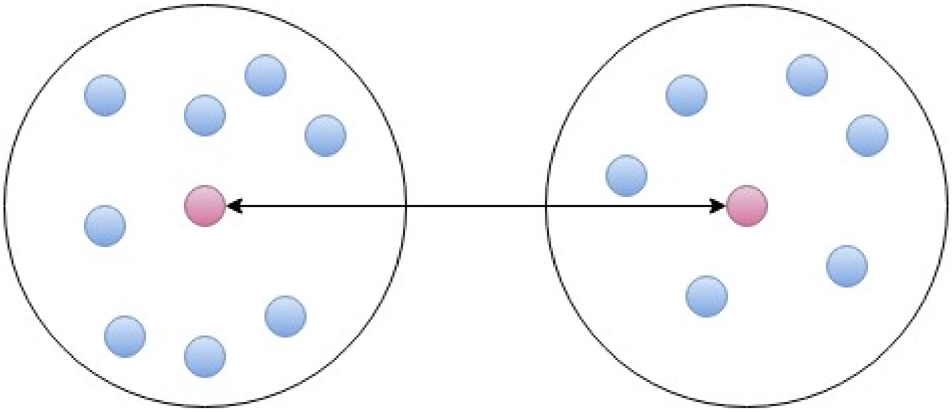

Figure 2.5: Comparing two clusters. If the identity between seed sequences of two clusters is less than or equal to 2 × M_G, then all the sequences in the two clusters are compared.

smaller dataset comprising 20 samples from four different sites, and a larger dataset consisting of 242 samples from 11 different sites were used to test the performance of RADProc on datasets of different sizes. Each library consisted of 22 samples identified by variable-length in-line barcodes ranging from 5 to 9 bp. The libraries were sequenced on a HiSeq 2000 (Illumina) as 100 bp paired end sequences. Each sample comprised approximately 2.5 million RAD-tags. We have also used another dataset of 16 brown trout (*Salmo trutta L.*) samples [82] from southwest England occupying clean (8 samples) and metal-impacted (8 samples) sites. The reads from the trout samples were 95 bp in length with an average of 2.9 million RAD-tags per sample.

The run-time of the RADProc nucleotide distance calculation method was compared with the traditional pairwise string comparison and k-mer counting methods. The comparisons were made using 10,000 and 50,000-sequence datasets and for M values 2, 4, 6 and 8 to compare the run-times when the number of sequences increases as well as M increases.

As was done in [82], we tested M values from 1 to 8, m from 3 to 6 and n = 1 (the default setting in the "cstacks" catalog-building program in Stacks) using Stacks and RADProc and the run-times were compared. The samples were processed using the process_radtags module in the Stacks pipeline. The *de novo* assembly and catalog building were then done using the ustacks and cstacks modules, respectively. RADProc was executed using $S = 0.10$, $D = 2$, $M_G = 10$ and $m = 6$ to perform *de novo* locus formation and catalog building for all the combinations of M and m. The analysis was performed using the smaller dataset of 20 samples in order to try all 32 different combinations of the M and m parameter values described above. We also compared the run times of Stacks and RADProc for the full dataset of 242 samples. The maximum graph distance M_G was set to 8 to try values 2 to 6 for M and m from 3 to 6. The runtime evaluation was done using a Macintosh laptop (Mac OS 10.13) with Intel Corei5-4260U CPU @ 1.40GHz processor that can support 4 parallel threads and 8 GB 1600KHz DDR3 RAM. The program was implemented using C++ 11 and used OpenMP 4.0 for parallelization.

To compare the *de novo* loci formed, and catalog built using Stacks and RADProc, we used the full dataset of 242 samples and the default Stacks parameter settings $M = 2$, $m = 3$ and $n = 1$ for *de novo* locus formation and catalog building. We used consensus sequences to identify *de novo* and catalog loci that were inferred by both methods. The catalog locus sequences built by Stacks and RADProc were then aligned to the green crab reference genome (Hleap et al., in preparation) using BLASTN version 2.2.28 [4] with a minimum of 90% sequence identity and a maximum E-value of 1e-20 to identify the number of uniquely mapping loci in both the Stacks and RADProc catalogs. We also compared the results generated by the *populations* program in Stacks using the *de novo* loci formed and catalog built using Stacks and RADProc. All the catalog loci may not be well represented within and among the populations and such catalog loci will be filtered out. The *populations* program was run with filtering settings: percent samples limit per population (r) = 0.75, which requires that a locus be present in at least the specified percentage of individuals in a population; locus population limit (p) = 11, the minimum number of populations in which a locus must be present; and minor allele frequency cutoff (*min_maf*) = 0.05, which sets a minimum threshold for the frequency of the minor allele (the second-most-frequent allele at a given locus) for each SNP in a locus.

We compared the pairwise F_{ST} [107, 52] values among the populations from the retained catalog loci for both Stacks and RADProc. The F_{ST} value is a measure of population substructure and is most useful for examining the overall genetic divergence among subpopulations. F_{ST} values up to 0.05 indicate negligible genetic differentiation whereas $F_{ST} > 0.25$ means very great genetic differentiation among the populations analyzed.

For the trout dataset, we compared the run-time performance of Stacks and RADProc for *de novo* locus formation and catalog construction for 32 parameter sets, from $M = 1$ to 8 and $m = 3$ to 6 by setting M_G =10. We also compared the *de novo* loci formed and catalog of loci built by Stacks and RADProc from each sample for each of the 32 different parameter sets (M and m). The locus consensus sequences were compared and the proportion of common loci between Stacks and RADProc was identified.

2.3 Results

After processing the raw RADSeq data from the smaller green crab dataset comprising 20 samples from four different sites using *process_radtags*, they were provided as input to RADProc. Approximately 70,000 unique stacks were identified per sample (Figure 2.6). 129,743 of the total 409,864 unique stacks identified were shared by at least two samples, with the remaining 280,121 found in one sample only (Figure 2.7). We eliminated unique stacks that did not satisfy S = 0.10 and D = 2, which left 135,309 unique stacks for further analysis.

The process was repeated with the large green crab dataset consisting of 242 samples from 11 different sites. There were 3,090,130 unique stacks loaded into the

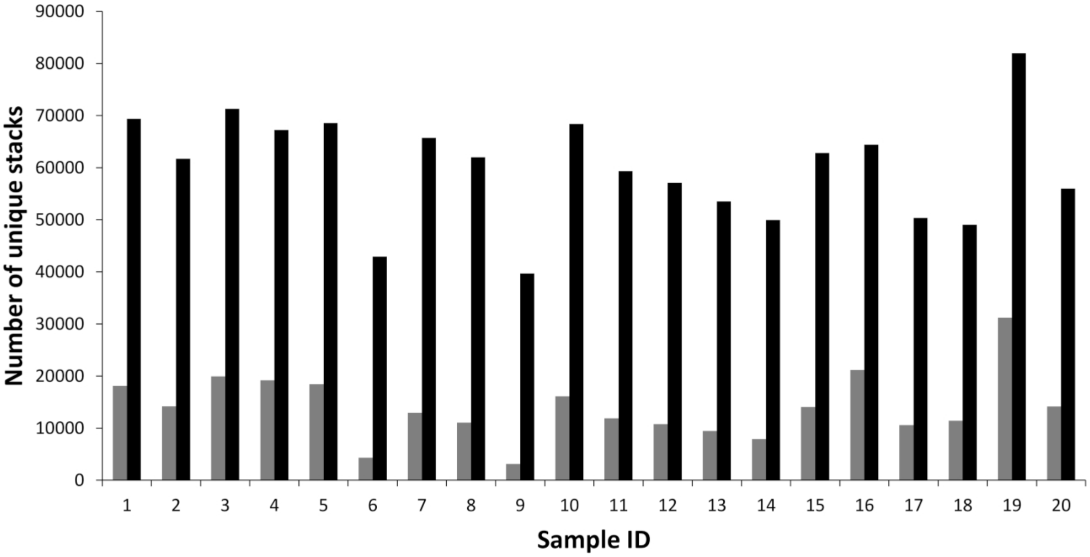

Figure 2.6: Total stacks (black bars) and unique stacks (gray bars) for each sample in the small green-crab dataset.

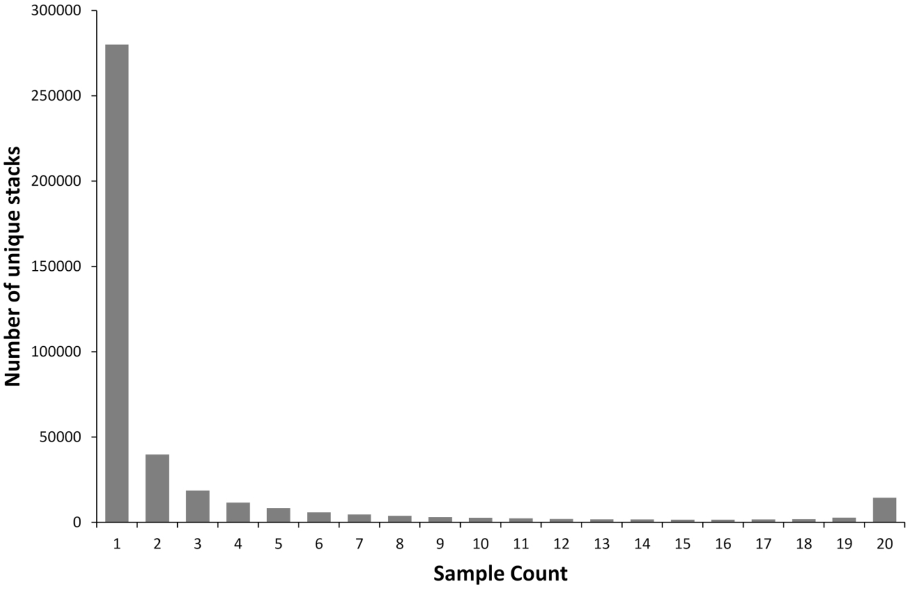

Figure 2.7: Distribution of unique stacks across samples in the 20-sample green-crab dataset. 280,121 stacks are present in one sample only, while 14,378 stacks are present in all 20 samples.

10000 Sequences			
M	**Pairwise String Comparison**	**K-mer Counting**	**RADProc Nucleotide Distance Calculation**
2	0m56.124s	0m6.486s	0m3.087s
4	0m59.856s	0m24.701s	0m2.565s
6	1m3.197s	0m42.916s	0m3.240s
8	1m7.251s	3m21.764s	0m3.982s
50000 Sequences			
2	25m45.530s	0m20.727s	0m48.643s
4	27m33.725s	1m5.604s	0m57.342s
6	33m38.622s	2m15.783s	0m59.241s
8	35m10.312s	76m46.063s	1m5.485s

Table 2.1: Run-time evaluation of the RADProc nucleotide distance calculation , pairwise string comparison, and k-mer counting methods for different values of M using datasets comprising 10,000 and 50,000 sequences.

RADProc graph data structure. As above, the unique stacks were formed from RAD-tags with coverage depth of at least two reads. After filtering out unique stacks that were not present in at least 22 samples and average coverage depth $D <= 2$, a total of 426,260 unique stacks were retained. RADProc took 50 minutes to load RADSeq data from all the 242 samples, calculate nucleotide distances with threshold M_G, and build the graph. The time taken to perform *de novo* locus formation and catalog building for one set of parameter values was 275 minutes. In total, to try all the 20 different parameter settings, RADProc took 92 hours 30 minutes (approximately four days). The large dataset was also processed using Stacks, which took 398 hours 40 minutes (approximately 17 days) for *de novo* locus formation and catalog building for the 20 different values of M and m.

Table 2.1 shows the run-time comparisons for the RADProc nucleotide distance calculation method and the pairwise sequence comparison and k-mer counting methods. For both the 10,000 and 50,000-sequence test datasets, the RADProc nucleotide distance calculation was faster than the pairwise string comparison and k-mer counting methods. Table 2.2 lists the run-time for the Stacks modules *ustacks* and *cstacks*

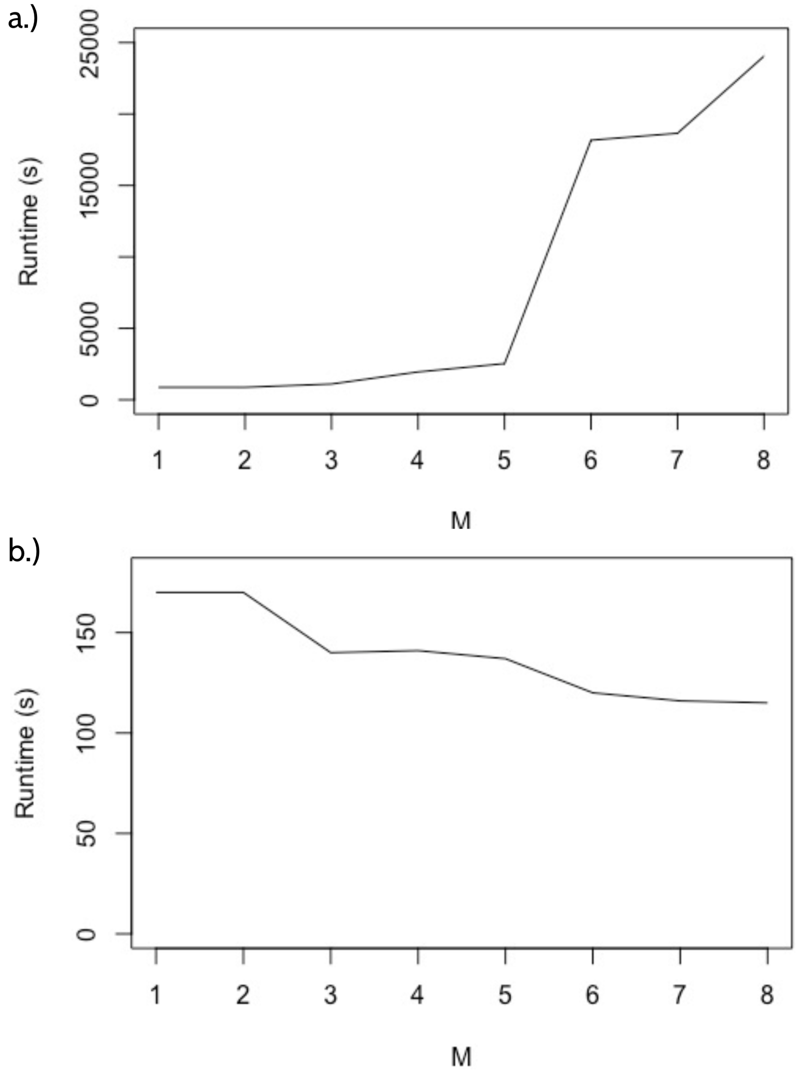

Figure 2.8: Runtime for values of $M = 1$ to 8 during a.) *de novo* locus formation and b.) catalog building.

m = 3					**m = 4**			
M	**ustacks**	**cstacks**	**Total**	**M**	**ustacks**	**cstacks**	**Total**
1	14 min 35 s	2 min 50 s	17 min 25 s	1	14 min 15 s	2 min 43 s	16 min 58 s
2	14 min 35 s	2 min 50 s	17 min 25 s	2	14 min 17 s	2 min 44 s	17 min 01 s
3	18 min 25 s	2 min 20 s	20 min 45 s	3	18 min 20 s	2 min 15 s	20 min 35 s
4	32 min 30 s	2 min 21 s	34 min 51 s	4	32 min 10 s	2 min 10 s	32 min 20 s
5	42 min 13 s	2 min 17 s	44 min 30 s	5	42 min 10 s	2 min 10 s	44 min 20 s
6	302 min 49s	2 min 00 s	304 min 49 s	6	302 min 35 s	2 min 03 s	304 min 38 s
7	310 min 54 s	1 min 56 s	312 min 19 s	7	308 min 30 s	1 min 45 s	310 min 15 s
8	400 min 50 s	1 min 55 s	402 min 45 s	8	390 min 40 s	1 min 40 s	392 min 20 s
m = 5					**m = 6**			
M	**ustacks**	**cstacks**	**Total**	**M**	**ustacks**	**cstacks**	**Total**
1	14 min 15 s	2 min 43 s	16 min 58 s	1	14 min 04 s	2 min 40 s	16 min 44 s
2	14 min 15 s	2 min 42 s	16 min 57 s	2	14 min 05 s	2 min 39 s	16 min 44 s
3	18 min 15 s	2 min 13 s	20 min 28 s	3	18 min 5 s	2 min 07 s	20 min 12 s
4	32 min 06 s	2 min 10 s	34 min 16 s	4	31 min 59 s	2 min 01 s	34 min 00 s
5	42 min 8 s	2 min 10 s	44 min 18 s	5	41 min 50 s	2 min 03 s	43 min 53 s
6	301 min 41s	2 min 05 s	303 min 46 s	6	298 min 03 s	2 min 01 s	300 min 04 s
7	306 min 13 s	1 min 45 s	307 min 58 s	7	305 min 23 s	1 min 44 s	307 min 07 s
8	385 min 30 s	1 min 46 s	387 min 16 s	8	382 min 50 s	1 min 42 s	384 min 32 s

Table 2.2: Run-time evaluation of ustacks and cstacks for different values of M and m. Total runtime of Stacks across all parameter sets was 78 hours, versus 2 hours 40 minutes for RADProc.

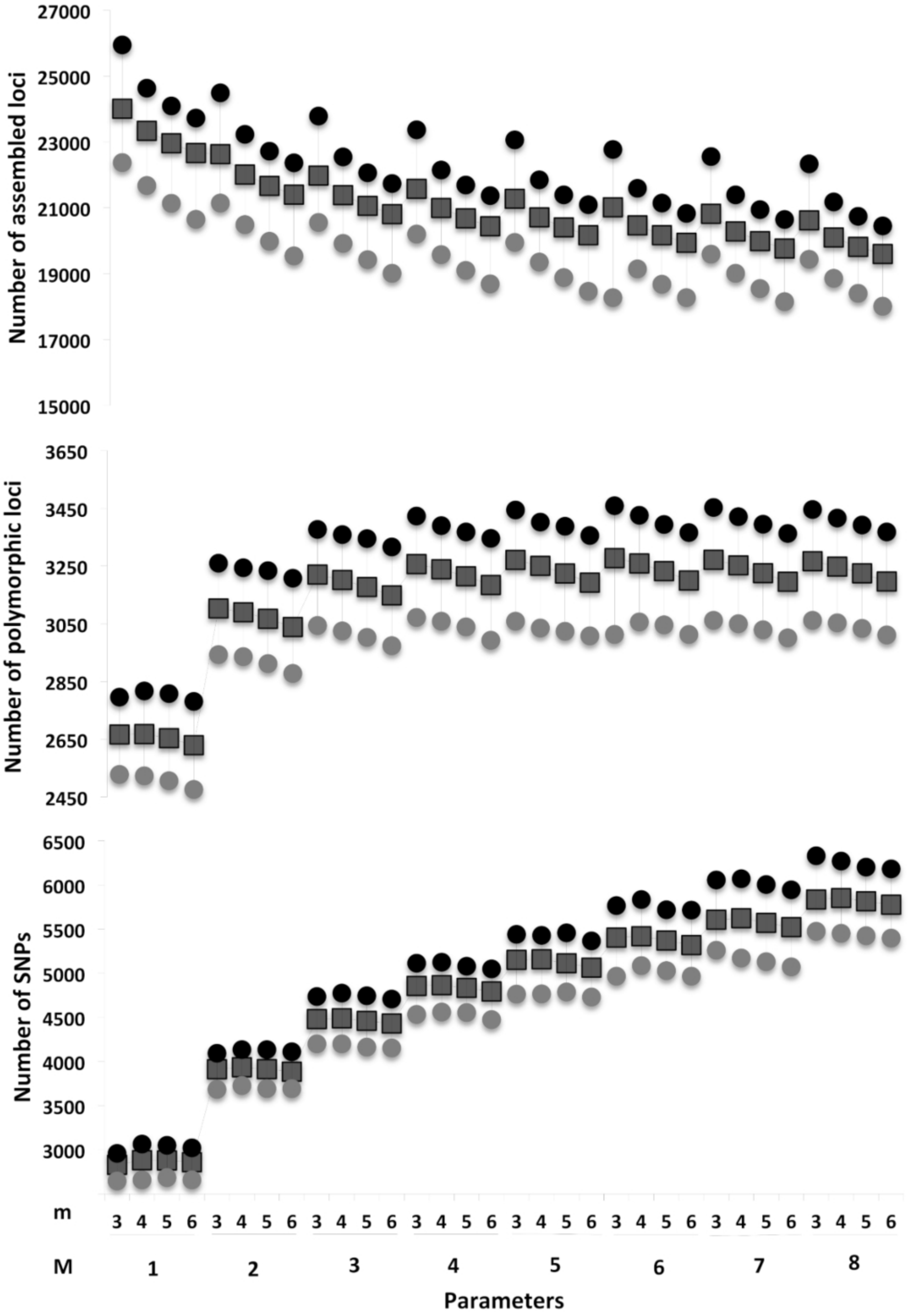

Figure 2.9: Comparisons using different parameter values for M and m. The minimum, average and maximum number of *de novo* loci formed, number of polymorphic loci, and SNPs identified for different values for M and m for the 20 samples are shown.

for all 32 different tested combinations of M and m. For *de novo* locus formation, the runtime increased as the M value increased, and for the catalog building the runtime decreased as the M value increased (Figure 2.8). Stacks took approximately 78 hours in total, while RADProc took just 2 hours 40 minutes to perform the same analysis. RADProc took 5 minutes in total to load RADSeq data from all 20 samples, perform nucleotide-distance calculation with threshold M_G and build the graph structure, and approximately 5 minutes for *de novo* locus formation and catalog building for each set of parameters. Figure 2.9 compares the total number of loci, polymorphic loci, and SNPs identified for the 32 different combinations of values for M from 1 to 8 and m from 3 to 6. In this example, the number of total loci formed and number of polymorphic loci identified plateaus after $M = 4$ but the number of SNPs identified keeps increasing as M increases suggesting that there could be loci with a high density of SNPs [82]. On average, approximately 97% to 98% of the loci formed by RADProc were identical to the ones formed by *ustacks*, and on average RADProc produced approximately 3% fewer *de novo* loci than ustacks for different values of M and m. Similarly, we also compared the consensus sequences of the catalogs of loci generated by RADProc and *cstacks* for the default Stacks settings i.e., $M = 2$, $m = 3$ and $n = 1$. The catalog generated by RADProc contained 91,825 loci and the *cstacks* program in Stacks built a catalog of 105,424 loci. Out of the 105,424 loci in the catalog built by Stacks and 91,825 loci in RADProc catalog, 86,375 loci were common between the two methods. Aligning the catalog locus consensus sequences from Stacks and RADProc to the green crab reference genome, 75,431 loci (71.5%) from the Stacks catalog and 68,125 loci (74.2%) from the RADProc catalog uniquely mapped to the reference genome.

The *populations* program filtered the catalog loci according to the parameter settings defined on Page 31 and retained 14,898 and 12,267 loci from the RADProc and Stacks catalogs respectively. RADProc had 22% more catalog loci that passed the stringent population filters than Stacks. Comparing the pairwise F_{st} values between the 11 sites (Table 2.3), the F_{st} values obtained by RADProc differed from Stacks by 0.82% on average, with a minimum and maximum of 0.04% and 3.14% respectively. Out of the 14,898 retained RADProc catalog loci, 2919 were private alleles, while 1942 out of 12,267 Stacks catalog loci were private alleles. Figure 2.10 shows the distribution of the private alleles in each of the 11 populations for RADProc and Stacks.

For the RADSeq data from the brown trout dataset of 16 samples, RADProc identified 364,611 unique stacks. We eliminated unique stacks that did not satisfy $S = 0.10$ and $D = 2$, which left 227,796 unique stacks for further analysis. From the run-times recorded for the 32 different parameter-value combinations of M and m (Table 2.2) using the Stacks modules *ustacks* and *cstacks.* processing all 16 samples for 32 parameter sets *ustacks* took approximately 252 hours and cstacks took approximately 11 hours, so in total Stacks required 263 hours. On the other hand, RADProc required only 23 hours to perform *de novo* locus formation and catalog building for all 32 parameter sets. Comparing the *de novo* loci formed by RADProc and Stacks, on average approximately 96% of the loci formed were common between RADProc and *ustacks* and RADProc produced approximately 7% fewer *de novo* loci on average than *ustacks* across all the parameter combinations of different values of M and m. We have also compared the consensus sequences of the catalogs of loci generated by RADProc and *cstacks* for the default Stacks settings i.e., $M = 2$, $m = 3$ and $n = 1$. The catalog generated by RADProc contained 149,381 loci and the cstacks program in Stacks built a catalog of 175,242 loci. Out of the 175,242 loci in the catalog built by Stacks and 149,381 loci in RADProc catalog, 124,045 loci were common between the two methods.

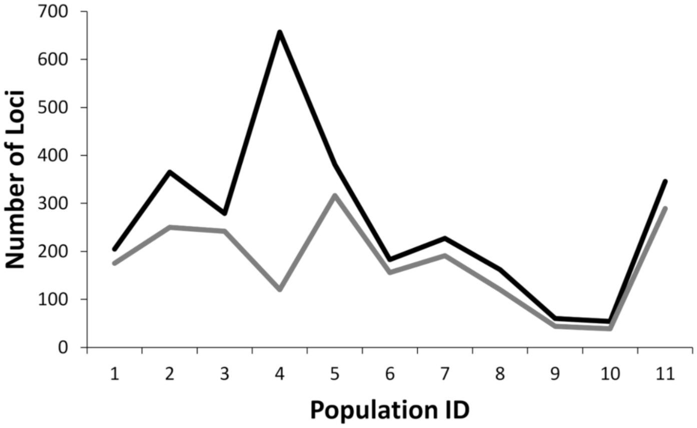

Figure 2.10: Private alleles retained by Stacks vs RADProc. The number of private alleles retained by the populations program using the *de novo* loci formed ($M = 2$ and $m = 3$) and catalog built ($n = 1$) for the 242-sample dataset using RADProc (black) and Stacks (grey) are plotted.

2.4 Discussion

RADSeq has enabled simultaneously examining tens of thousands of genetic loci for hundreds of individuals for a variety of ecological and evolutionary applications. Available tools for processing RADSeq data are useful in *de novo* loci assembly of the RAD-tags and genotyping those loci for evolutionary analysis. However, the parameter settings during *de novo* formation can significantly affect the analytical results [74]. The uncertainty in choosing the optimal minimum sequence depth required to consider the reads as potential alleles and the maximum distance allowed between such reads to be merged into candidate loci requires running the program multiple times and comparing the results to determine the optimal values for these parameters. The optimal values are determined based on the number of loci formed and the number of SNPs identified per sample (Figure 2.9). In practice, this could be achieved by using a smaller dataset to sweep through the parameter values. But as shown in the run-time comparisons above, trying different parameters could be

	BDB	BRN	CBI	CLH	KJI	MBO	NWH	PLB	SGB	SYH	TKT
BDB		0.02	0.05	0.01	0.02	0.01	0.05	0.03	0.01	0.01	0.05
BRN	0.02		0.05	0.02	0.02	0.02	0.06	0.03	0.02	0.02	0.05
CBI	0.05	0.05		0.05	0.04	0.05	0.01	0.03	0.06	0.06	0.01
CLH	0.01	0.02	0.05		0.02	0.01	0.05	0.03	0.01	0.01	0.05
KJI	0.02	0.02	0.04	0.02		0.02	0.04	0.02	0.02	0.02	0.04
MBO	0.01	0.02	0.05	0.01	0.02		0.05	0.03	0.01	0.01	0.05
NWH	0.05	0.06	0.01	0.06	0.04	0.05		0.03	0.06	0.06	0.01
PLB	0.03	0.03	0.03	0.03	0.02	0.03	0.03		0.03	0.03	0.03
SGB	0.01	0.02	0.06	0.01	0.02	0.01	0.06	0.03		0.01	0.06
SYH	0.01	0.02	0.06	0.01	0.02	0.02	0.06	0.03	0.01		0.06
TKT	0.05	0.05	0.01	0.05	0.04	0.05	0.01	0.03	0.06	0.06	

Table 2.3: Pairwise F_{ST} values between the 11 sites generated by the populations program using the *de novo* loci formed ($M = 2$ and $m = 3$) and catalog built ($n = 1$) for 242 green crab samples dataset using RADProc (above black diagonal) and Stacks (below black diagonal).

extremely time-consuming for even a small dataset of 20 samples. RADProc reduces the number of unique reads per sample by retaining only the unique reads that are present in a minimum number of samples and also have an average minimum coverage depth across all the samples in which it is present. RADProc accelerates the process by eliminating redundant calculations and using a faster nucleotide distance calculation method. The acceleration is observed across the two tested datasets, which suggests that it will work similarly well for other datasets too.

RADProc was able to calculate the nucleotide distances for increasing values of M in similar run-times (ranging from 48 seconds to 1 minute for 50,000 sequences). By contrast, the run-times for pairwise sequence comparison and k-mer counting methods increased as the value of M increased. The proposed method out-performed the other methods as the number of sequences increased as well. The graph data structure stores all the unique reads across all samples in the dataset and connects them if they are within a maximum nucleotide distance threshold, thus eliminating the need for reformation of unique reads when the value of m changes and the need for recalculating the nucleotide distance when M changes and reducing the run-time.

The constant time required for each parameter combination after building the graph with RADProc contrasts with run-time for Stacks, which increases as the value of M increases.

Although RADProc is efficient in processing the RADSeq data and decreases the run-time manyfold in comparison to Stacks, RADProc also depends on reducing the number of unique stacks using the abundance (S) and average coverage depth (D) based filtering of unique stacks. In our evaluation, even with the S and D based filtering of unique stacks there was a large proportion of common loci between the two methods. More lenient values of S and D, could increase the proportion of common loci between the two methods, as RADProc would be filtering more number of unique stacks, which are retained by Stacks.

2.5 Summary

RADProc is a RADSeq data processing software package that can be used for *de novo* locus formation and genotyping the formed loci. RADProc is different from other RADSeq tools by providing options to sweep different parameter set values for *de novo* locus formation and catalog building. By accelerating nucleotide distance calculations and separating distance calculation from locus inference and catalog construction, RADProc can efficiently process large RADSeq datasets containing several hundred samples or more. This is highly advantageous considering the uncertainty in choosing the parameter values for *de novo* locus formation and the need to process large datasets for population studies. By accelerating locus formation and catalog building, RADProc allows rapid processing of large RADSeq datasets, and efficient evaluation of different parameter combinations to identify suitable values for analysis.

Chapter 3

PMERGE: Computational filtering of paralogous sequences from RADSeq data

3.1 Introduction

Duplication events at the gene, chromosome or genome level [53] can create two or more paralogous DNA sequences from a single ancestral sequence and complicate genome assemblies and estimation of genetic variations [27]. The identification of loci and true SNPs from short sequences remains especially challenging for species with a duplicated genome because the duplicated sequences can be wrongly merged into a single locus, causing difficulty in identifying true allelic variations [50]. A well-known example is that of the salmonid fishes, which underwent a whole-genome duplication event approximately 80 Million Years Ago (MYA) [71, 68]. In the absence of a reference genome, there are several methods available to filter paralogous loci from the genome data. Some approaches augment the genetic data with other information, such as linkage mapping based on pedigrees [106], and removing heterozygous SNPs from double-haploid individuals [69]. Computational filtering approaches rely solely on the DNA sequence data, and can be done either during assembly and genotyping [31, 33] or on the assembled data, for example, by retaining only those loci with the expected number of alleles and by retaining only those putative loci whose inferred genotype frequencies conform to Hardy-Weinberg equilibrium (HWE) expectations [50, 17]. For populations in HWE, the expected heterozygosity can never be more than 0.50 at any bi-allelic locus. HDplot [76] uses read depths and excess heterozygosity to identify putative paralogs. HDplot works by plotting the relative proportion of heterozygotes in a population (H) and the deviation of allele-specific reads of each locus from a 1:1 ratio (D). One more approach is haplotyping [109], which relies on the fact that closely linked SNPs can constitute haplotypes of which a diploid individual can have no more than two.

An important challenge in distinguishing paralogs is the choice of percent identity used to delineate loci; typically, a value of 98-99% among reads is used (e.g., [18, 70]). However, a stringent similarity threshold carries the risk of splitting divergent alleles into separate loci ("over-splitting") if the orthologs differ by an amount greater than the similarity threshold, whereas lower similarity thresholds can allow paralogous sequences to be incorrectly merged into one orthologous locus ("under-splitting") [45, 91]. Stacks identifies the erroneously merged sequences and tries to break them into multiple loci using a deleveraging algorithm, which calculates a minimum-spanning tree out of the graph representing the locus, using the stacks from the locus as nodes and the distance between them as edge weights. A minimum-spanning tree requires that there is only one path between each pair of nodes in the graph (no cycles) and that all edges in the graph are of minimal weight. The graph is traversed and the edge weights are recorded and ordered. The graph is sheared at all edges with weight equal or greater than the second smallest edge weight. However, if the erroneous locus is formed from only 2 or 3 paralogous stacks, it will not be considered an over-merged locus.

Given the potentially confounding effects of paralogous loci, new methods are needed to identify them and allow removal prior to the inference of population-level statistics. Here we describe PMERGE, a new method that identifies candidate paralogs or duplicated loci in the catalog loci built by the Stacks program. PMERGE works by building networks of catalog loci that share high levels of nucleotide similarity and flagging highly similar sequences as potential paralogs. Our approach is able to successfully identify the majority of paralogous loci generated from a RADSeq analysis of two species, first 150 sampled Atlantic salmon (*Salmo salar*); and second 242 green crab (*Carcinus maenas*) samples. By embedding PMERGE in the analysis pipeline of the widely used Stacks software [17], it is straightforward to apply it as an additional filter in population-genomic studies using RADSeq data and can also be used in addition to other existing approaches.

3.2 Methods

3.2.1 Identification of putative paralogs using PMERGE

The PMERGE software (Figure 3.1) is run after *sstacks* and before *populations* to generate a "whitelist" of loci from the catalog based on population-level filtering conditions and our new paralog-detection method. The populations program then uses only the whitelisted loci to generate population-genetic statistics. Apart from the paralog filter, PMERGE includes the following filters that are also used by the populations program: percent samples limit per population (r), which requires that a locus be present in at least the specified percentage of individuals in a population; locus population limit (p), the minimum number of populations in which a locus must be present; minor allele frequency cutoff (a), which sets a minimum threshold for the frequency of the minor allele (the second-most-frequent allele at a given locus) for each SNP in a locus; maximum observed heterozygosity (q) for each SNP in a locus; and minimum stack depth (m) at a given locus.

Paralogous sequences that have arisen from recent duplication events will exhibit high similarity with more than one region in the genome. The catalog built from the *de novo* loci formed from each individual gives us a pool of loci from all the individuals in the population, and our hypothesis is that highly similar groups within the pool of *de novo* loci have a high probability of being derived from multiple sites

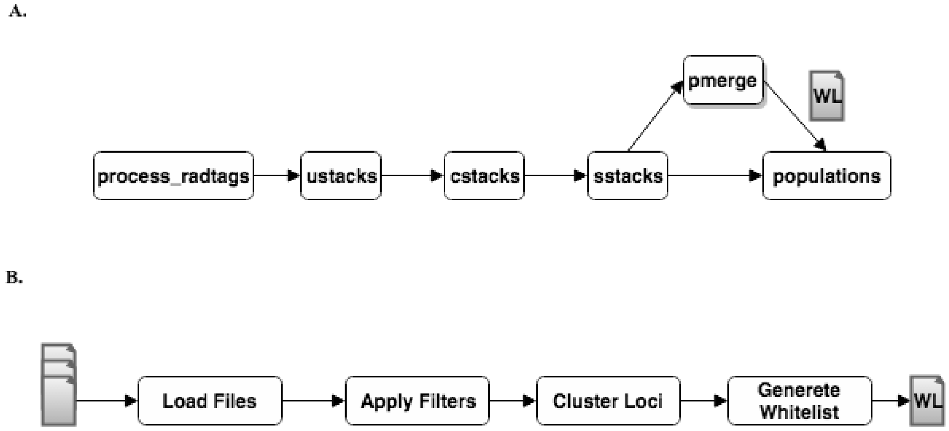

Figure 3.1: Use of PMERGE in the Stacks workflow to generate whitelisted (WL) loci. (A) The modified Stacks pipeline with PMERGE invoked immediately after sstacks is used to search loci against the reference catalog. The populations program can use the whitelist file generated by the PMERGE module to include only those whitelisted loci. (B) The PMERGE module reads the catalog, tags, SNP, and match files, applies user-defined filters and performs clustering to identify and eliminate paralogs. The retained loci are then written to a whitelist file.

in the reference genome. PMERGE is applied to the catalog of loci and not to the *de novo* loci formed in each sample separately: the paralogs may be merged into a single locus in some but not all samples, allowing us to cluster them and identify based on similarity. PMERGE flags the polymorphic (heterozygous) catalog loci that are clustered with at least one other catalog locus.

To identify probable paralogs, we construct a graph or network where each node corresponds to a locus, which is represented by its consensus sequence. The consensus sequence represents the sequence of major allele nucleotides at each position in the locus sequence. For efficiency, we represent all sequences of a given locus with a consensus sequence, which greatly simplifies the network; in practice this reduction has minimal impact on the inference of paralogs. Using the consensus sequences can allow us to capture the cumulative mismatches of all the sequences in the locus. PMERGE uses a more-lenient similarity threshold than M for clustering the catalog locus consensus sequences, which allows us to identify loci that are less similar than their constituent stacks but similar enough to be flagged as duplicates.

A cluster similarity threshold parameter C is set (default value 90%), and all pairs of loci whose representative sequences that are C% similar are connected with an edge. Sets of loci that are connected by at least one path in the network define connected components; each of these components is interpreted as a putative set of paralogous sequences. These sequences can then be removed from the dataset prior to calculation of population parameters, or set aside for further analysis.

PMERGE subdivides catalog loci into probable paralogs by (i) determining the consensus sequences from the catalog loci, (ii) buliding a network that connects the consensus sequences within a similarity threshold C, and (iii) flagging all catalog loci with consensus sequences clustered with at least one other consensus sequence as probable paralogs.

3.2.2 Validation of the proposed method

We validated the performance of PMERGE on an Atlantic salmon data set (see [13])and a green crab data set (see [58]). For reference comparisons, polymorphic catalog loci consensus sequences were aligned to version ICSASG_v2 of the Atlantic salmon reference genome [68] and the Green crab reference (Hleap et al., in preparation) using BLASTN version 2.2.28 [4] with a minimum of 90% sequence identity and a maximum E-value of $1e - 20$.

The Atlantic salmon data set contained RADSeq data obtained from 150 individuals from 15 different locations along the south coast of Newfoundland, Canada. The dataset comprised samples with approximately 2,500,000 to 14,000,000 RAD-tags per individual trimmed to 80bp. The genomic DNA was digested using restriction enzyme SbfI, and the resulting fragments were sequenced. Individually barcoded RAD samples were jointly sequenced on the Illumina GAIIx platform with single-end sequencing 100-bp chemistry. The Atlantic salmon genome underwent a whole genome duplication (WGD) event 80 million years ago and is in the process of reverting to diploid state (Ohno et al. 1968; Allendorf and Thorgaard 1984; Macqueen and Johnston 2014; Lien et al. 2016). As such, salmon is a good choice for validation of our proposed method, as there is a reference genome available that allows us to verify the majority of our predictions by mapping loci back to the genome. We evaluated the number of paralogs identified by PMERGE with the Stacks *de novo* assembly similarity parameter M set to 2 and 4 in separate runs. The parameter C was varied from 90% to 50% in intervals of 10%, to compare the number of paralogs identified. [56] recommended an M value of 2 to reduce the merging of putative paralogs into one locus. *De novo* locus formation with $M = 4$ was also done to demonstrate the effect of over-merging in identifying the paralogs using the proposed approach. The catalog of loci was built using *cstacks* with the maximum nucleotide distance allowed between catalog loci to merge $n = 1$. The resulting catalog of assembled *de novo* loci was passed to our filtering software, with parameter settings $a = 0.05$, $p = 12$ and $r = 0.75$.

The validation involved identifying the efficiency of PMERGE in correctly identifying and removing the paralogs in both data sets by aligning the identified paralogs to their corresponding reference genomes. Firstly, all the polymorphic catalog loci were aligned to their reference genome and the loci with multiple hits to the reference genome flagged as candidate paralogs. Secondly, constituent alleles of each polymorphic catalog locus were aligned to the reference genome and alleles mapping to different regions in the reference genome were identified to flag the wrongly merged paralogs in the catalog. By comparing the polymorphic catalog loci flagged as paralogs by PMERGE with the candidate paralogs and the candidate PSVs the proportion of duplicated loci and PSVs identified by PMERGE was determined.

The impact of the parameter M on the results was also examined. The number of paralogous loci and PSVs identified by PMERGE was also compared with HDplot and filtering by deviations from H-W expectations. The populations program in Stacks generates a VCF format output that was used as input for HDplot. The deviations from H-W equilibrium were analysed using VCFtools, which can use the VCF format output from Stacks. Loci that significantly deviate from HWE ($p = 0.001$) were flagged as paralogs. A combination of PMERGE and the two approaches was also performed to evaluate the possibility of improvement in the proportion of paralogs detected.

The error rates for different values of C were calculated by determining the number of paralogs (polymorphic loci with multiple hits to the reference genome) correctly identified in the clustered loci, and the number of non-paralogous sequences that were rejected in the analysis (false positives); the error rate for a given C was calculated as the ratio of the number of false positives to the total number of clustered loci. We used ROC (Receiver Operating Characteristic) curves to assess the performance of PMERGE in filtering paralogs. In a ROC curve, the true-positive rates (in our case, detected paralogous loci) are plotted against the corresponding false-positive rates (single-locus alleles incorrectly classified by PMERGE as paralogous) for different values of a parameter, in our case C. The area under the resulting ROC curve (AUC) gives a measure of how well the method can distinguish between paralogous and non-paralogous loci.

One of the statistics calculated by the populations program in Stacks is pairwise F_{ST} values between all pairs of populations under study. Since paralogs can affect population divergence estimates [109], we compared the pairwise F_{ST} before and after paralog filtering by PMERGE on pairwise F_{ST}. Dendrograms were generated from the pairwise F_{ST} distance matrices obtained with and without application of the proposed filter for M=2 and C=90%. The generated pairwise F_{ST} distance matrices were clustered using the “hclust” function in the R package “stats” [88], which uses an agglomerative hierarchical clustering approach to construct relationships among different populations. In this analysis, the pairwise F_{ST} distance between populations was used as the distance metric, with clusters constructed based on the average-linkage criterion, where the distance between two clusters of populations is defined as the average pairwise F_{ST} distance between each of their populations. The dendrograms created were mapped to the actual geographical locations using GenGIS [83]. Differences in the topologies of the dendrograms created before and after PMERGE filtering were evaluated by calculating the Robinson-Foulds distance (RF: [90]), as implemented in T-REX [11] and the rooted subtree prune-and-regraft (rSPR: [47, 108]) distances. The RF distance is the measure of number of bipartitions in one tree that are absent in the other tree. Migration of a single branch to a different part of the tree can affect many bipartitions, which inflates the RF distance and may overemphasize the distance between the two corresponding trees. An SPR operation cuts a subtree from the rest of the tree and reattaches it in a different location. The rSPR distance between two trees is equal to the minimum number of SPR operations required to reconcile two rooted trees, and is influenced less strongly by single branch migrations. RF and SPR therefore provide two contrasting views of tree similarity.

We also tested the ability of PMERGE to detect paralogs in a species that has no historical genome duplication. The green crab (Carcinus maenas) dataset consists of RADSeq data extracted from 242 individuals from 11 locations in eastern North America. Each library consists of 22 samples identified by variable length in-line barcodes ranging from 5 to 9 bases. The libraries were sequenced on a HiSeq 2000 (Illumina) as 100 bp paired end sequences. The dataset comprises samples with approximately 3,000,000 RAD-tags per individual trimmed to 80bp. The RADSeq data from each individual sample were cleaned, demultiplexed and *de novo* assembled using the default Stacks parameters $M = 2$ and $m = 3$. The catalog of loci was built using cstacks with maximum nucleotide distance allowed between catalog loci to merge $n = 1$. The resulting catalog of loci was then filtered using PMERGE with the parameter settings $a = 0.05$, $p = 11$ and $r = 0.75$. In separate runs, the parameter C was varied from 90% to 40% with intervals of 10%, to compare the number of paralogs identified. Contrasting the Atlantic salmon genome, the absence of recent whole-genome duplication in green crab lowers expectations of the prevalence of paralogs. The inclusion and comparison of both species allows the utility of PMERGE to resolve paralogs under two very different contexts to be evaluated.

3.3 Results

We examined the effectiveness of PMERGE on data sets from two different species with distinct evolutionary histories. Both species have reference genomes available, which allow validation of paralogs predicted by PMERGE. First, we examined a set of RADSeq data from Atlantic salmon (*Salmo salar*), which has a recent (80 MYA) whole-genome duplication and consequently a large proportion of expected paralogs. The analyses included identifying the impact of filtering paralogs on the inferred population structure and random subsampling of loci to show that the differences in population structure after filtering the paralogs is not random. We also examined the removal of paralogs from a European green crab (*Carcinus maenas*) dataset, which was first used to study their population structure in Northwest Atlantic [58] and has no known historical genome duplication.

3.3.1 Atlantic salmon analysis

For $M = 2$, after applying the filters (see methods), 25,209 polymorphic catalog loci were retained and alignment of the locus consensus sequences to the Atlantic salmon reference genome using BLASTN revealed that 13,510 of these 25,209 were putative paralogs and mapped to multiple locations in the genome. Similarly, aligning the constituent allele sequences from the 13,510 catalog loci revealed that 4,852 loci (36%) had their allele sequences mapped to multiple regions in the reference genome. Out of the 13,510 putative paralogs, 5447 (40%) were unplaced and 8063 were chromosome-positioned (Figure 3.2). Approximately 36% of the 8063 loci mapped to the homeologous blocks with high similarity (>90%) and 52% of the 8063 loci are from the other homeologous blocks specified in [68].

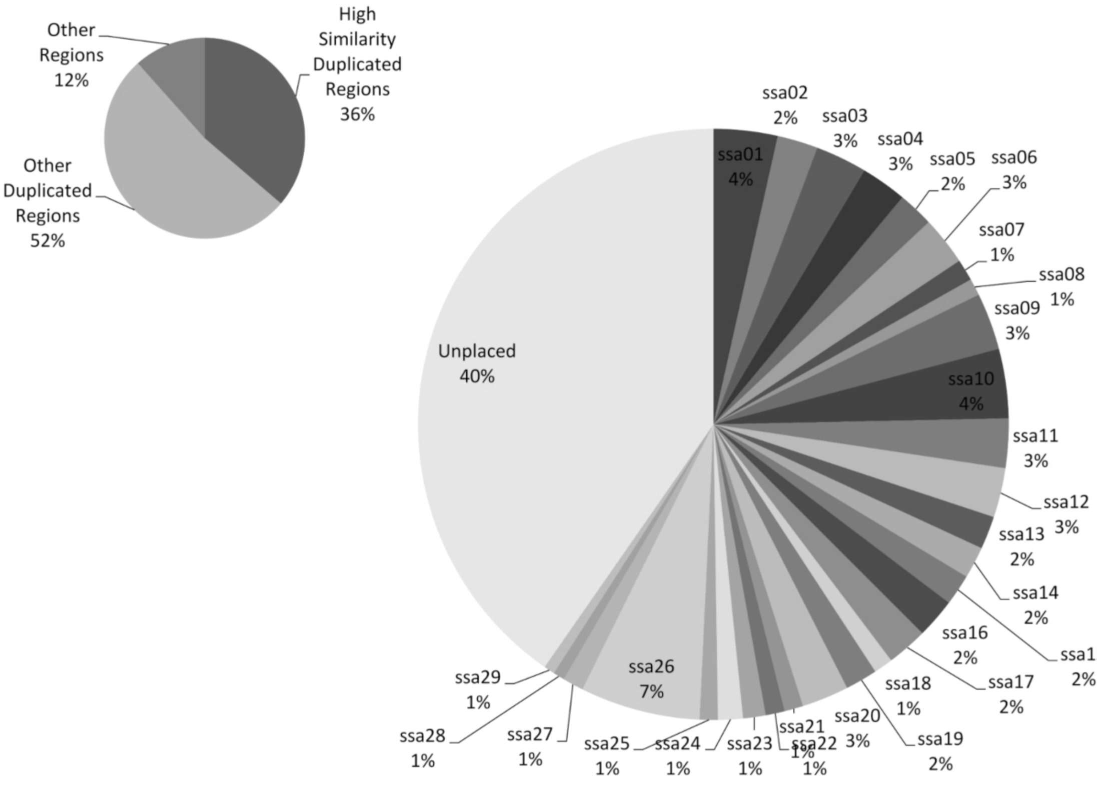

Figure 3.2: Distribution of putative paralogs (chromosome-positioned) with respect to the chromosome regions. The putative paralogs are flagged from the catalog formed from Atlantic salmon data set with *de novo* locus formation using $M = 2$ by aligning to the reference genome.

As Figure 3.3 shows, at $C = 90\%$, out of the 25,209 loci, 8,226 (32.63%) were clustered by PMERGE (i.e., potential paralogs) and 16,983 loci remained non-clustered.

Of the 8,226 clustered loci, 8,214 (99.85%) mapped to multiple locations in the reference genome. Reducing C to 80% increased the number of loci clustered to 10,667 with 10,268 (96.26%) loci mapping to multiple locations in the genome. The cluster similarity threshold C was varied between 50% and 90% at intervals of 10% and the error rates recorded. Approximately 81% and 85% of the total putative paralogs were identified by $C = 70\%$ and $C = 60\%$, respectively. From $C = 90\%$ to $C = 60\%$, the error rates varied from 0.01 to 0.10. At $C = 50\%$, all the 25,209 polymorphic loci were flagged as paralogs by PMERGE. On the other hand, using the HDplot approach, loci with the proportion of heterozygous individuals (H) > 0.6 and read-ratio deviation (D) between -7 and 7 were flagged as paralogous. The HDplot approach identified 1996 loci as paralogs, out of which 167 loci uniquely mapped to the reference genome (false positives). The HWE filter identified 2,499 loci as paralogs, in which 566 loci were false positives. Approximately 36% and 45% of paralogs flagged by HDplot and deviations from HWE overlapped with paralogs identified by PMERGE ($C = 60\%$), respectively. PMERGE identified 1938 loci with wrongly merged PSVs at $C = 90\%$, and the proportion increased as the similarity threshold C decreased. PMERGE identified a maximum of approximately 60% of the 4,852 merged PSVs at $C = 60\%$. HDplot identified 31% and deviation from HWE identified 30% of the loci with wrongly merged PSVs.

When $C = 90\%$, out of the 4573 chromosome-positioned loci, 1211 (26.5%) were from high-similarity duplicated regions and 2649 (58%) were from other duplicated regions. Reducing C to 80%, out of the 5667 chromosome-positioned loci, 1494 (26.4%) mapped to high-similarity duplicated regions and 3340 (59%) mapped to other duplicated regions. At $C = 60\%$, a similar trend was observed, with 1943 (28%) and 4111 (59%) of the 6981 chromosome-positioned loci from high-similarity duplicated regions and other duplicated regions respectively. From $C = 90\%$ to $C = 60\%$, approximately 12% to 16% of the loci were unplaced in the chromosome.

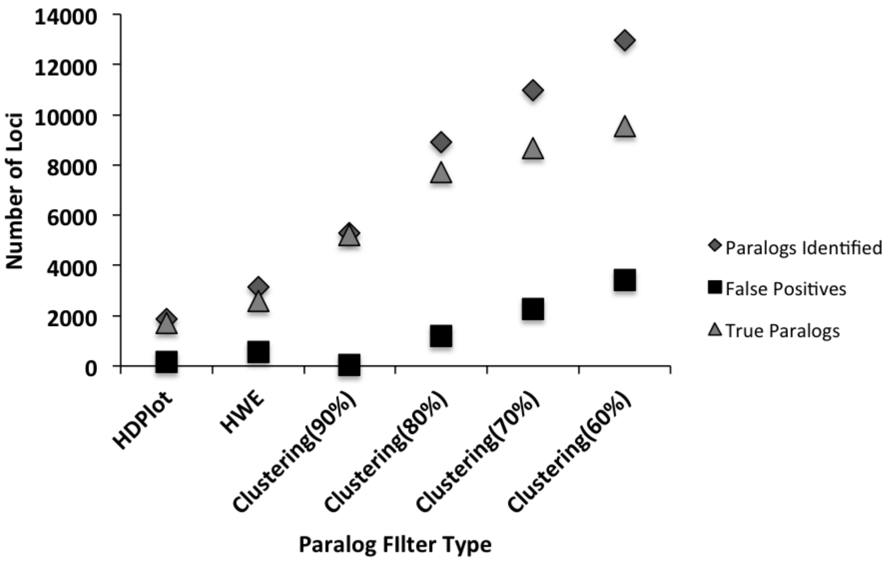

Figure 3.3: Comparing the effectiveness of the HDplot, HW approach and PMERGE for Atlantic salmon data set with *de novo* locus formation using $M = 2$. Showing the number of putative paralogs identified, false positives and true paralogs identified by the HDplot, HW approach and PMERGE with different settings of cluster similarity C.

Approximately 60% of the 8226 loci that clustered at $C = 90\%$ mapped to exactly two locations in the reference genome. Figure 3.4 shows the distribution of the loci mapped exactly to two locations with respect to the number of mismatches, including gaps. About 43% of the clustered loci that mapped exactly to two locations in the reference genome are part of clusters of size 2, and 71% of the loci with exactly two hits belong to clusters of size from 2 to 20. This illustrates the correlation between the size of the clusters and the number of matching locations in the genome.

Combining PMERGE with HDPlot or deviation from HWE methods increased the proportion of paralogs and loci with wrongly merged PSVs identified (Figure 3.5). We observed an approximately 8% to 10% increase in the putative paralogs identified (Figure 3.5A) and a 22% to 26% increase in the loci with merged PSVs detected (Figure 3.5B). At $C = 60\%$, using only PMERGE we were able to identify 85% of the putative paralogs and 60% of loci with merged PSVs, whereas combining PMERGE

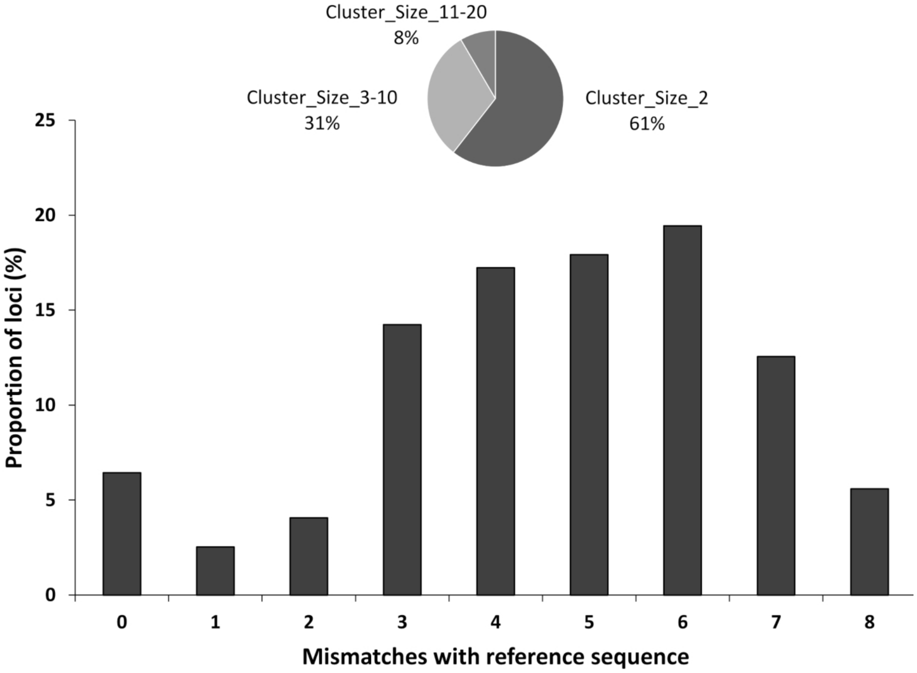

Figure 3.4: Distribution of filtered loci when C = 90% that mapped exactly to two locations in the reference genome based on number of mismatches and on cluster size.

with HDPlot or HWE approaches we were able to detect 93% of the putative paralogs and 81% of loci with merged PSVs.

When $M = 4$, 25,775 polymorphic loci were retained in the catalog for further analysis after applying filters. 12,473 loci out of the 25,775 mapped to multiple locations in the reference genome. Aligning the constituent alleles from the 12,473 loci revealed 4,316 loci (35%) had allele sequences that mapped to different regions in the genome. At $C = 90\%$, out of the 25,775 loci, 5254 (20.38%) were clustered by PMERGE (i.e., potential paralogs) and 16,983 loci remained non-clustered. Of the 5254 clustered loci, 5239 (99.71%) mapped to multiple locations in the reference genome. Reducing C to 80% increased the number of loci clustered to 7963 with 7548 (94.79%) loci mapping to multiple locations in the genome. The error rates ranged from 0.01 to 0.13 for $C = 90\%$ to $C = 60\%$, identifying 42% to 72% of the total paralogs respectively. Using the HDplot approach, loci with the proportion of

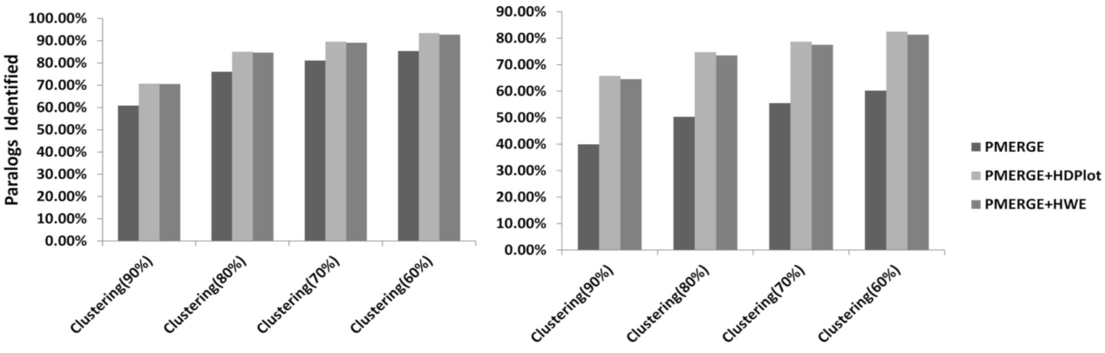

Figure 3.5: Paralogous loci identified by using only PMERGE and in combination with HDPlot and HWE approaches using Atlantic salmon data set with *de novo* locus formation using M=2. A. Putative paralogs idenitified. B. Loci with merged paralogs detected.

heterozygous individuals (H) > 0.6 and read-ratio deviation (D) between −7 and 7 were flagged as paralogous. The HDplot approach identified 1880 loci as paralogs, out of which 124 loci uniquely mapped to the reference genome. The HWE filter identified 3,143 loci as paralogs, out of which 862 loci were false positives. Approximately 32% and 36% of paralogs flagged by HDplot and deviations from HWE overlapped with paralogs identified by PMERGE (C = 60%), respectively. Approximately 22% and 32% of paralogs flagged by HDplot and deviations from HWE overlapped with paralogs identified by PMERGE (C = 60%), respectively. PMERGE identified a maximum of approximately 50% of the 5,444 loci with merged PSVs at C = 60%. Both HDplot and deviation from HWE identified 28% of the loci with merged PSVs.

The ROC curve obtained using different values of C for M = 2 (Figure 3.6A), the AUC was 0.92, which means PMERGE is good at separating paralogous loci from non-paralogous loci (Zweig & Campbell, 1993). Reducing the value of C below a certain limit leads to clustering of non-paralogous sequences (i.e., false positives). This is likely to happen because of short length of these sequences and the reduced amount of similarity required to cluster them. Also, the proposed method works based on only the similarity among the catalog of loci assembled from the set of samples,

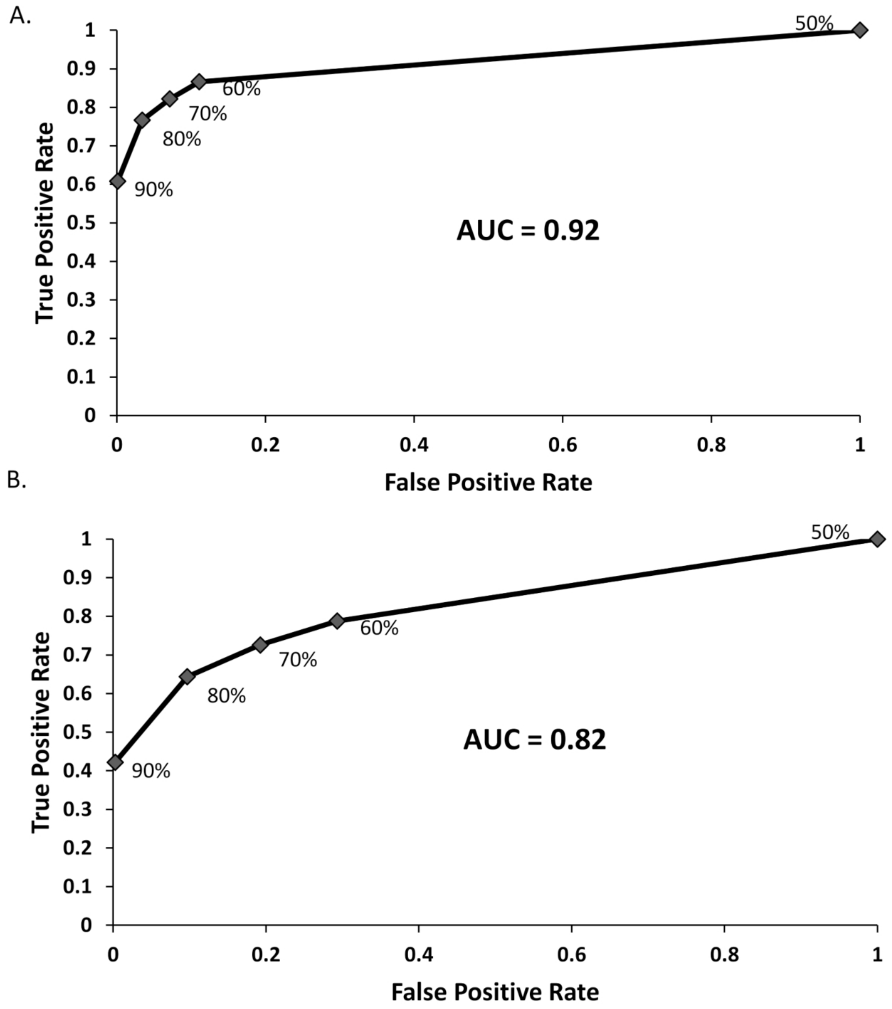

Figure 3.6: Identification of paralogous sequences by PMERGE. ROC curve generated using A. Atlantic salmon data set with *de novo* locus formation using $M = 2$, B. Atlantic salmon data set with *de novo* locus formation using $M = 4$ The ROC curves are generated from the observed true positives (paralogs), true negatives (non-paralogs), false positives and false negatives. The percentage labels on the curves are the similarity thresholds C used. The area under the ROC curve demonstrates the accuracy of the proposed method.

hence the number of paralogs identified is highly influenced by the proportion of similar loci available in the catalog. For $M = 4$ (Figure 3.6B), the AUC reduced to 0.82, indicating the accuracy of PMERGE in separating paralogous loci from non-paralogous loci reduces as the value of M increases.

3.3.2 Impact of paralog filtering on population structure

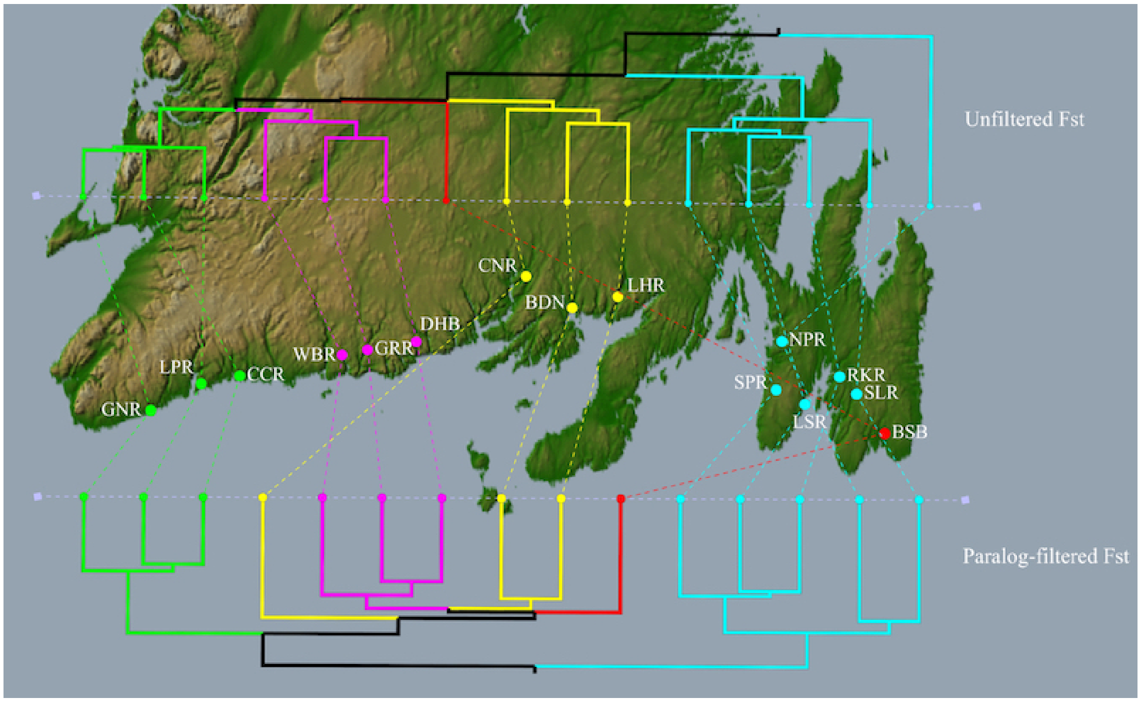

Figure 3.7: Dendrograms constructed from pairwise F_{ST} values between sites, before (top) and after (bottom) paralog filtering, with dendrogram leaves assigned to the sampled geographical locations along the Southern coast of Newfoundland. The C parameter was set to 90% and there were 40,618 loci before paralog filtering and 32,392 loci after paralog filtering.

Pairwise F_{ST} values after applying the paralog filtering generally increased between the populations. For the site “NPR” the pairwise F_{ST} values with other sites generally decreased after applying the paralog filtering, except with the sites “BSB”, “LSR”, “RKR” and “SPR” where the pairwise F_{ST} values increased. While the percentage difference in the pairwise F_{ST} values after applying PMERGE filtering were as low as 0.96% between “NPR” and “LSR”, it was as high as 28.95% between “SLR” and “BSB”. Dendrograms obtained from the pairwise F_{ST} values generated between all pairs of populations under study before and after applying the PMERGE filter differed in topology. Figure 3.7 shows variations in subpopulation structures between the unfiltered and PMERGE-filtered trees: one notable pattern is the increased genetic differentiation between the east and west coast populations with the paralog-filtered data. In the paralog-filtered dendrogram, there are two major clusters separating the east and west coast populations, and the five populations “SPR”, “SLR”, “RKR”, “NPR” and “LSR” from the Avalon Peninsula in the east are grouped into one cluster. However, with the unfiltered data one of those five east coast populations (“NPR”) is an outlier in the generated dendrogram. The clustering of “NPR” with the rest of the populations from the Avalon peninsula is a result of the differences in their pairwise F_{ST} values after applying the PMERGE paralog filter, as opposed to using F_{ST} values obtained from the unfiltered data.

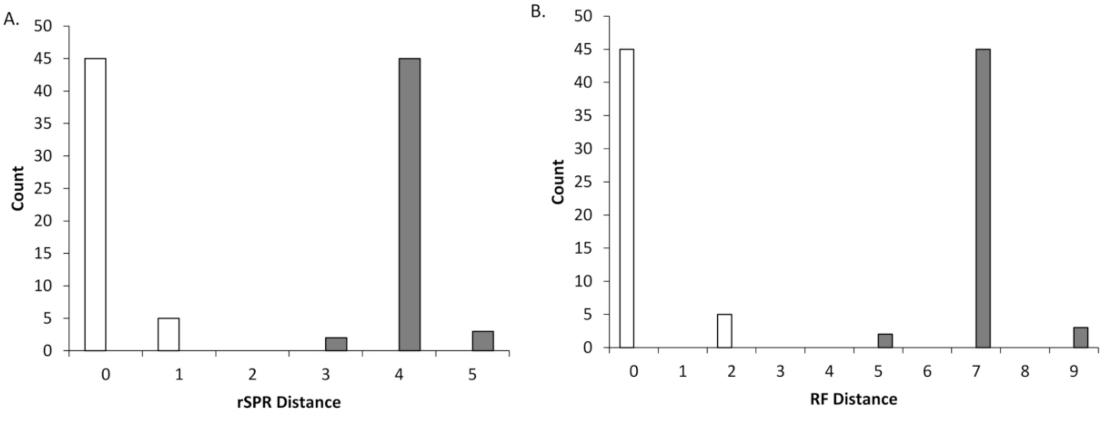

Figure 3.8: Distribution of rSPR (A) and RF (B) distances between trees constructed from randomly subsampled loci, and trees obtained from pairwise F_{ST} values before (white bars) and after (grey bars) PMERGE filtering. The RF distance is the measure of number of bipartition in one tree that are absent in the other tree and rSPR distance is minimum number of SPR operations required to reconcile two rooted trees.

RF and rSPR distances (Figure 3.8) calculated to compare the topology of the dendrograms were 10 and 4, respectively, indicating differences between the dendrograms. Since the PMERGE-filtered dendrogram was based on fewer loci than the unfiltered tree, we assessed the impact of choosing random subsamples of 32,392 loci from the unfiltered tree. Fifty replicate trees based on random subsamples of loci were constructed. If the effect of paralog filtering is greater than that of random subsampling, we expect that the paralog-filtered tree should differ more from the reference tree than do the dendrograms obtained from random subsamples. The rSPR distances between the unfiltered F_{ST} dendrogram and the dendrograms constructed from randomly subsampled loci were between 0 and 1, whereas the corresponding distances for the filtered F_{ST} dendrogram ranged between 3 and 5 (Figure 3.8A). The RF distances showed a similar trend but with distance values of 0 to 2 for unfiltered F_{ST} dendrogram and 5 to 9 for filteredF_{ST} dendrogram (Figure 3.8B).

3.3.3 Green crab analysis

In contrast with the Atlantic salmon genome, the green crab has no evidence of ancestral genome duplication; consequently, far fewer paralogs are expected. Assembling these RAD-tags using ustacks yielded approximately 25,000 loci per individual. The complete catalog contained 156,272 unique loci, which decreased to 12,435 by applying the locus filters as described in the Methods section. Of these 12,435 loci, 6,695 were polymorphic and alignment to the green crab reference (Hleap et al., in preparation) genome using BLASTN revealed that 913 of these 6,695 mapped to multiple locations in the genome (putative paralogs) and 360 loci with alleles that mapped to different locations in the genome (loci with merged PSVs). We have also compared the effectiveness of PMERGE with other approaches and performed ROC curve analysis (Figure 3.9). At $C = 90\%$, out of the 12,435 loci, 330 (32.63%) were clustered by PMERGE (i.e., potential paralogs) and 12,105 loci remained non-clustered. Of the 330 clustered loci, 307 (93%) mapped to multiple locations in the reference genome

(Figure 3.9A). Reducing C to 80% increased the number of loci clustered to 546 with 426 (78%) loci mapping to multiple locations in the genome. PMERGE identified a maximum of 62% of total paralogs and 37% of total loci with merged PSVs at $C = 60\%$. The error rates ranged from 0.07 to 0.40 for $C = 90\%$ to $C = 60\%$. The HDplot approach flagged 153 loci with proportion of heterozygous individuals (H) > 0.6 and read-ratio deviation (D) between -10 and 10 as paralogous. 50 out of the 153 loci mapped uniquely to the reference genome and 40 loci were PSVs (11% of the 360 loci). Using the HWE method, 963 loci were flagged as paralogs in which, 782 loci were false positives and 75 were merged PSVs (21% of the 360 loci). The AUC for the ROC curve obtained using different values of C was 0.71 (Figure 3.9B)

3.4 Discussion

The splitting of paralogous loci depends on the choice of maximum nucleotide distance parameter (M in the Stacks software); as M increases, paralogous loci are merged together [45, 91]. The putative paralogs flagged by PMERGE are the catalog loci with high sequence similarity. PMERGE identifies the wrongly merged PSVs by considering the entire catalog of loci constructed from all samples, rather than focusing on one sample at a time. If paralogs are merged into a single locus in one or more samples and not in others, the resulting pattern is used by PMERGE to properly subdivide loci.

Using the Atlantic salmon dataset we were able to assess the extent to which paralogs identified by PMERGE mapped to two or more genomic regions. When $M = 2$, 36% of the putative paralogs that mapped to chromosomes were situated in homeologous blocks with high similarity (>90%), and an additional 52% mapped to other homeologous blocks specified in [68]. Mapping the paralogs flagged by PMERGE for different values of C revealed that approximately 26% to 28% of the chromosome-positioned loci were from the high-similarity regions. Around 58% to 59% of the

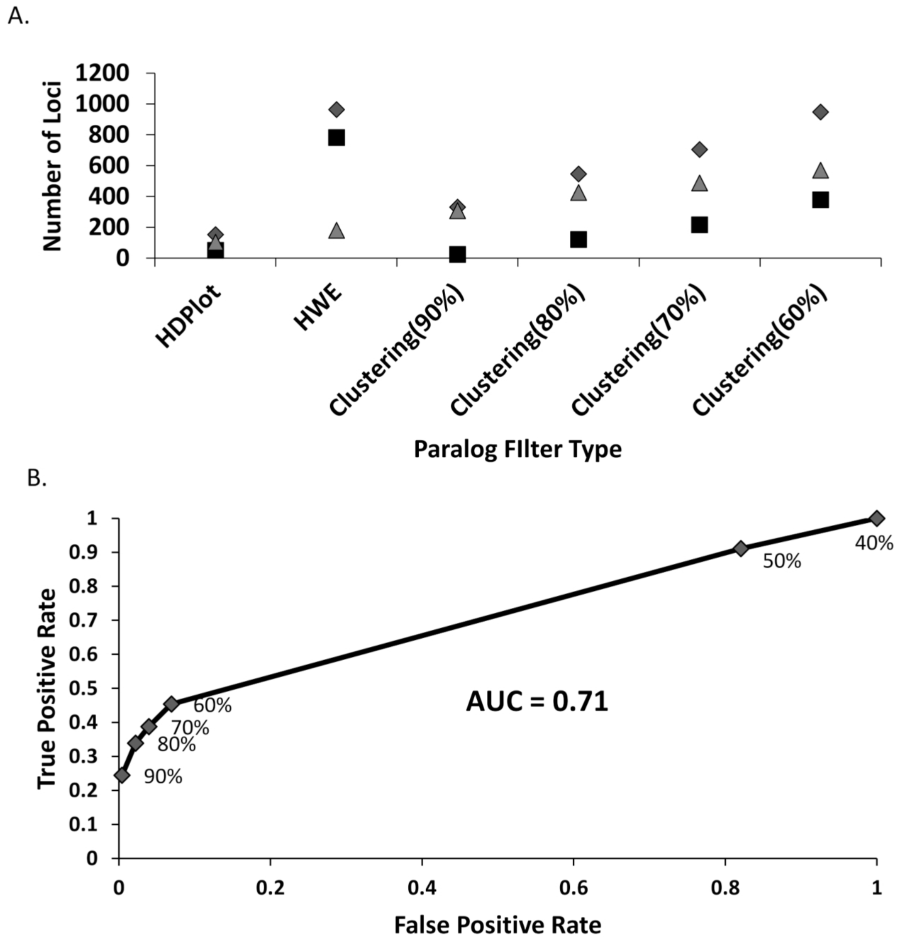

Figure 3.9: Paralog filtering in green crab data set using HDplot, HW approach and PMERGE. A. Comparing effectiveness of PMERGE and other methods. B. ROC curve generated using the observed true positives (paralogs), true negatives (non-paralogs), false positives and false negatives.

loci mapped to the less similar duplicated regions. Since the sequences used in the analysis are as short as 80 bp, we see high similarity among them even though they are from less similarity duplicated regions and wrongly merged into a locus.

Comparing the paralogs identified by HDplot, deviations from HWE and PMERGE with different values of C, it is evident that PMERGE identifies more paralogs and merged PSVs than the other two methods. HDplot and deviations from HWE focus on identifying the merged PSVs by analysing individual polymorphic loci, whereas PMERGE identifies paralogous loci in the catalog using their similarity. While HD-Plot and HWE tests are applied to the VCF format output generated by the *populations* program in Stacks, PMERGE is applied to the catalog loci before the *populations* program even processes them. Most of the PMERGE flagged loci with merged PSVs were unique to PMERGE. Combining PMERGE with the other two approaches increased the proportion of paralogs detected. Approximately 7% of the paralogous loci and 19% of the wrongly merged PSVs were not detected by any of the three approaches. Since PMERGE cannot subdivide loci that are merged across all samples, the best use case, explored above and worthy of further development, is to combine the PMERGE approach with other methods such as HDplot which can examine distributional patterns within loci from even a single sample.

By observing the number of loci clustered for different values of C, we can identify an optimal cut-off value for this parameter, depending on the species and dataset. The accuracy of PMERGE analysed using AUC obtained from ROC curves suggests that PMERGE can best perform when the species has more duplicated regions and the M value used in Stacks is low. The AUC obtained for the Atlantic salmon data at $M = 2$ was 0.92 and for $M = 4$ it was 0.82, whereas we obtained an AUC of 0.71 for the green crab dataset. Unlike the Atlantic salmon, the green crab does not have large proportions of highly similar regions in the genome. Hence the accuracy of PMERGE in separating paralogous loci from non-paralogous loci is less than salmon data. For large values of M and species with less duplicates, set the similarity threshold C to high values (more than 80%) and for smaller values of M and species with genome duplication, set C as low as 60% .

Applying PMERGE with a C value as high as 90% eliminated at least 61% of paralogous loci and 40% of the loci with wrongly merged paralogs. The resulting population structure is more consistent with the previous study by [13] involving microsatellites, SNP arrays and RADSeq data from southern Newfoundland, that also showed strong evidence of subdivision of salmon populations into eastern and western groups. In the RADSeq data used for their analysis the PSVs were eliminated by removal of SNPs with three or more alleles as well as SNPs that mapped to multiple locations in the reference genome [27]. As expected, the populations were clustered into two large east-west groups. Analyzing the dendrogram obtained without applying PMERGE filtering, the "NPR" population was unusually distinct, contradicting the results obtained in previous studies. The RF and rSPR distance comparisons between the paralog-filtered dendrogram, the dendrogram obtained without applying paralog filtering and the dendrograms obtained from the random subsample showed that paralog filtering applied using PMERGE has a significant non-random effect on the topology of the pairwise F_{ST} dendrogram.

3.5 Conclusion

We have demonstrated the effectiveness of PMERGE in filtering paralogous loci from two species with different genome structures. Depending on the species under study and the expected proportion of paralogs, different values of C may be examined for an optimal value based on the proportion of detected paralogs. Also, for non-model species we will not know the expected proportion of paralogs and in that case the best option will be to set high values for C. The results from the Atlantic salmon and green crab datasets show that we can detect large number of paralogs even with high values of C using PMERGE.

Chapter 4

A novel method to infer relative gene flow among populations using exclusively shared alleles.

4.1 Introduction

Gene flow (often used synonymously with "effective migration") between populations can introduce new alleles into either or both populations and change their respective allele frequencies. Gene flow is the result of both dispersal of individuals in space and the successful reproduction of the migrants [99]. Gene flow can be restricted by physical barriers separating the populations and also by incompatible reproductive behaviours between the individuals of the populations [20]. Gene flow between genetically dissimilar populations can reduce the genetic difference between the populations by increasing homogeneity and reducing the degree of genetic differentiation [12]. Gene flow may lead to interbreeding between individuals from genetically differentiated populations (hybridization) and incorporation of new alleles into existing lineages (admixture) [92].Gene flow can also rescue populations experiencing demographic decline and high rates of inbreeding. Hence it is important to understand the gene flow patterns and how they shape population structure.

One of the primary applications of RADSeq in population studies is to infer population structure [91]. The population structure and the level of genetic differentiation can be estimated using measures like F_{ST} [107, 52], Nei's G_{ST} [80], G'_{ST} [46] and D [61] using allele frequency data. The pairwise F_{ST} values between populations can be calculated from allele frequencies and used as a distance metric to cluster the populations and infer their structure. Gene flow has traditionally been estimated using genetic differentiation measures, which rely on the fact that gene flow reduces divergence and inbreeding in populations. Assuming an island model of migration and symmetric migration rates, migration can be estimated using these measures [111, 113]. Gene flow is symmetric between populations when the rate of gene flow is same in both the directions. In nature, gene flow can also be asymmetric if the rate of gene flow is not same in both directions. Asymmetric gene flow is common in systems influenced by physical processes like wind or water currents (e.g., [86]) and competition-driven directional dispersal. In such cases, genetic differentiation estimates between the populations can be significantly skewed [29]. Hence it is important to understand the gene flow patterns and the processes that lead to genetic structuring of populations.

Gene flow patterns have been estimated from mathematical models using maximum-likelihood or Bayesian approaches (e.g., [110, 8]), which involve different assumptions and require the estimation of large numbers of parameters. More recently a simpler approach to estimate relative gene flow between pairs of populations was introduced in *divMigrate* [98] from the *diversity* package [63]. The method defines a hypothetical pool of migrants for a given pair of populations, and estimates the directional components of genetic differentiation between each of the two populations and the hypothetical pool using measures of genetic differentiation such as multilocus D [23], multilocus G_{ST} [80] or Nm (effective number of migrants; [2]). Relative migration levels are estimated from the directional genetic differentiation, where the population with a larger relative migration value is considered the source of gene flow and the one with the smaller value as recipient [98] .

Though methods like Migrate [8] and BayesAss [110] can be used to estimate gene flow patterns in asymmetric systems, they are computationally expensive and have a large number of parameters and options that need to be adjusted to the data set under consideration [98]. Mis-specification of parameters can lead to misleading results with high associated confidence scores [38]. Some studies have also mentioned issues with convergence and repeatability of results using Migrate and BayesAss [57, 37, 77] and they are also computationally demanding sometimes requiring impractical amounts of time to run [98]. On the other hand, estimating gene flow pattern using symmetric measures of genetic differentiation such as F_{ST}, G_{ST} and D could generate misleading results in asymmetric systems. The method introduced in *divMigrate* overcomes the highlighted issues in the above listed tools and methods, specifically identifying relative gene flows in asymmetric systems and also being less computationally demanding [98] . Although *divMigrate* is relatively simple and used by many different studies [59, 22, 85], the method is not validated for scenarios such as varied population sizes or recent common ancestry. Recent common ancestry will lead to low level of genetic differentiation between the populations and may cause divMigrate to generate misleading results as the method is based on genetic differentiation measures. The original paper describing the method [98] does not provide any guidelines on how the relative gene-flow network looks when applied to a system with absence of gene flow, i.e., when the genetic structuring is shaped only by genetic drift or natural selection.

All these methods rely on the genetic differentiation measures to identify gene flow patterns among populations. We explore the possibility of obtaining the relative gene flow patterns using the distribution of differences in the allele frequencies between pairs of populations. Using alleles present in only one population (“private alleles”) is a common strategy to estimate gene flow, as the logarithmic average frequency of private alleles is approximately linearly related to the logarithm of Nm [96]. In the proposed method we use frequencies of alleles that are present only in the (”exclusive to the”) two populations for which the relative gene flow is estimated. RADProc uses the graph components to perform de novo locus formation for each sample/individual and then builds a catalog of loci from all the samples/individuals. Each catalog locus represents all the alleles that are identified for that locus from all individuals and populations in the dataset. Further, the catalog can also serve as the input to the proposed method, as each catalog locus contains information like the populations in which the alleles are present and their corresponding allele frequencies.

4.2 Methods

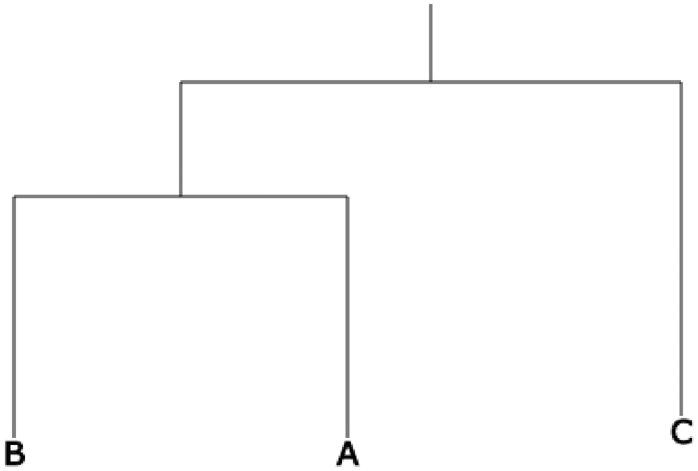

Figure 4.1: Population structure used to simulate the smaller datasets of 3 populations, where populations A and B are closely related and population C diverged from them at an earlier point in time. Different gene flow scenarios, varying number of loci and number of samples were simulated using this population tree.

When individuals migrate from one population to another, different scenarios can arise. The frequencies of existing alleles can increase or decrease based on whether the population is a source or a sink of gene flow and new alleles may be introduced into recipient populations. Let us consider three populations A, B and C with an underlying branching structure (Figure 4.1). Alleles in the resulting catalog can be categorized based on their presence in these three populations, such as "private alleles" (alleles that are present only in one population e.g., $\{A\}$-$\{B \cup C\}$) "common alleles" (alleles that are present at least in two populations e.g., $\{A \cap B\}$) and "exclusive alleles"

(alleles that are present only in two of the populations and not present in any other population e.g., {A ∩ B}-{C}) (Figure 4.2). By this definition, all exclusive alleles are common alleles, but the reverse is not true. In the absence of gene flow, we can expect a pair of populations with low genetic differentiation to have a larger number of exclusive alleles. In the presence of gene flow, the pair of populations with high gene flow relative to other populations is expected to have more exclusive alleles.

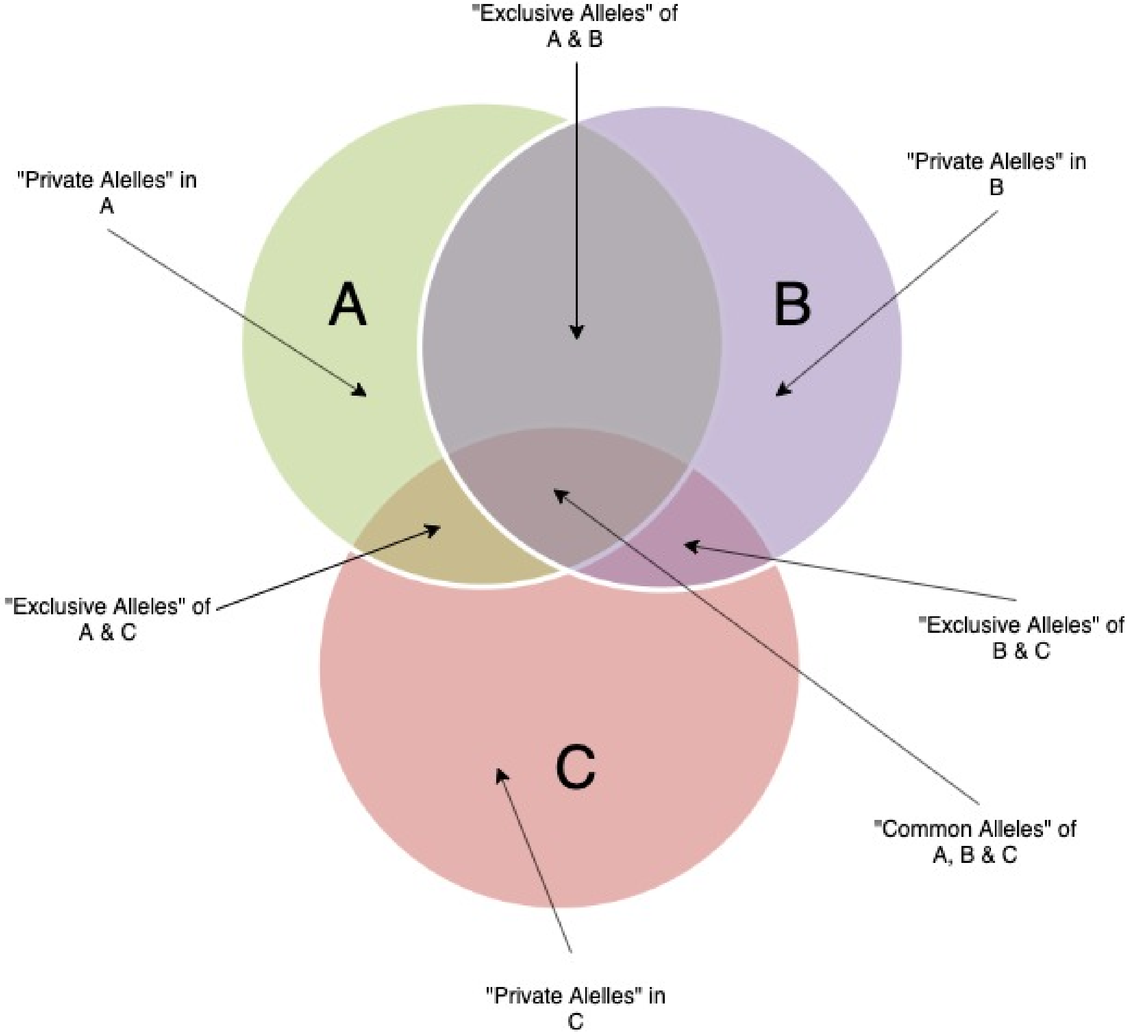
"Exclusive Alleles" of A & B
"Private Alelles" in A
"Private Alelles" in B
A
B
"Exclusive Alleles" of A & C
"Exclusive Alleles" of B & C
C
"Common Alleles" of A, B & C
"Private Alelles" in C

Figure 4.2: Venn diagram representing the private, common and exclusive alleles among the three populations.

Figure 4.3 illustrates the expected impacts of gene flow from population A to

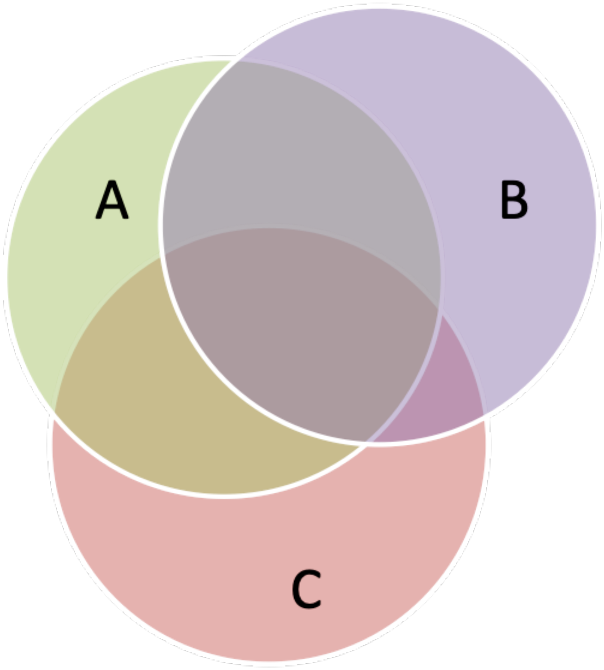

Figure 4.3: Venn diagram representing the private, common and exclusive alleles among the three populations in the presence of gene flow from A to C.

population C. In this case, the number of common alleles between the two populations is increased compared to the no gene flow scenario (Figure 4.2). But the gene flow from A to C also increases the number of common alleles between populations B and C. Since populations A and B are genetically similar due to recent divergence, they have more alleles in common, and the gene flow from A to C introduces common alleles into C. For this reason using common alleles between populations to understand the gene flow pattern and genetic differentiation will be misleading. Because of the confounding impact of gene flow on common alleles, our proposed method focuses on the exclusive alleles that are shared between pairs of populations. In our example of three populations in the presence of gene flow from A to C, although the number of common alleles between population B and C increased, the number of alleles exclusive to B and C is not expected to increase. This is because the gene flow from A to C has no influence on the number of exclusive alleles between populations B and C. In this chapter we propose a method using such exclusive alleles between populations to identify the direction and relative rates of gene flow.

4.2.1 Proposed method

The proposed method uses exclusive alleles between a given pair of populations to detect gene flow direction and the relative magnitude with respect to the other populations. Since the exclusive alleles are only present in the pair of populations, it could be either because of less genetic differentiation between the populations or due to gene flow. Hence, only using the proportion of exclusive alleles between populations is not enough to distinguish between populations with low level of genetic differentiation and the ones with gene flow; we consequently focus on the frequencies of exclusive alleles. We can also expect that in case of low level of genetic differentiation between the populations, the exclusive alleles between them would have similar allele frequencies [52]. Whereas exclusive alleles that arise due to gene flow will likely have frequencies that are more dissimilar. If the populations are homogenized due to high gene flow rates, the proposed method might fail to detect gene flow since the alleles are expected to have similar frequencies. In the three-populations scenario illustrated in Figure 1, populations A and B have less genetic differentiation compared to population C due to more-recent divergence, but gene flow between population A and C increases their degree of similarity. This gene flow can be unidirectional (A to C or C to A) or bidirectional (A to C and C to A), and in the latter case migration rates can be symmetric or asymmetric. For simplicity and clarity gene flow between a pair of populations is expressed using the notation $\gamma_{\text{[source][destination]}} = m$. For example, [$\gamma_{\text{AC}} = m$,$\gamma_{\text{CA}} = 0$] indicates a unidirectional gene flow from A to C, [$\gamma_{\text{AC}} = 3m/4$,$\gamma_{\text{CA}} = m/4$] represents asymmetric bidirectional gene flow between populations A and C, with migration rate from A to C greater than migration rate from C to A, and [$\gamma_{\text{AC}} = \gamma_{\text{CA}} = m/2$] represents symmetric bidirectional gene flow between populations A and C.

In general, we can expect that the frequencies of the newly introduced exclusive alleles in the recipient population will be much lower than their corresponding frequencies in the source population. However, in the extreme case, migration rates could be so high as to homogenize the allele frequencies in the two populations, making them one effective population and erasing any historical signal in the data. We calculate the differences in allele frequencies by combining the allele frequencies for each exclusive allele between a given pair of populations and estimate the ratio r, the frequency of the exclusive allele in a population to the combined allele frequency. Since the combined allele frequency function is arithmetic mean, the r values will range between 0 and 2.

$$r_A = \frac{f_A}{\left(\frac{f_A+f_B}{2}\right)} \tag{4.1}$$

$$r_B = \frac{f_B}{\left(\frac{f_A+f_B}{2}\right)} \tag{4.2}$$

Where,

f_A = Frequency of allele ‘a’ in population ‘A’

f_B = Frequency of allele ‘a’ in population ‘B’

r_A = Ratio of f_A to the combined allele frequency

r_B = Ratio of f_B to the combined allele frequency

The ratios are calculated for all exclusive alleles in a given pair of populations using equations 4.1 and 4.2. Alleles with equal frequencies in both populations will have $r_A = r_B = 1$. Alleles with a higher frequency in population A than in population B will have $r_A > 1$ and $r_B < 1$, and vice versa. In the absence of gene flow we can expect that most of the exclusive alleles would be concentrated around 1 and in the presence of gene flow they are more concentrated on the lower and upper ends of the r values. In the case of low and medium level migration rates, the frequency distributions of exclusive alleles should be influenced by the rates of gene flow. For each pair of populations we can plot the distribution of r values across the entire set of exclusive alleles, with r_A and r_B reflections of each other at the line $r = 1$. We can express our expectations under different scenarios:

No gene flow (γ^{**}=0). In the absence of gene flow between a pair of populations, we can expect most of the exclusive alleles to be concentrated around $r = 1$.

Unidirectional gene flow from A to C ($\gamma_{AC} > 0$, γ_{CA}=0). In the unidirectional case, we can expect population C to have a relatively high number of exclusive alleles with frequencies less than the corresponding allele frequencies in population A. The r-value distribution for population C would have more exclusive alleles on the $r < 1$ side than $r > 1$ and the other way round for population A.

Asymmetric bidirectional gene flow between A and C ($\gamma_{AC} > \gamma_{CA}$). Here the gene flow is in both directions, so we can expect the r-value distribution for population A also to have more number of exclusive alleles on the $r < 1$ side than $r > 1$. The proportion of exclusive alleles on the $r < 1$ side would be greater for population C than A, since the relative migration rate is higher from A to C than C to A.

Symmetric bidirectional gene flow between A and C ($\gamma_{AC} = \gamma_{CA}$). We would expect both populations A and C to have similar a proportion alleles with frequencies less than the corresponding allele frequencies in the other population. Hence the r-value distribution for both A and C would have almost same number of alleles on the $r < 1$ side.

4.2.2 Estimating gene flow direction and relative migration

The allele frequencies are estimated using the catalog loci and for each locus the frequencies of alleles add up to 1. The presence of gene flow, the direction of gene flow and the relative migration can be predicted from the distribution of the r values obtained from the exclusive alleles for each pair of populations. Since we are using the arithmetic mean as the function to combine the allele frequencies, 0 and 2 bound the range of proportions. To predict gene flow from population i to j, we need to see if the percentage of alleles with $r = 1$ is less than the percentage of alleles with $r < 1$. We define a p x p matrix **D**, where p is the total number of populations and each value $\mathrm{X}_{\mathrm{i,j}}$ represents the relative gene flow value from population i to j obtained from the r-value distributions between j and i. The diagonal in **D** corresponds to comparisons between identical populations is set to 0.

$$\mathbf{D} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,p} \\ \vdots & \ddots & \vdots \\ X_{p,1} & \cdots & X_{p,p} \end{bmatrix} \tag{4.3}$$

$\mathrm{X}_{\mathrm{i,j}}$ = (% of alleles with $r < 0.9$ in population j) – (% of alleles with $0.9 \leq r \leq 1$ in population i)

If two populations share very few exclusive alleles, their distribution of r values may be unstable due to the small sample size. To overcome this, we scale every entry in **D** with a value S, where S is the proportion of exclusive alleles between a pair of populations out of the number of common alleles between them.

$$S_{ij} = E_{ij}/C_{ij} \tag{4.4}$$

Where,

E_{ij} = Number of exclusive alleles between i and j.

C_{ij} = Number of common alleles between i and j.

For visualizing the gene flow patterns, we build an adjacency matrix G that is the product of **D** and S obtained from equation 4.4. In the adjacency matrix, the cell G[i,j] contains the value for gene flow from population i to j and the cell G[j,i] contains the value for gene flow from population j to i. The adjacency matrix is then used to build a network of populations to visualize the gene flow patterns.

$$G = \begin{bmatrix} X_{1,1} * S_{1,1} & \cdots & X_{1,p} * S_{1,p} \\ \vdots & \ddots & \vdots \\ X_{p,1} * S_{p,1} & \cdots & X_{p,p} * S_{p,p} \end{bmatrix} \tag{4.5}$$

In the case of bidirectional gene flows, for [$\gamma_{\mathrm{AC}} > \gamma_{\mathrm{CA}}$] the cell G[A,C] would have higher value than G[C,A] and for $\gamma_{\mathrm{AC}} = \gamma_{\mathrm{CA}}$ both G[A,C] and G[C,A] would have similar values. Unidirectional gene flow from A to C is a special case of asymmetric gene flow with γ_{AC} =m and $\gamma_{\mathrm{CA}} = 0$. In this case, we can expect G[A,C] value to be much higher than G[C,A], where G[C,A] can be zero or negative, depending on the percentage of exclusive alleles between A and C with similar frequencies (defined here as 0.9 $\leq$ r $\leq$). In general, we can differentiate [$\gamma_{\mathrm{AC}} > 0$, γ_{CA}=0] and [$\gamma_{\mathrm{AC}} > \gamma_{\mathrm{CA}}$] by the magnitude of G[C,A] value relative to G[A,C]. Even though we cannot define any strict threshold on the value of G[C,A], it can still provide insights into possible gene flow pattern. For example, if G[C,A] $= 0.2 *$ G[A,C], then we can say the probability of gene flow from C to A is very low compared to G[C,A] $= 0.7 *$ G[A,C]. In the network the vertices V represent the populations and edge weights W are assigned the corresponding values in **G**.

Confidence intervals are computed for all entries in **G** (and the corresponding edge weights) using a bootstrapping procedure. For each pair of populations, a distribution is estimated by resampling with replacement the catalog of loci 100 times. For each of the resampled data sets, a new G is calculated by identifying the exclusive alleles in the resampled data set. 95% confidence intervals (lower = 0.025, upper = 0.975) are constructed in both directions of gene flow between each pair of populations.

The matrix **G** can be normalized by the largest value in the matrix, so that

W ranges from 0 to 1. But in case of systems with absence of gene flow between any of the populations, W can be negative or positive (almost zero) values. The normalization would increase the magnitude of the positive values leading to wrong interpretations of the relative gene flow network. To avoid this the matrix **G** is normalized by the largest value only when there is at least one positive value in **G** with a 95% confidence intervals that does not include 0. We might also not want to normalize the edge weights when there are some positive edges and their values are near zero.

4.2.3 Simulated datasets

The performance of the proposed method was first assessed using simulations. The simulations were performed using the coalescent simulator *simul* provided by EggLib [28] and the *simrrls* [33] Python module for RADSeq-like data. The **egglib.simul. CoalesceParamSet** class was used to specify the parameters like migration rates (M), the relative size of all populations (N) and to provide the input tree structure. The migration rate M is expressed as $4Nm$, where N is the effective population size and m is the probability that a given individual migrates from the source to recipient population. For all the simulations the per site mutation rate was fixed at 1×10^{-9} (default in *simrrls*) and the base population size fixed at N = 100,000. The coalescent model for the most recent common ancestor of A and B was set to $0.5 * 4N$ generations ago, A and B with C to $1 * 4N$ generations ago and the root to $1.5 * 4N$ generations ago, providing the topology(Figure 4.1). Three different migration levels are used: 1 migrant per 10 generations ('low', $m = 1e-06$), 1 migrant per generation ('medium', $m = 1e-05$) and 10 migrants per generation ('high', $m = 1e-04$) [34]. Datasets were simulated with gene flow under three regimes: unidirectional [$\gamma_{\mathrm{AC}} > m$, γ_{CA}=0], bidirectional asymmetric [$\gamma_{\mathrm{AC}} = 3m/4$, $\gamma_{\mathrm{CA}} = m/4$], and bidirectional symmetric [$\gamma_{\mathrm{AC}} = m/2$, $\gamma_{\mathrm{CA}} = m/2$], for different migration levels, number of loci (L), number of samples (n) and population sizes (N). To evaluate the impact of sample size, each migration level and gene flow type combination was simulated with sample size $n = 20$, 30, and 40 with the number of loci L set to 5,000. To evaluate the impact of sample size, each migration level and gene flow type combination was simulated L=10k, 20k, and 30k with n set to 10. Datasets were simulated by varying population size (N_e) for population C to evaluate the impact of differences in population sizes of source and recipient. The N_e value for population C was set to , , and the population size of A (i.e., 100,000). A larger dataset of 12 populations (Figure 4.4) was also simulated, each with 15 samples and 15,000 loci per sample. Population 'A' was set as the recipient and all other 11 populations as source of gene flow [γ_{A^*} =0, $\gamma_{*\mathrm{A}} = m$] with m set to $1e - 05$ (medium level).

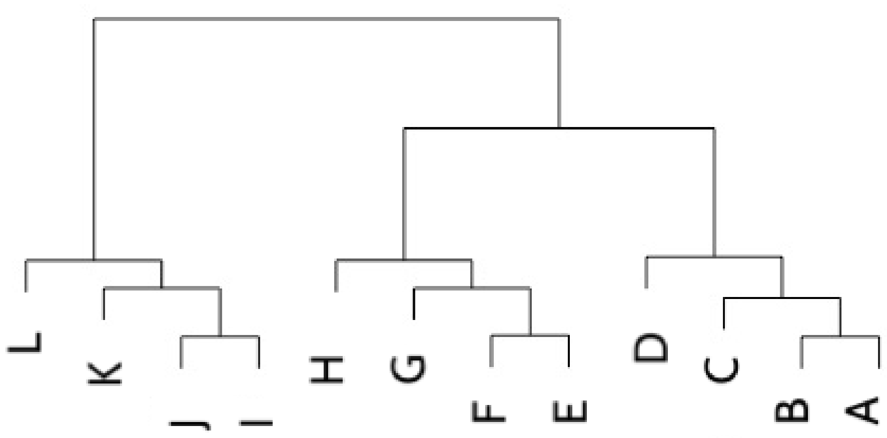

Figure 4.4: Population structure used to simulate the larger dataset of 12 populations. In the simulation the gene flow is unidirectional from all other populations to A.

4.2.4 Empirical dataset

Evaluation of the proposed method was also performed using RADSeq data extracted from green crab (*Carcinus maenas*) samples, which were first used to study population structure in the Northwest Atlantic [58](Figure 4.5). The dataset consists of 242

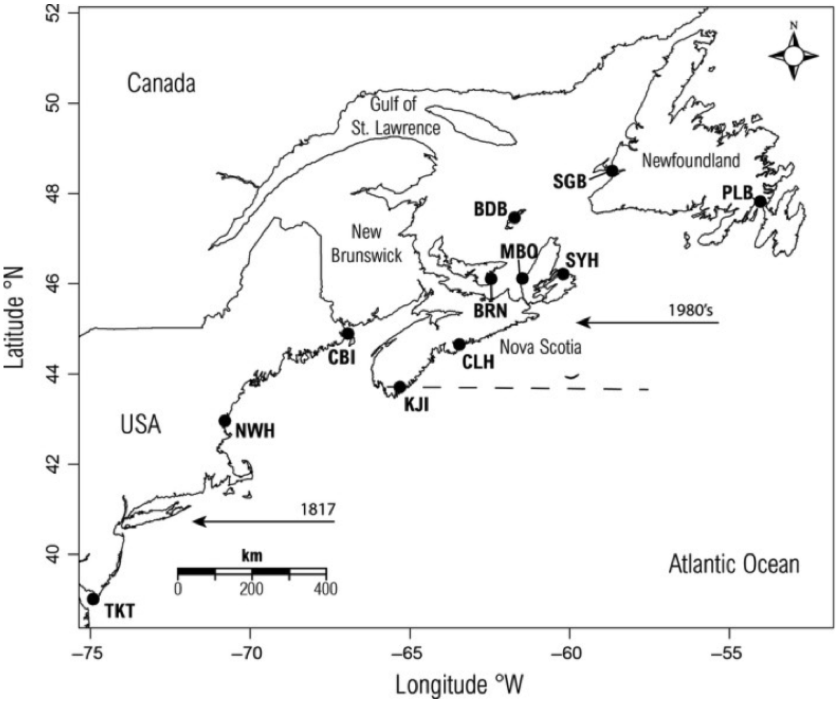

Figure 4.5: The 11 sampling locations in the eastern North America in USA and Canada for the green crab dataset. The map was accessed from [58].

samples from 11 different sites (Table 2.1). Each library consisted of 22 samples identified by variable length in-line barcodes ranging from 5 to 9 bp. The libraries were sequenced on a HiSeq 2000 (Illumina) as 100 bp paired-end sequences sequences. Each sample comprised approximately 2.5 million RAD-tags. Using this dataset, previous studies have observed two population groups: locations TKT, NWH and CBI constitute southern populations, while CLH, BRN, MBO, SYH, BDB and SGB are identified as northern populations. The two locations KJI and PLB were intermediate between the northern and southern groups with possible hybridization and admixture.

4.2.5 Evaluation

Multiple datasets were simulated to evaluate the impact of sample size, number of loci and population sizes. The simulations were done using the 3 populations and input tree structure (Figure 4.1). The gene flow patterns predicted from the simulated

Location code	**Location name**
SGB	St. George's Bay, NL
PLB	Placentia Bay, NL
BDB	Baie de Bassin, QC
SYH	Sydney Harbour, NS
MBO	Mabou, NS
BRN	Brudenell River, PE
CLH	Cole Harbour, NS
KJI	Kejimkujik, NS
CBI	Campobello Island, NB
NWH	New Hampshire
TKT	Tuckerton, NJ

Table 4.1: Location codes and their names for the 11 different sampling sites.

datasets for different gene flow scenarios with $m = 1e - 05$ were resampled 100 times and the 95% confidence intervals were determined to evaluate the significance of the gene flow patterns. The gene flow patterns predicted by the proposed method were compared with the ones obtained using *divMigrate*. Although there are other methods to infer gene flow patterns as listed in the introduction, *divMigrate* is more relevant to compare with the proposed method, as both are non-parametric approaches to infer relative gene flow in asymmetric systems. The ability of the two methods to correctly identify underlying gene flow patterns by comparing the relative gene flow values generated by them. We also compared the relative gene flow networks obtained for the no gene flow scenario, to evaluate how the two methods worked when the genetic structuring among the populations is not influenced by gene flow.

The *divMigrate* tool is also hosted as a web application and requires a 'genepop' format input file. To obtain the genepop format file, the simulated and green crab datasets were processed using the Stacks pipeline (Stacks v1.42). The RADSeq data from each individual sample were cleaned, demultiplexed and de novo assembled using the default ustacks parameters $M = 2$ and $m = 3$. The catalog of loci was built using *cstacks* with maximum nucleotide distance allowed between catalog loci to merge $n = 1$. The *ustacks* and *cstacks* files are then passed to the *populations* program

Population pairs	**Common Alleles**	**Exclusive Alleles**
AB	9749	2821
AC	7524	596
BC	7519	591

Table 4.2: Number of common alleles and exclusive alleles between different pairs of populations in the absence of gene flow.

to obtain the required genepop format output file. The genepop format file was then input to divMigrate-online and the relative migration networks were obtained. divMigrate-online provides options to choose a migration statistic from a list of three different genetic differentiation measures such as multilocus D, multilocus G_{ST} or Nm as migration statistic. In our study we used multilocus G_{ST} as the migration statistic, because it performed better than the other measures in evaluations done by [98]. Similarly, the relative gene flow networks obtained from the green crab dataset was also compared with those obtained from *divMigrate*. The gene flow network obtained from the proposed method is also visualized using the GenGIS software [83].

4.3 Results

4.3.1 Simulated datasets

No gene flow

In the first scenario, no gene flow was allowed in the simulated RADSeq dataset. The dataset contained 30 individuals from the three populations with 10,000 loci per individual. After de novo locus formation and catalog building, the catalog contained 11,356 loci; of these, 9885 loci were present in at least two of the populations and retained for the analysis. These loci were represented by 32,827 alleles in total. Table 4.2 shows the number of common alleles between each pair of populations and the corresponding number of exclusive alleles. Due to their more recent divergence, population pair AB had more common and exclusive alleles than population pairs

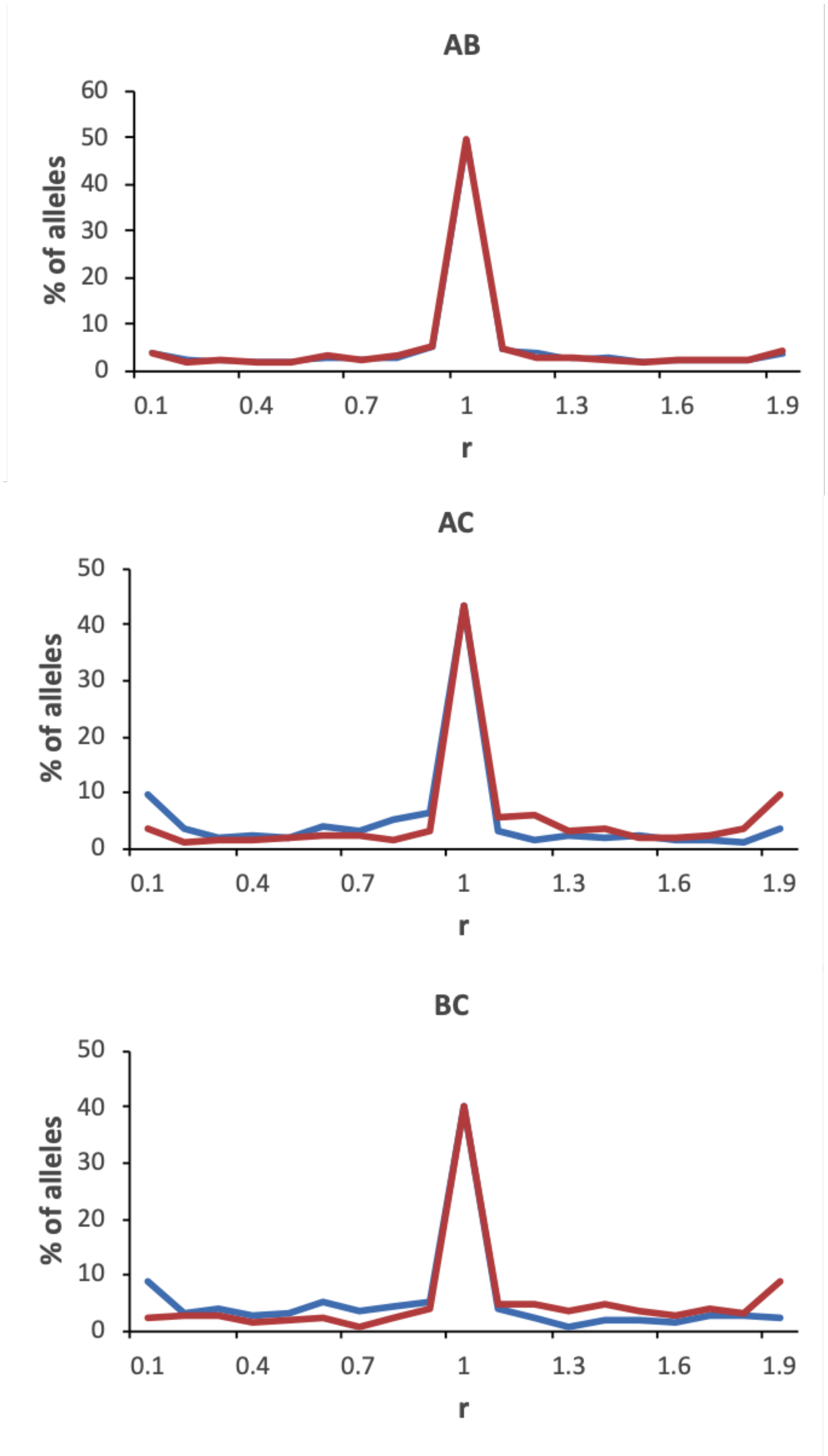

Figure 4.6: Allele distribution for values of r when $\gamma_{**} = 0$ for all three pairs of populations (AB, AC and BC). The blue line represents the first population and red line represents second population in each pair of populations.

AC and BC.

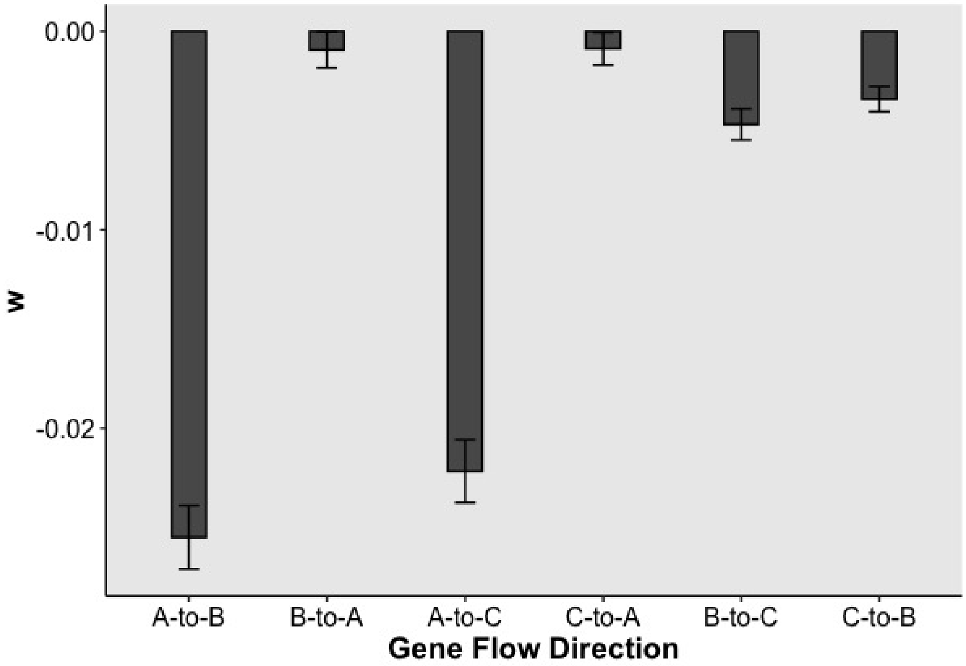

Figure 4.7: The values of W for all gene flow directions in the absence of gene flow between all pair of populations ($\gamma_{**} = 0$).

Using equations 4.1 & 4.2, the values of ratio r for all the pairs of populations was derived and their distribution was recorded. Figure 4.6 shows the plots obtained from the distributions for the pairs of populations. Since there was no gene flow, as expected all the pairs of population had balanced distributions and the majority of the alleles had a frequency ratio ~ 1. From the adjacency matrix $\mathbf{G}$, the plots of W and the 95% confidence intervals were obtained. In all cases, $W \leq 0$, which indicates that most exclusive alleles have similar allele frequencies in both populations. Figure 4.7 shows that W ranged from -0.03 to 0, suggesting no gene flow between any of the pairs of populations.

Population pairs	**Common Alleles**	**Exclusive Alleles**
AB	9718	123
AC	13672	4077
BC	10244	649

Table 4.3: Number of common and exclusive alleles between different pairs of populations simulated with [$\gamma_{\text{AC}} = 0.00001$,$\gamma_{\text{CA}} = 0$].

Unidirectional gene flow

In the simulated case of unidirectional gene flow [γ_{AC} =0.00001,γ_{CA} =0], the gene flow happens only from population A to C. The dataset contained 30 individuals from the three populations with 10,000 loci per individual. After de novo locus formation and catalog building, the final catalog had 10,851 loci that were present in at least two of the populations. There were 31,739 alleles in total; Table 4.3 shows the number of common alleles between each pair of populations and the corresponding number of exclusive alleles. Unlike in the no-gene-flow scenario, population pair AC had more common and exclusive alleles than the population pairs AB and BC, because of the gene flow from population A to C. We can also observe that the number of common alleles and exclusive alleles between populations B and C is higher than between populations A and B. This pattern arises because populations A and B are more closely related due to recent common ancestry, and gene flow from A to C increases the number of common alleles between populations B and C. The frequencies of the exclusive alleles were obtained and the values of ratio r for all the pair of populations were calculated and their distribution was recorded.

Figure 4.8 shows the imbalances in the distributions of values of r for populations A and C. In population A, the proportion of exclusive alleles on $r < 1$ is less than $r > 1$, whereas in population C the proportion of exclusive alleles on $r < 1$ is greater than $r > 1$. From the W-plot (Figure 4.9) the values of W_{AC} and W_{CA} were 0.10 and 0.25 respectively, and W_{AC} was approximately four times higher than W_{CA} suggesting unidirectional gene flow from A to C.

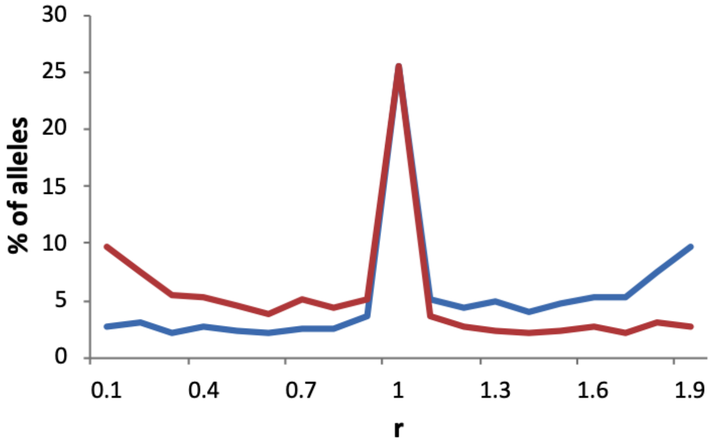

Figure 4.8: Allele distribution for values of r, in case of $[\gamma_{\text{AC}} = 1e-05, \gamma_{\text{CA}} = 0]$. The blue line represents population A and the red line represents population C.

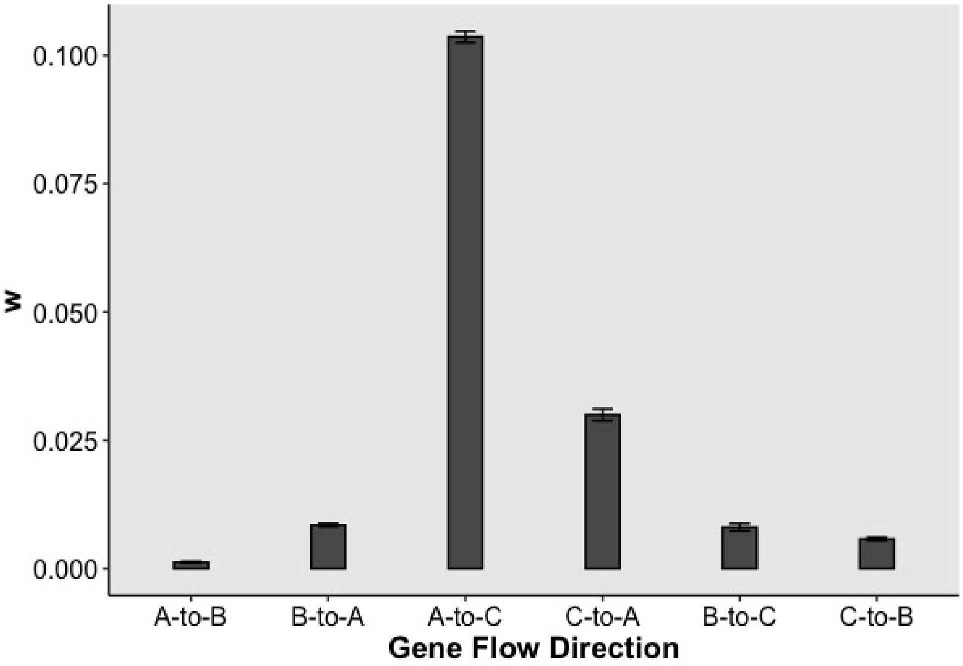

Figure 4.9: The values of W for different pairs of populations in case of $[\gamma_{\text{AC}} = 1e-05, \gamma_{\text{CA}} = 0]$. The high value for W_{AC} than others indicates that gene flow is only in the A to C direction.

Population pairs	**Common Alleles**	**Exclusive Alleles**
AB	10108	470
AC	16810	5885
BC	10246	657

Table 4.4: Number of common alleles and exclusive alleles between different pairs of populations in case of [$\gamma_{\mathrm{AC}} = 3m/4$,$\gamma_{\mathrm{CA}} = m/4$].

Asymmetric bidirectional gene flow

In the case of asymmetric bidirectional gene flow, the gene flow happens in both directions but with higher migration rates in one of the directions. In our simulated dataset the migration rate m was set to $1e - 05$ (medium level), with $3m/4$ from population A to C and $m/4$ from C to A. The catalog contained 10,342 loci that were present in at least two of the populations and consisted of 33,273 alleles in total. From Table 4.4, we observed that similar to the unidirectional gene flow from A to C, population pair AC had more common and exclusive alleles than population pairs AB and BC, because of the gene flow between populations A and C. The frequencies of the exclusive alleles were obtained and the combined allele frequencies for the alleles were calculated for each pair of populations.

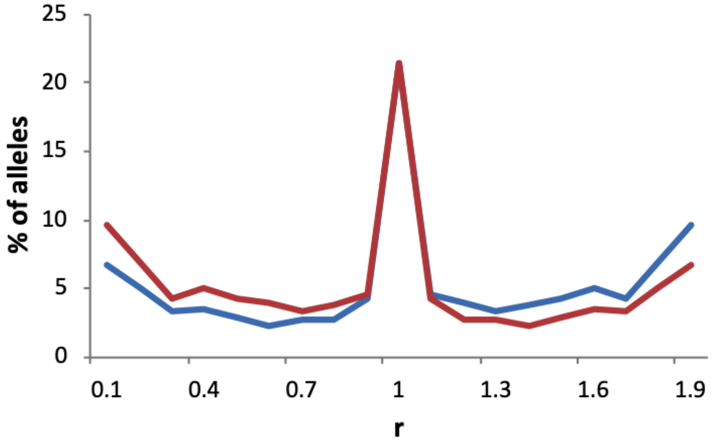

Figure 4.10: Allele distribution for values of r, in case of [$\gamma_{\mathrm{AC}} = 3m/4$,$\gamma_{\mathrm{CA}} = m/4$]. The blue line represents population A and the red line represents population C.

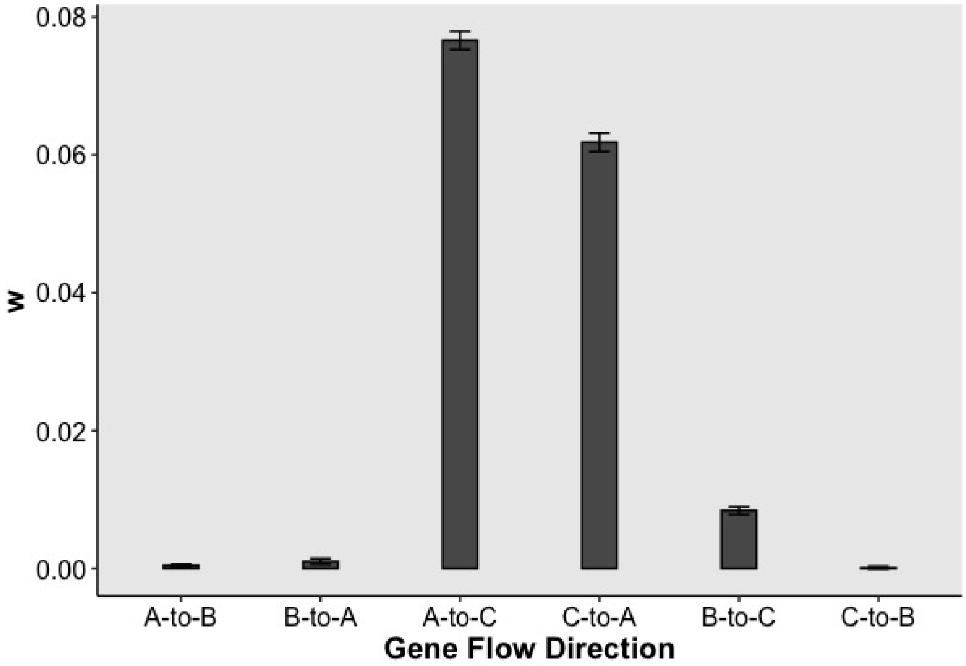

Figure 4.11: The values of W for different pairs of populations in case of $[\gamma_{\text{AC}} = 3m/4, \gamma_{\text{CA}} = m/4]$. The high values for W_{AC} and W_{CA} indicate bidirectional gene flow between A and C. $W_{AC} > W_{CA}$ denotes that gene flow from A to C is higher than C to A.

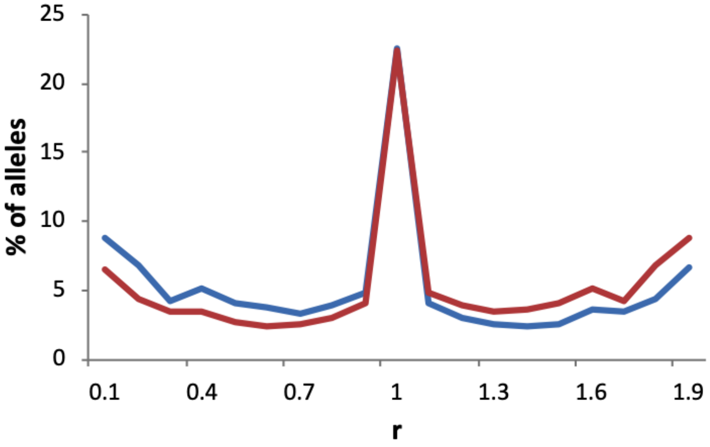

Figure 4.12: Allele distribution for values of r, in case of $[\gamma_{\text{AC}} = m/4, \gamma_{\text{CA}} = 3m/4]$. The blue line represents population A and the red line represents population C.

Figure 4.10 shows the plots obtained from the distributions of alleles for the pairs of populations. The plot for AC showed imbalances in the distribution of the alleles with respect to the values of r. Unlike the unidirectional gene flow scenario, the imbalance is seen in the plots for populations C and A, indicating a bidirectional gene flow between A and C. From the W-plot (Figure 4.11), gene flow is observed from A to C and also from C to A. The values indicate that the migration rate from A to C ($W_{AC} = 0.08$) is higher than the migration rate from C to A ($W_{CA} = 0.06$). When $\gamma_{\text{AC}} > \gamma_{\text{CA}} > 0$, the r-plot (Figure 4.12) and W-plot (Figure 4.13) reflected that the migration rate is relatively higher from C to A ($W_{CA} = 0.08$) than from A to C ($W_{AC} = 0.06$). The proposed method was able to capture the asymmetric bidirectional gene flow in both cases.

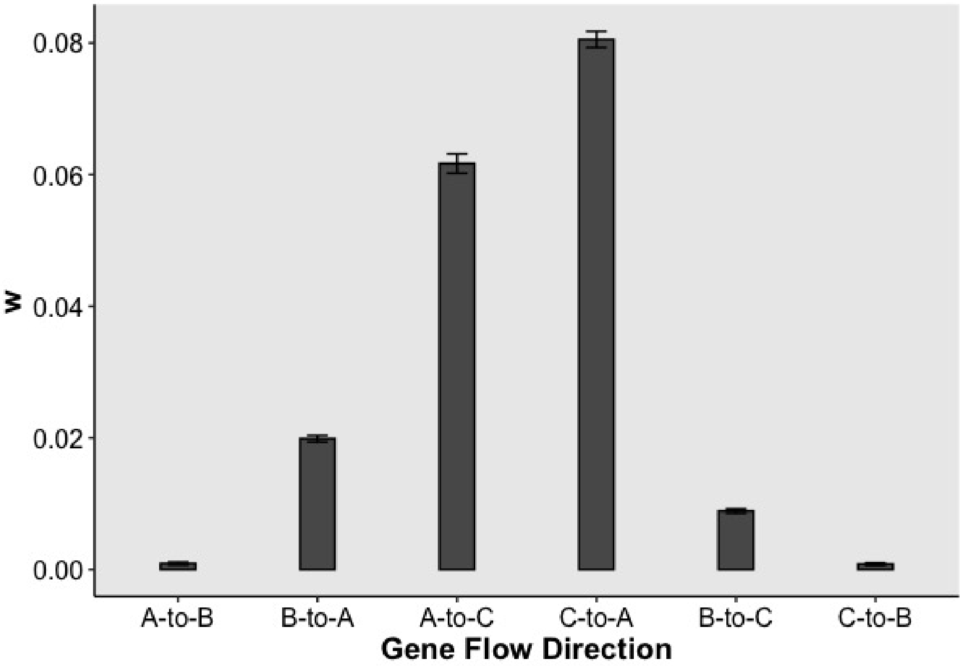

Figure 4.13: The values of W for different pairs of populations in case of [$\gamma_{\text{AC}} = m/4$,$\gamma_{\text{CA}} = 3m/4$]. The high values for W_{AC} and W_{CA} indicate bidirectional gene flow between A and C. $W_{AC} < W_{CA}$ denotes that gene flow from C to A is higher than A to C.

Population pairs	**Common Alleles**	**Exclusive Alleles**
AB	10185	363
AC	16005	6183
BC	10151	329

Table 4.5: Number of common alleles and exclusive alleles between different pairs of populations in case of [$\gamma_{\text{AC}} = m/2$,$\gamma_{\text{CA}} = m/2$].

Symmetric bidirectional gene flow

When the gene flow is symmetric, migration rate is same in both directions between a pair of populations. In our simulated dataset the migration rate m was set to $1e-05$ (medium level) with $m/2$ from population A to C and from C to A. We obtained 33,793 alleles in total from the 10,393 retained catalog loci.The observations from the number of common alleles and exclusive alleles (Table 4.5) were similar to the other gene-flow scenarios , i.e., population pair AC had more common and exclusive alleles than population pairs AB and BC.

Similar to the asymmetric bidirectional gene flow plots (Figure 4.14), the plots here too indicated gene flow in both directions between populations A and C. From the W-plot (Figure 4.15), gene flow is observed from A to C ($W_{AC} = 0.081$) and from C to A ($W_{CA} = 0.083$). The values indicate that the migration rate is nearly equal in both directions.

Low and high level migration rates

Figure 4.16 shows the W-plot obtained for different gene flow patterns when the migration rates are set low ($m = 1e-06$) and high ($m = 1e-04$). For low value of m, when [$\gamma_{\text{AC}} = m$,$\gamma_{\text{CA}} = 0$], the W_{AC} and W_{CA} values were 0.16 and 0.02 respectively (Figure 4.16a) and W_{AC} was eight times higher than W_{CA}, indicating unidirectional gene flow from A to C. Similarly, when [$\gamma_{\text{AC}} = 3m/4$,$\gamma_{\text{CA}} = m/4$], W_{AC} and W_{CA} values were 0.13 and 0.09 respectively (Figure 4.16b) and when [$\gamma_{\text{AC}} = m/2$,$\gamma_{\text{CA}} = m/2$], both W_{AC} and W_{CA} were approximately equal to 0.14 (Figure 4.16c).

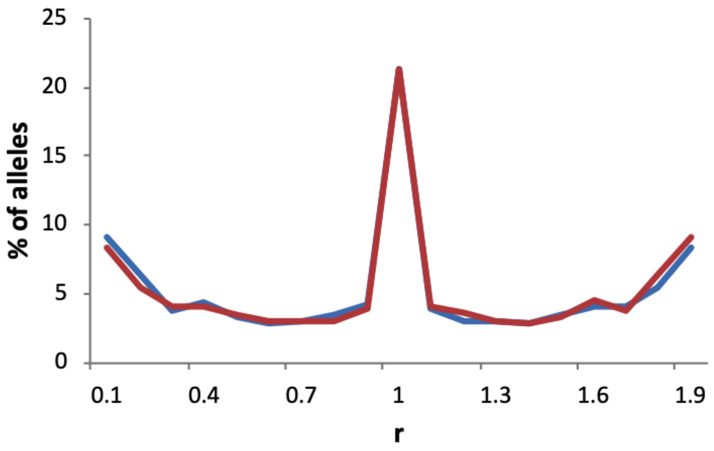

Figure 4.14: Allele distribution for values of r, in case of $[\gamma_{\mathrm{AC}} = m/2, \gamma_{\mathrm{CA}} = m/2]$. The blue line represents population A and the red line represents population C.

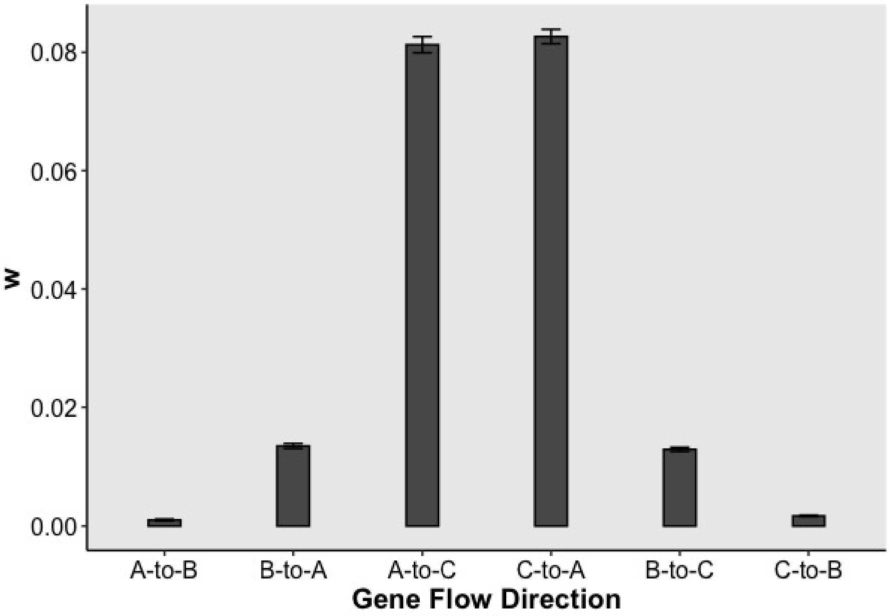

Figure 4.15: The values of W for different pairs of populations in case of $[\gamma_{\mathrm{AC}} = m/2, \gamma_{\mathrm{CA}} = m/2]$. The high values for W_{AC} and W_{CA} indicate bidirectional gene flow between A and C, and the almost similar values for W_{AC} and W_{CA} indicates symmetric migration rates.

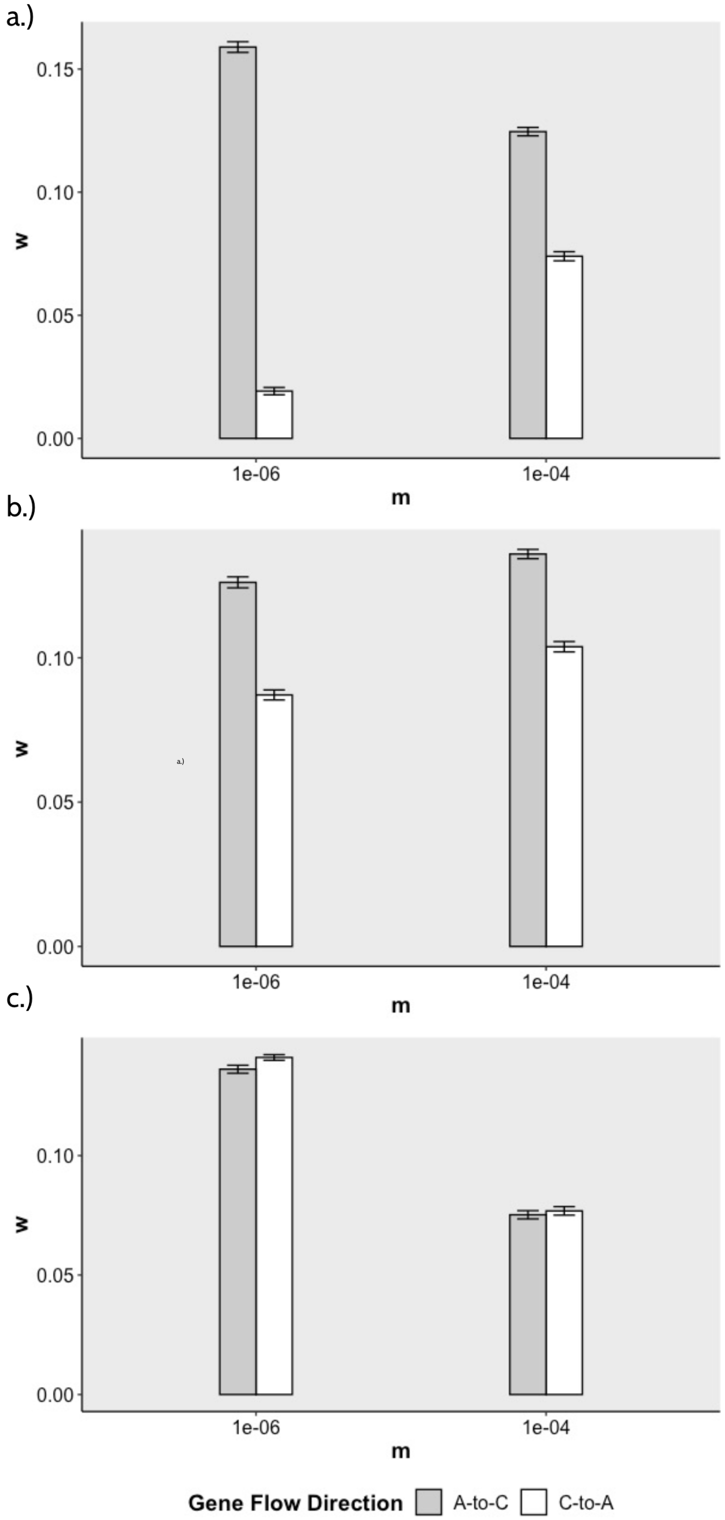

Figure 4.16: The values of W for different pairs of populations in case of $m = 1e-06$ (low) and $m = 1e-04$ for (a) [$\gamma_{\text{AC}} = m$,$\gamma_{\text{CA}} = 0$] (b) [$\gamma_{\text{AC}} = 3m/4$,$\gamma_{\text{CA}} = m/4$] and (c) [$\gamma_{\text{AC}} = m/2$,$\gamma_{\text{CA}}= m/2$].

The results show that the method was able to correctly identify the gene flow patterns, even in case of low migration rates. In case of high migration rate, for both [γ_{AC} = m,γ_{CA}= 0] (Figure 16a) and [γ_{AC} = $3m/4$,γ_{CA} = $m/4$] (Figure 4.16b) the W_{AC} and W_{CA} values reflected the asymmetry in the migration rate. But for [γ_{AC}= m,γ_{CA} = 0] the W_{AC} was only 1.7 times higher than W_{CA}, making it look like bidirectional gene flow, likely due to the homogenization of the populations due to high gene flow. when [$\gamma_{\mathrm{AC}} = m/2$,$\gamma_{\mathrm{CA}} = m/2$], both W_{AC} and W_{CA} were approximately equal to 0.07 (Figure 4.16c), indicating symmetric gene flow.

Comparisons with ***divMigrate***

We generated network visualizations for the simulated three-population datasets using both *divMigrate* and the proposed method. Figure 4.17a displays the relative migration network obtained for [$\gamma_{\mathrm{AC}} = 0$,$\gamma_{\mathrm{CA}} = 0$]. Although there is no gene flow among the populations, the relative migration network obtained using *divMigrate* suggested the existence of gene flow between A and B. This could be because the populations A and B are less differentiated due to their relatively recent divergence. When [γ_{AC} = m, γ_{CA} = 0] (Figure 4.17b), although *divMigrate* correctly identified the gene flow from A to C (1), the network also suggested a lesser amount of gene flow from C to A (0.32). In unidirectional gene flow (Figure 4.17b) and bidirectional asymmetric gene flow [$\gamma_{\mathrm{AC}} = 3m/4$,$\gamma_{\mathrm{CA}} = m/4$] (Figure 4.17c) the network indicated gene flow from C to A, 0.32 and 0.49 respectively. Thus making the results ambiguous and hard to distinguish between unidirectional gene flow and bidirectional asymmetric gene flow. In the case of bidirectional symmetric gene flow [$\gamma_{\mathrm{AC}} = m/2$,$\gamma_{\mathrm{CA}} = m/2$](Figure 17d) almost similar values were observed in both directions, A to C (1) and C to A (0.96). Interestingly the method generated similar networks for the no-gene-flow

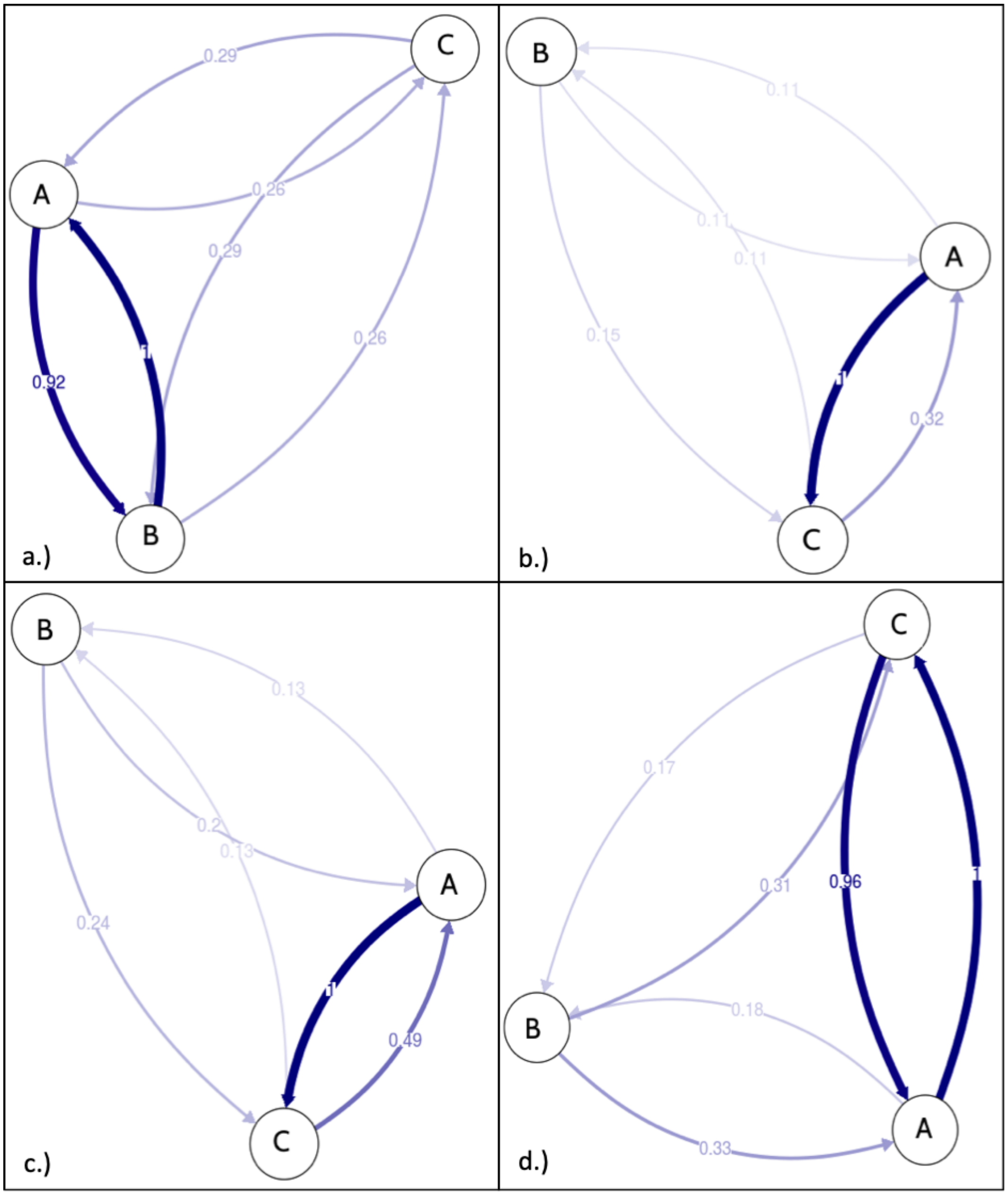

Figure 4.17: Relative gene flow network obtained using *divMigrate* for the 3 populations dataset with varying gene flow patterns at medium level migration rate. (a) [$\gamma_{\mathrm{AC}} = 0$,$\gamma_{\mathrm{CA}} = 0$] (b) [$\gamma_{\mathrm{AC}} = m$,$\gamma_{\mathrm{CA}} = 0$] (c) [$\gamma_{\mathrm{AC}} = 3m/4$,$\gamma_{\mathrm{CA}} = m/4$] and (d) [$\gamma_{\mathrm{AC}} = m/2$,$\gamma_{\mathrm{CA}}= m/2$].

(Figure 4.17a) and bidirectional symmetric gene flow (Figure 17d) scenarios. The edges with high edge weights even in the absence of gene flow (Figure 4.17a) makes it difficult to differentiate between presence and absence of gene flow.

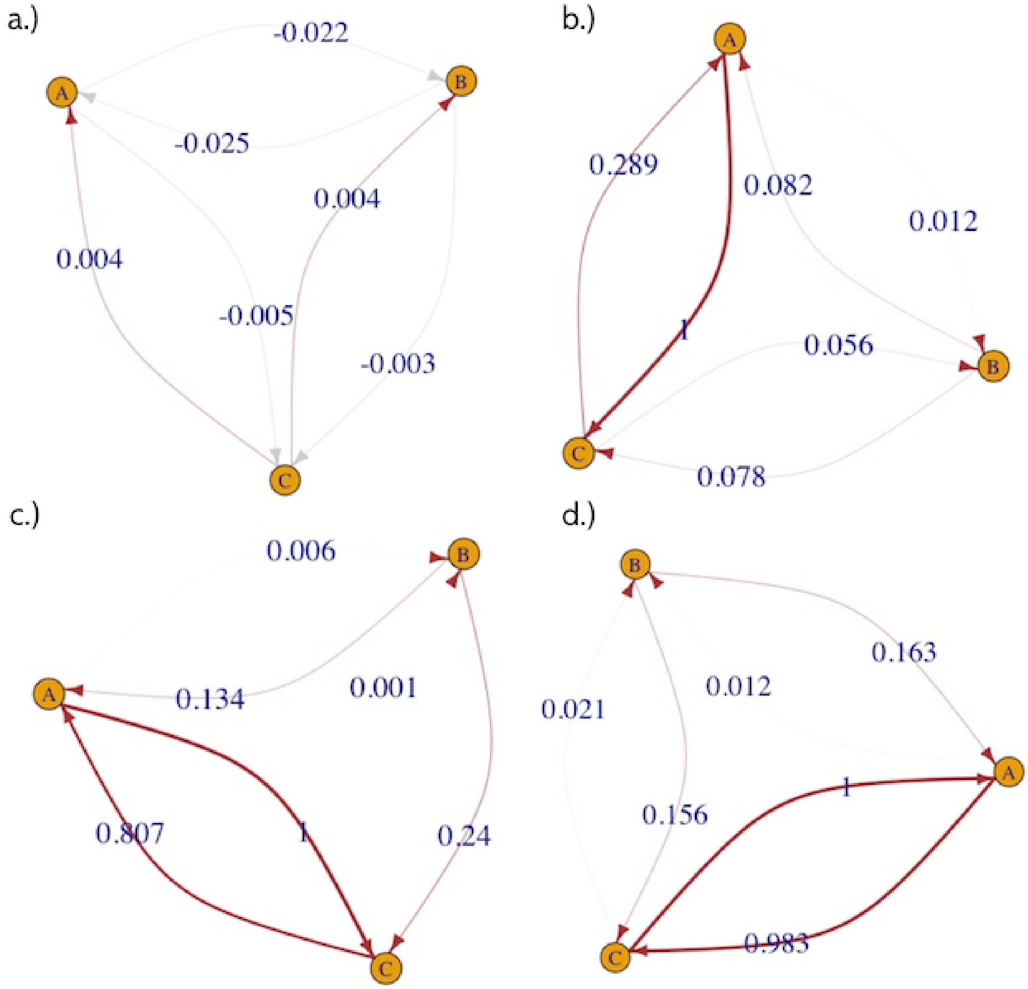

Figure 4.18: Relative gene flow network obtained using the proposed method for the three-population dataset with varying gene flow patterns when alleles with $r \geq 0.9$ is considered to have similar frequencies. (a) [$\gamma_{\mathrm{AC}} = 0$,$\gamma_{\mathrm{CA}} = 0$] (b) [$\gamma_{\mathrm{AC}} = m$,$\gamma_{\mathrm{CA}} = 0$] (c) [$\gamma_{\mathrm{AC}} = 3m/4$,$\gamma_{\mathrm{CA}} = m/4$] and (d) [$\gamma_{\mathrm{AC}} = m/2$,$\gamma_{\mathrm{CA}} = m/2$].

The network visualization obtained using the proposed method (Figure 4.18) the edges are not normalized by largest value for no gene flow scenario (Figure 4.18a)

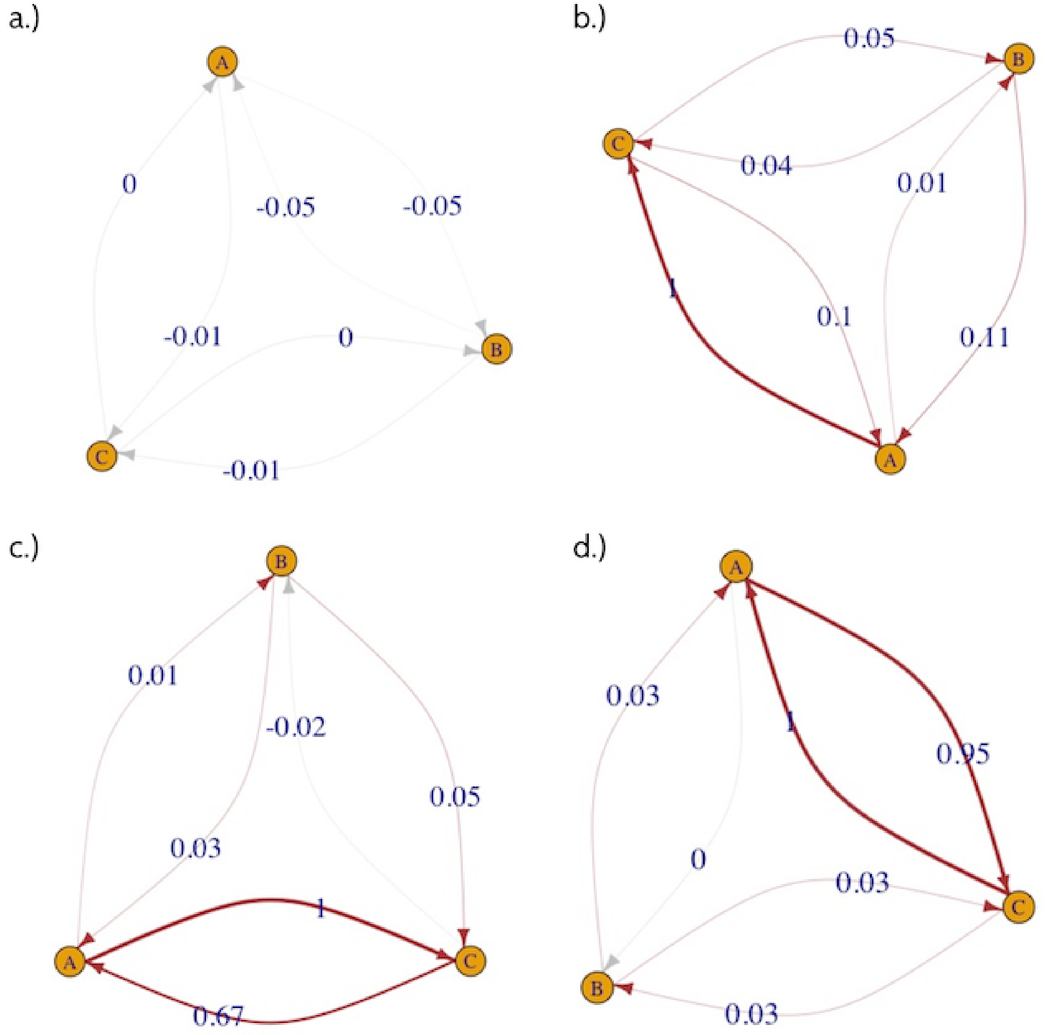

Figure 4.19: Relative gene flow network obtained using the proposed method for the three-population dataset with varying gene flow patterns when alleles with $r \geq 0.8$ is considered to have similar frequencies. (a) [$\gamma_{\mathrm{AC}} = 0$,$\gamma_{\mathrm{CA}} = 0$] (b) [$\gamma_{\mathrm{AC}} = m$,$\gamma_{\mathrm{CA}} = 0$] (c) [$\gamma_{\mathrm{AC}} = 3m/4$,$\gamma_{\mathrm{CA}}= m/4$] and (d) [$\gamma_{\mathrm{AC}} = m/2$,$\gamma_{\mathrm{CA}}= m/2$].

and normalized by largest value for other three scenarios. The W values were calculated using $r < 0.9$ for dissimilar allele frequencies and $0.9 \leq r \leq$ for similar allele frequencies. In the case of no gene flow, the W values for all pair of populations were either negative or the confidence intervals included zero (Figure 4.7), hence the edge weights are not normalized by largest value.The gene flow network was also able to detect no gene flow scenario (Figure 4.18a) correctly unlike the network obtained using *divMigrate*. The relative gene flow network clearly distinguishes unidirectional gene flow and bidirectional asymmetric gene flow (Figures 4.18b and 4.18c). But there were few edges with W values substantially higher than zero even for non-gene flow making it ambiguous to differentiate between presence and absence of gene flow. When the W values were calculated using $r < 0.8$ for dissimilar allele frequencies and $0.8 \leq r \leq$ for similar allele frequencies, the W values were almost zero for no gene flow edges (Figure 4.19). Thus implying, adjusting the r thresholds while calculating W can help to differentiate between presence and absence of gene flow.

Evaluation by varying number of samples and number of loci

To demonstrate the robustness of the proposed method, the method was applied to differing numbers of loci and samples simulated. The number of samples was set to 10, when the number of loci was varied and the number of loci was set to 10,000, when the number of samples was varied. The migration rate was set to medium ($m = 1e-05$) for the different gene flow patterns. In case of $[\gamma_{\mathrm{AC}} = m, \gamma_{\mathrm{CA}} = 0]$, for increasing values of L, the W_{AC} values ranged from 0.096 to 0.10 and W_{CA} ranged from 0.025 to 0.030 (Figure 4.20a). The values for W_{AC} were at least three times higher than that of W_{CA}, indicating unidirectional gene flow from A to C. On the other hand, the W values increased as the n values increased, and the values of W_{AC} ranged from 0.14 to 0.17 and W_{CA} ranged from 0.035 to 0.045 (Figure 4.20b). But the W_{AC} values were still at least three times higher than that of W_{CA}.

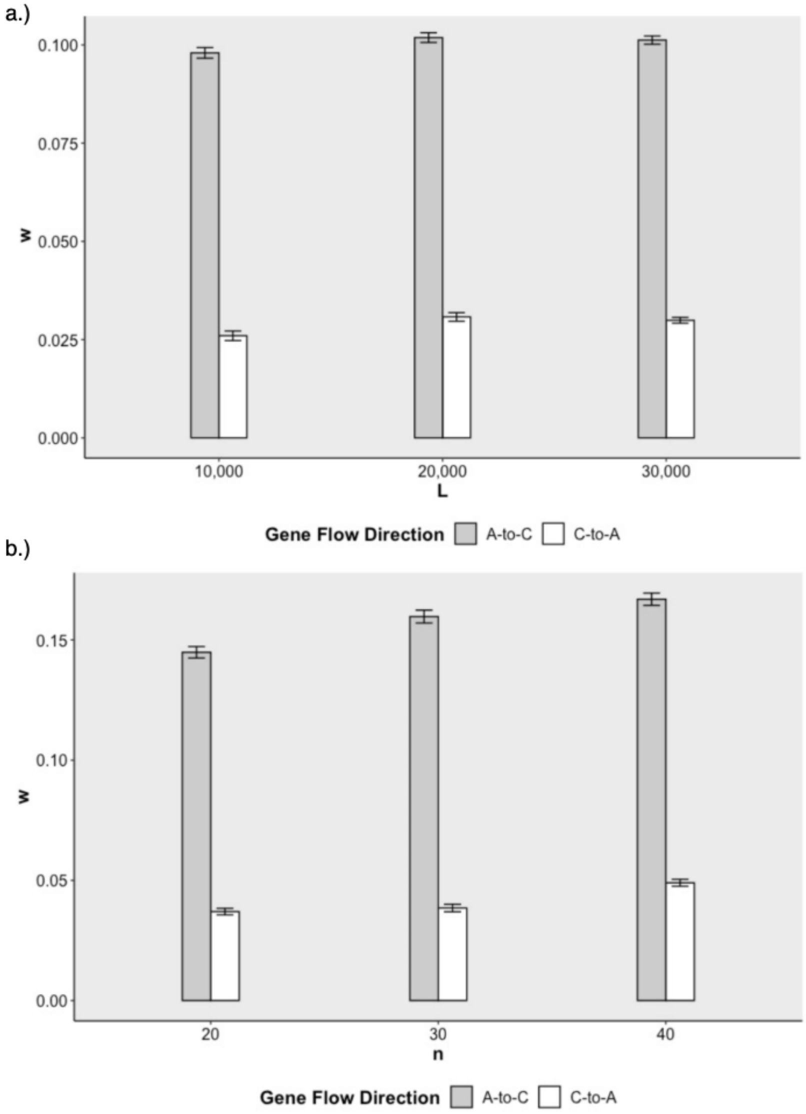

Figure 4.20: Plots using W values when $[\gamma_{\text{AC}} = m, \gamma_{\text{CA}} = 0]$ for different (a) Number of Loci and (b) Number of samples.

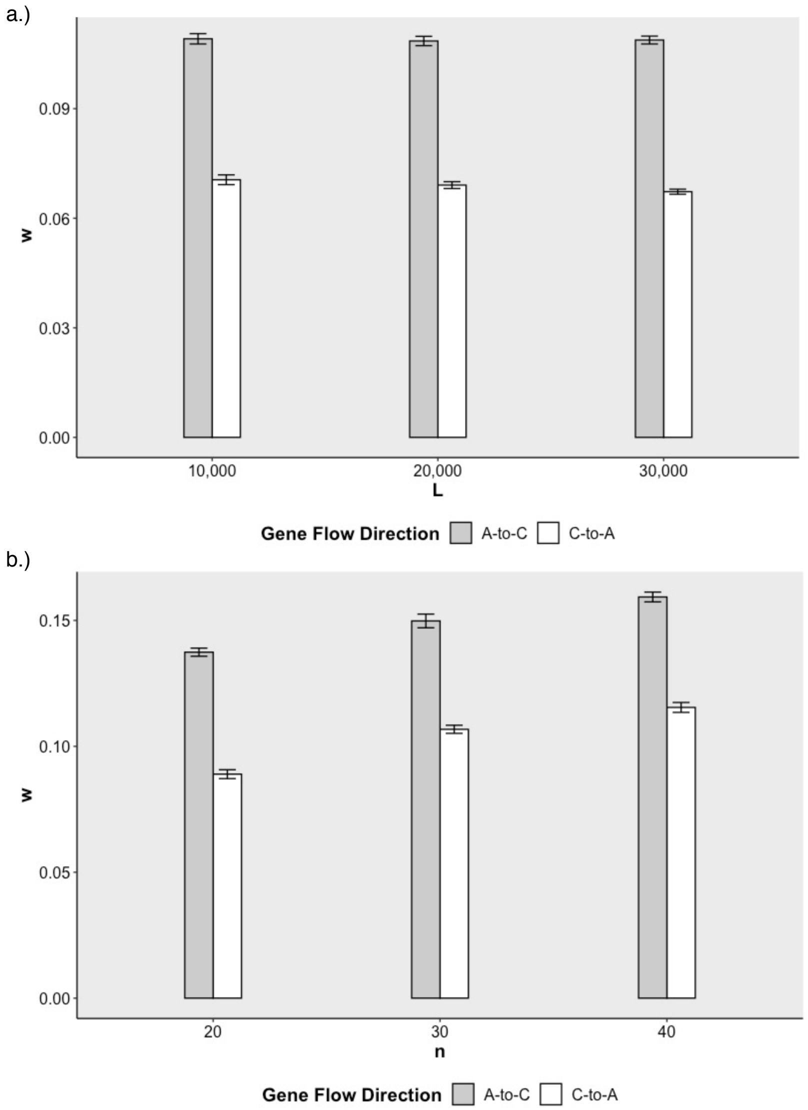

Figure 4.21: Plots using W values when [$\gamma_{\mathrm{AC}} = 3m/4$,$\gamma_{\mathrm{CA}} = m/4$] for different (a) Number of Loci and (b) Number of samples.

For asymmetric gene flow [$\gamma_{\mathrm{AC}} = 3m/4$,$\gamma_{\mathrm{CA}} = m/4$] (Figure 4.21) we obtained higher values of W in the direction of A to C (migration rate $= 3m/4$) and lower values in the direction of C to A (migration rate $= m/4$). Unlike [$\gamma_{\mathrm{AC}} = m$, $\gamma_{\mathrm{CA}} = 0$], W_{AC} values were only 1.5 times higher than W_{CA}, differentiating the unidirectional and bidirectional gene flow scenarios. For different values of L and n, the method was able to detect the asymmetric bidirectional gene flow between A and C, with higher migration rates in the direction of A to C.

In the case of symmetric gene flow [$\gamma_{\mathrm{AC}} = m/2$,$\gamma_{\mathrm{CA}} = m/2$] (Figure 4.22), the values of W were almost similar in both directions indicating symmetric migration rates. For increasing values of L, the values of W_{AC} were around 0.1 and the values of W_{CA} were around 0.095. For varying values of n, the values of W_{AC} ranged from 0.12 to 0.15 and the values of W_{CA} ranged from 0.14 to 0.15. As the value of n increased, the W_{AC} and W_{CA} values became more similar, in fact when $n = 40$, $W_{AC} = W_{CA}$, suggesting that high number of samples results in more accurate predictions.

Evaluation by varying population size of C

Here we evaluated the impact of population size on the values of W, by varying the effective population size N_e for population C. The values of N_e were set to $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{3}{4}$ the population size of A. The gene flow scenarios were simulated for [$\gamma_{\mathrm{AC}} = m$, $\gamma_{\mathrm{CA}} = 0$] and [$\gamma_{\mathrm{AC}} = 0$, $\gamma_{\mathrm{CA}} = m$] to examine the ability of the method to detect unidirectional gene flow from relatively small to relatively large populations and vice versa. In case of [$\gamma_{\mathrm{AC}} = m$, $\gamma_{\mathrm{CA}} = 0$] (Figure 4.23a) the plot indicated unidirectional gene flow from A to C for all the different population sizes of C. On the other hand, for [$\gamma_{\mathrm{AC}} = 0$, $\gamma_{\mathrm{CA}} = m$] (Figure 4.23b) the plot indicated unidirectional gene flow from C to A. In both the cases, the range of range of W in the no-gene-flow direction was considerably lower than the one with gene flow. The method correctly identified the unidirectional gene flow direction irrespective of the population size of the source

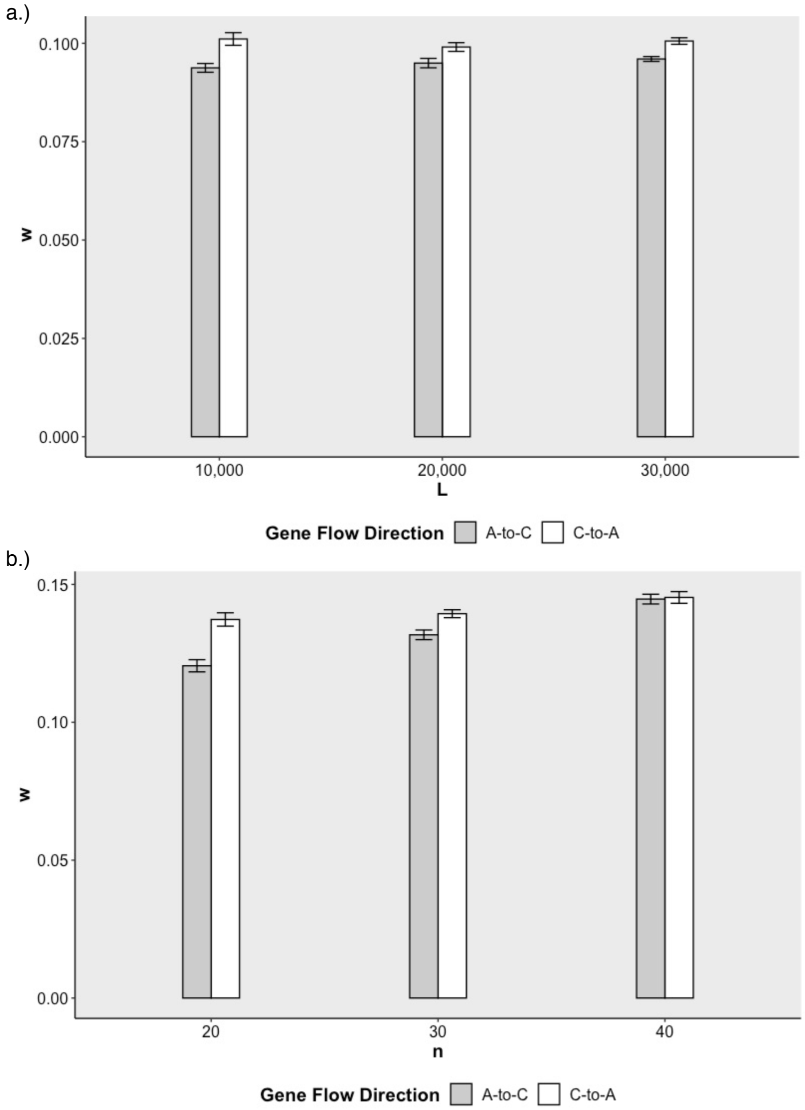

Figure 4.22: Plots using W values when [$\gamma_{\mathrm{AC}} = m/2$,$\gamma_{\mathrm{CA}} = m/2$] for different (a) Number of Loci and (b) Number of samples.

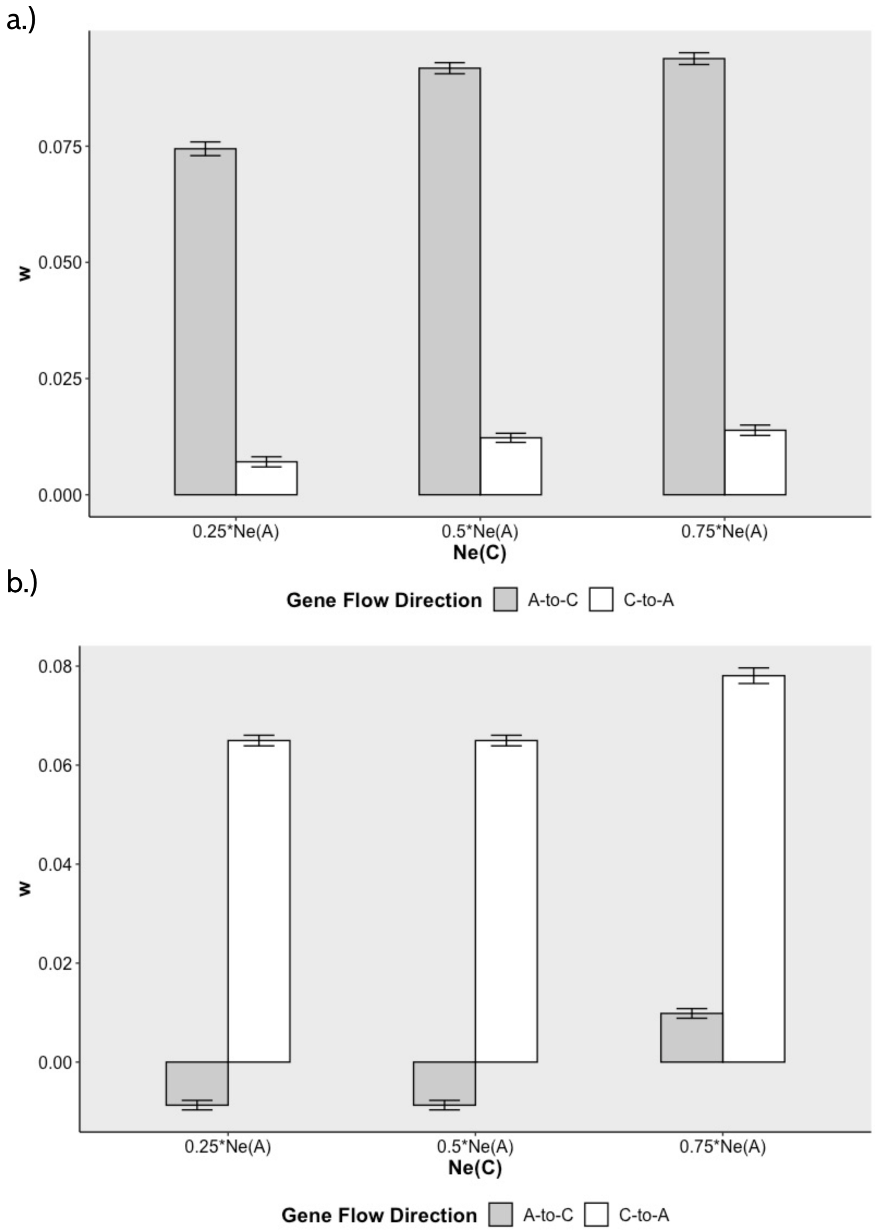

Figure 4.23: Plots using W values for varying population sizes for C when (a) [$\gamma_{\text{AC}} = m$, $\gamma_{\text{CA}} = 0$] and (b) [$\gamma_{\text{AC}} = 0$, $\gamma_{\text{CA}} = m$].

and recipient populations.

In the case of [$\gamma_{\mathrm{AC}} = 3m/4$,$\gamma_{\mathrm{CA}} = m/4$], where the gene flow is bidirectional and the migration rate from the smaller population C is much lower than the larger population A (Figure 24a), the method correctly identified the asymmetric gene flow for $N_e(\mathrm{C}) = 0.75 * N_e(\mathrm{A})$, and $N_e(\mathrm{C}) = 0.5 * N_e(\mathrm{A})$. When $N_e(\mathrm{C})$ is further lowered to $0.25 * N_e(\mathrm{A})$, the method mislabeled the gene flow direction with higher migration rate. This could be because the higher gene flow from the larger population could increase the allele frequencies in the smaller population. When [$\gamma_{\mathrm{AC}} = m/4$,$\gamma_{\mathrm{CA}} = 3m/4$], the W plot correctly labeled the gene flow direction with higher migration rate irrespective of the population sizes, but the range of W values decreased with decreasing population size of C (Figure 4.24b). Similarly, for [$\gamma_{\mathrm{AC}} = m/2$,$\gamma_{\mathrm{CA}} = m/2$] the values of W reflected the gene flow correctly, and the values decreased as the size of population C decreased (Figure 4.25).

Twelve-population dataset

In the dataset comprising twelve populations, the gene flow rate was set to $\gamma_{*\mathrm{A}} = 0.00001$, with all other rates set to zero. In the full network connecting all the populations, W ranged from -0.04 to 1. By setting a threshold for $W > 0$, the edges with $W \leq 0$ were filtered out. Figure 4.26 shows that the proposed method was able to capture this gene flow pattern. The migration rates used for the simulation were exactly same, but the W values did not reflect that, because the proportion of exclusive alleles between each pair of populations also influences the W values. The results demonstrate that the proposed method works even for datasets with large number of populations.

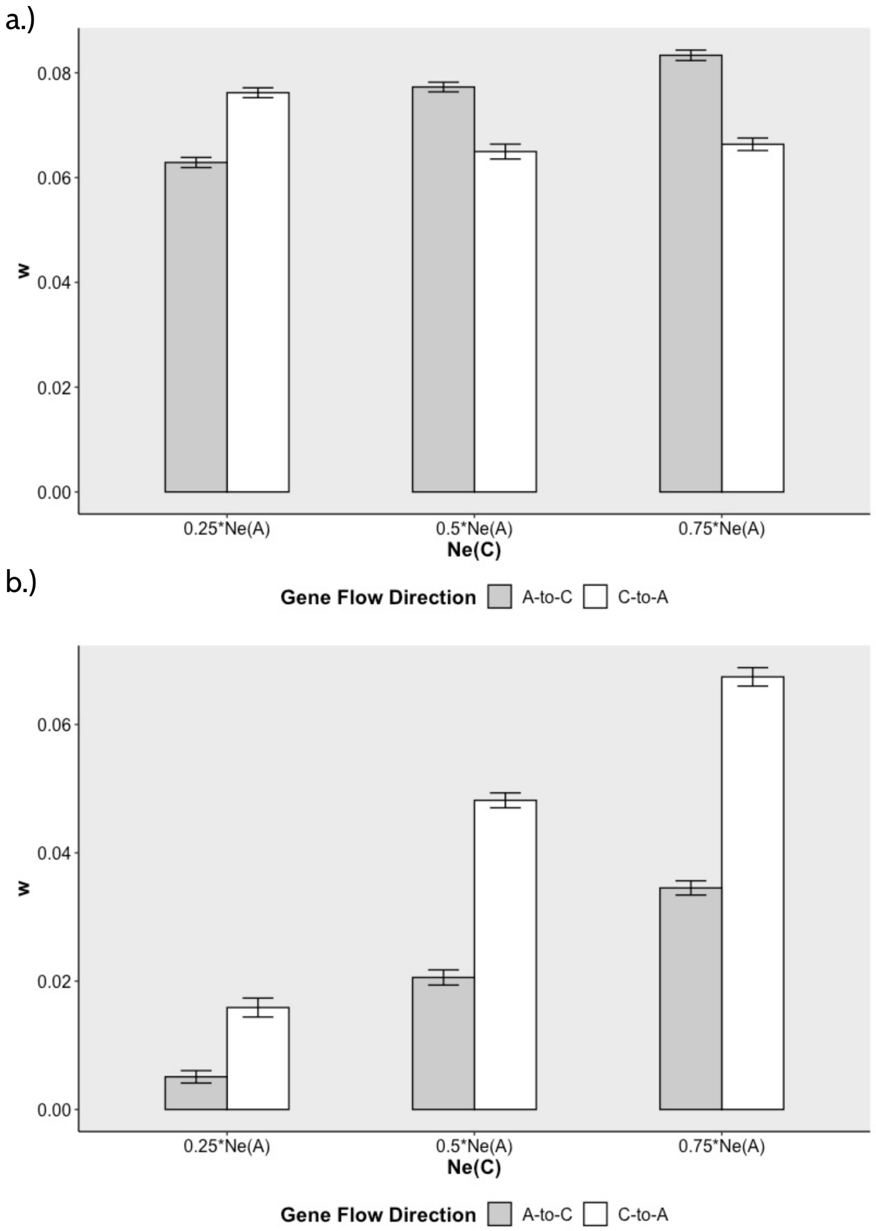

Figure 4.24: Plots using W values for varying population sizes for C when (a) [$\gamma_{\text{AC}} = 3m/4$,$\gamma_{\text{CA}} = m/4$] and (b) [$\gamma_{\text{AC}} = m/4$,$\gamma_{\text{CA}} = 3m/4$].

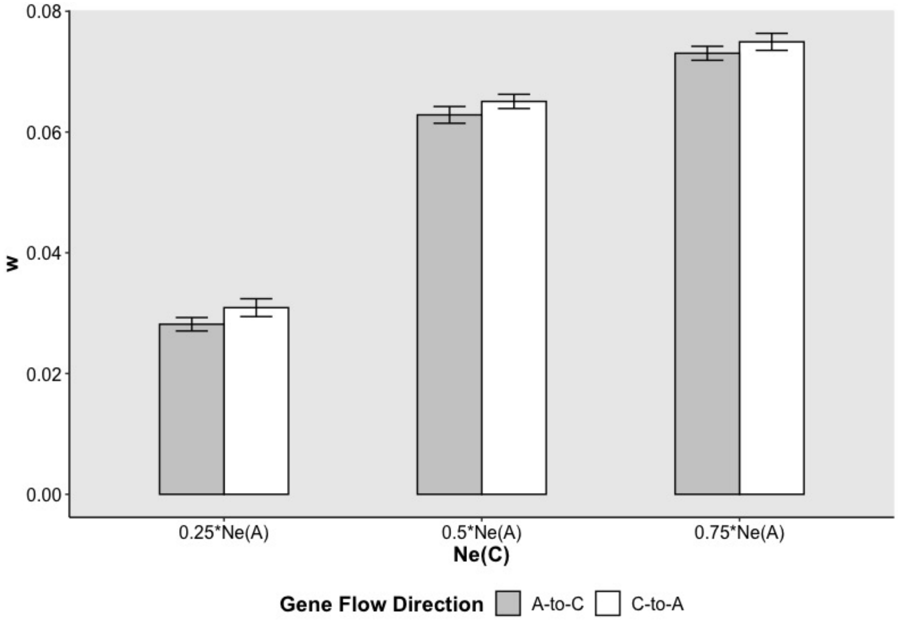

Figure 4.25: Plots using W values for varying population sizes for C when [$\gamma_{\mathrm{AC}} = m/2$,$\gamma_{\mathrm{CA}} = m/2$].

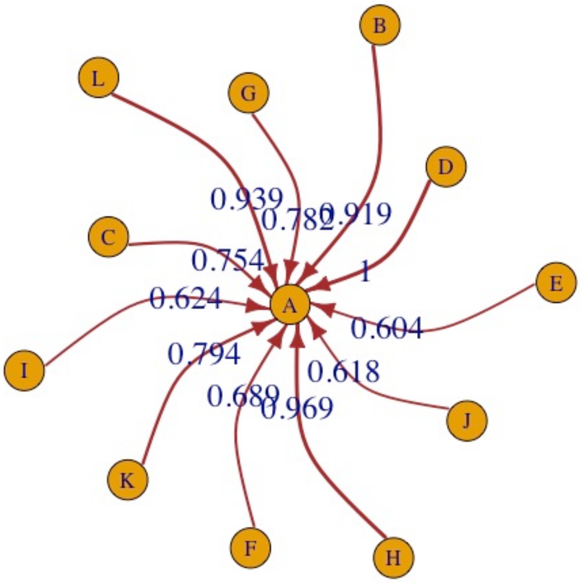

Figure 4.26: Relative gene flow network for the larger dataset with 12 populations (greater than zero edges). The network displays only the edges with $W > 0$; all other edges in the complete graph had $W \leq 0$

4.3.2 Green crab dataset

The green crab dataset was processed using RADProc with default parameter settings ($M = 2$, $m = 3$ and $n = 1$) and *de novo* locus formation and catalog building was performed. The catalog contained a total of 38,435 loci across the 11 populations and only loci that were present in at least 50% of individuals in each population and present in at least 6 populations were retained, to make sure only loci represented in good number of individuals and populations were used for the analysis. Previous studies on the dataset had revealed that two locations (PLB and KJI) intermediate between the northern and southern clusters of populations showed strong evidence of admixture and hybridization [24, 100]. Introduction of a previously admixed population from the Scotian Shelf into PLB due to heavy shipping traffic is likely being maintained [10]. KJI represents a secondary contact region, where the two invasions (north and south) are coming into contact [24].

The relative migration network obtained using *divMigrate* (Figure 4.27) shows two distinct groups (northern and southern populations) as observed in [58]. Apart from these genetic structuring patterns, the network visualization did not show any significant gene flow to populations KJI and PLB as suggested by the length, shading, and thickness of the edges to KJI and PLB from other populations. The edge weights from other populations to KJI ranged from 0.06 to 0.31 and the edge weights to PLB ranged from 0.05 to 0.16. Whereas, the edges among the northern populations and southern populations have high edge weights with most of the edges with weights above 0.5.

On the other hand, the proposed method was able to identify gene flow within the northern and southern population (Figure 4.28) and gene flow from the southern and northern populations to KJI and PLB. The gene flow network indicated gene flow from the southern and northern populations to KJI, given it is geographically intermediate

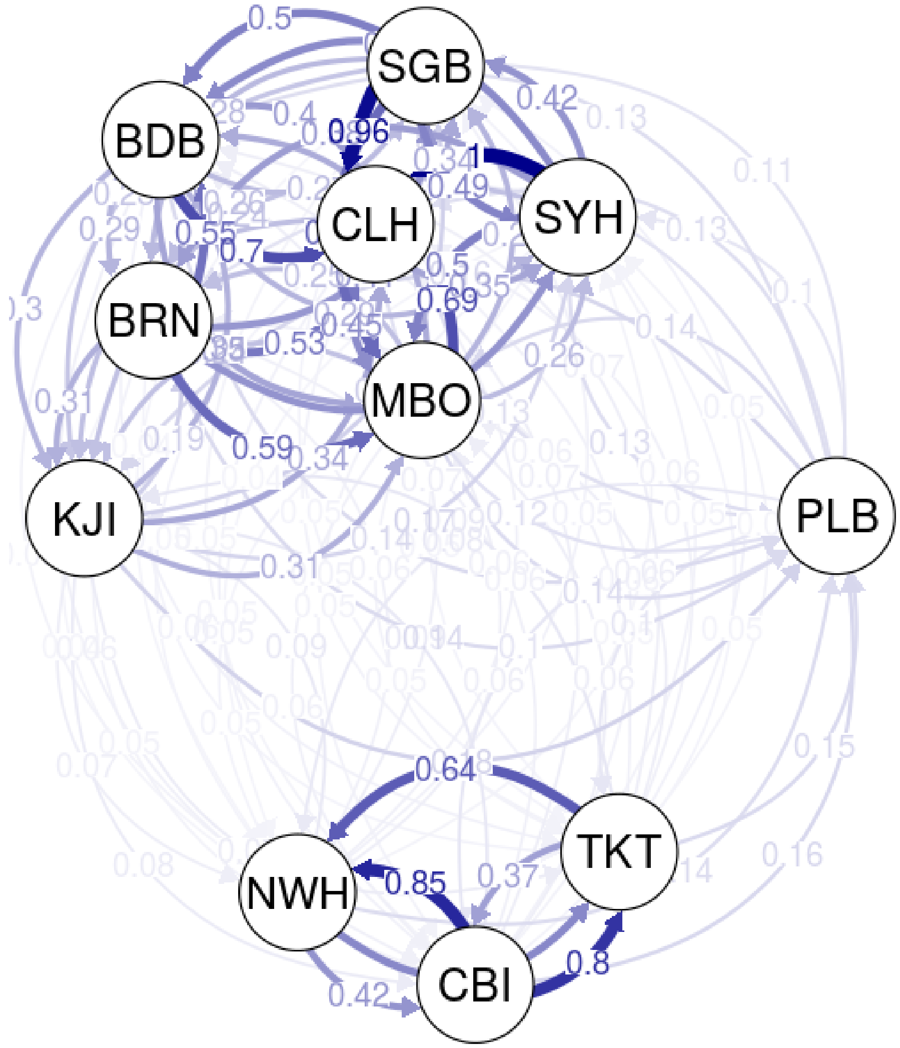

Figure 4.27: Relative gene flow network obtained using *divMigrate* from the green crab dataset.

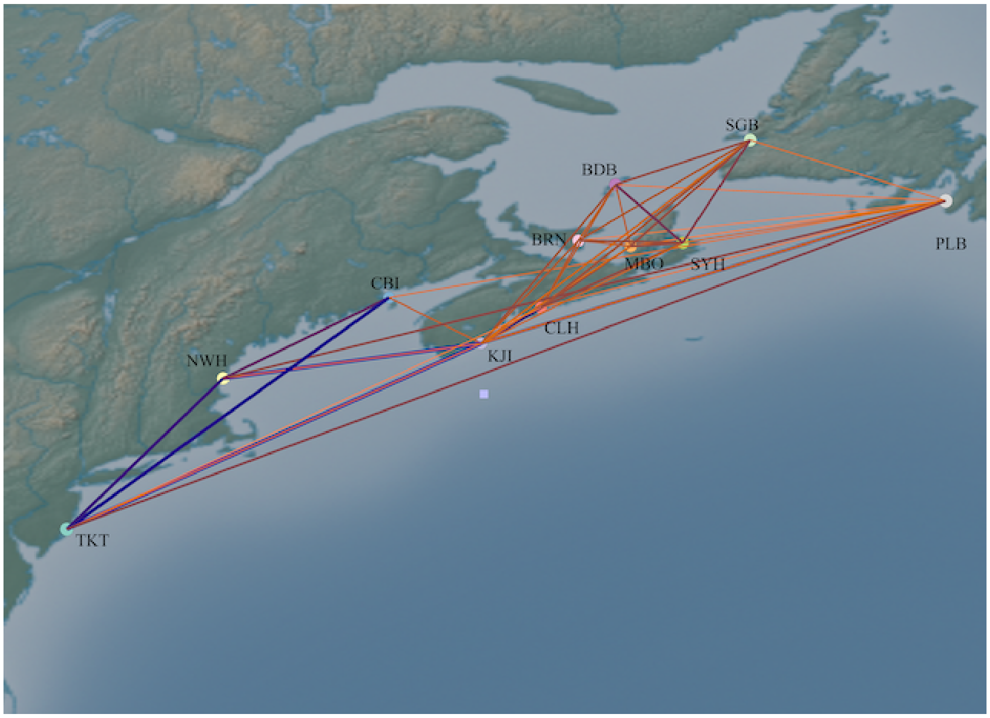

Figure 4.28: Population network generated using GenGIS, showing 11 population sites connected by edges with W greater than 0.3.

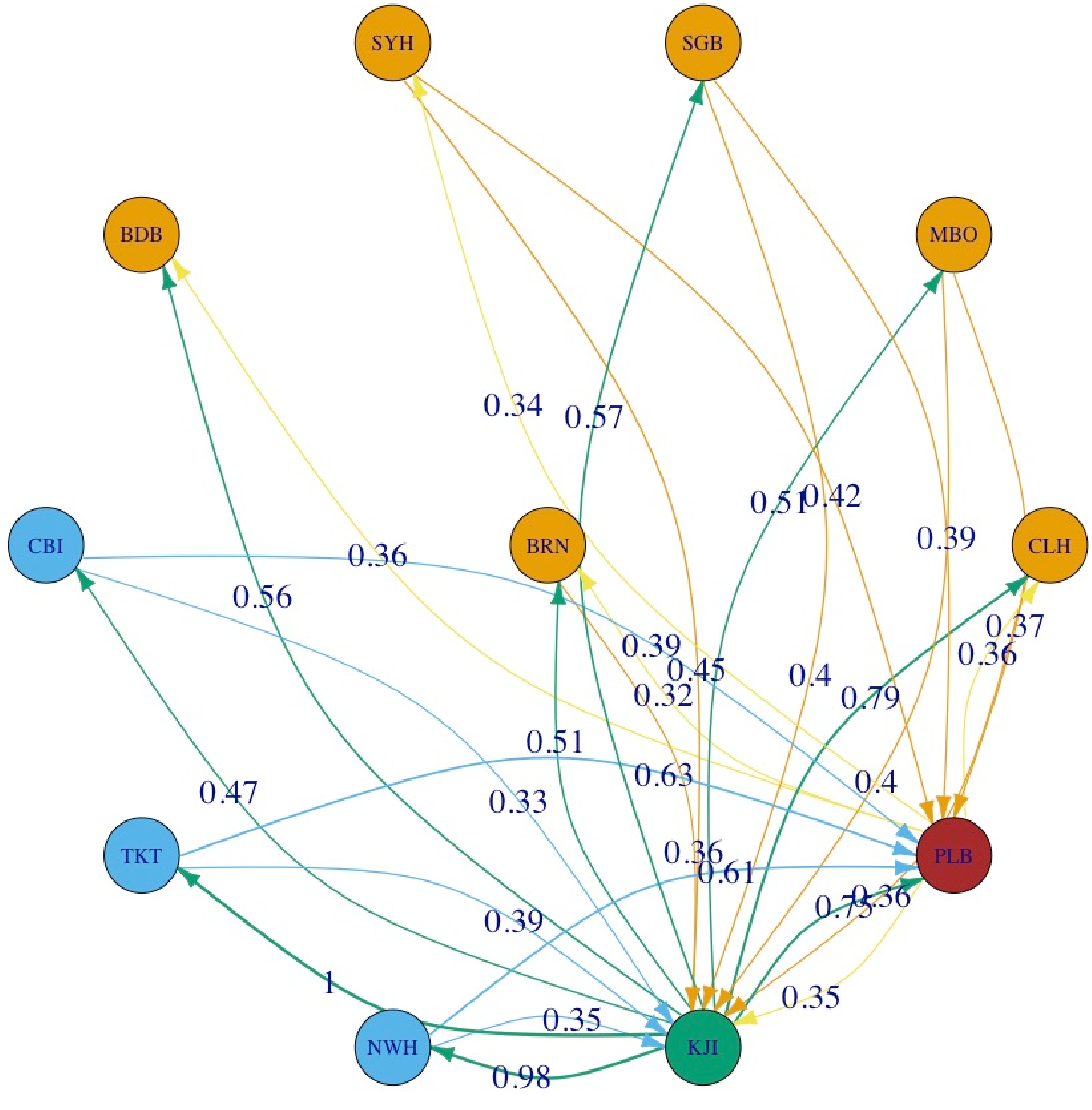

Figure 4.29: Relative gene flow network obtained using the proposed method from the green crab dataset. The W threshold set to 0.3 and only edges to and from KJI and PLB were retained. Orange and blue coloured vertices represent the northern and southern populations respectively.

to the northern and southern populations and also consistent with secondary contact at the region [59]. The gene flow network also revealed gene flow from the southern and northern populations to PLB. Out of the 11 sites in the dataset, the KJI and PLB locations have shown a high level of introgression in a study done by [59]. Figure 4.29 shows the relative gene flow with respect to KJI and PLB, in case of KJI the W values indicate higher level of gene flow from KJI to the souther populations compared to the opposite direction. In case of PLB, we can observe unidirectional gene flow from the southern populations. The relative migration network generated by the proposed method reflects the observations from previous studies and provides additional information about relative rate of gene flow among the populations.

4.4 Discussion

We have demonstrated the possibility of extending the RADProc graph components to estimate relative gene flow among populations. The idea is based on the expectation that if there are alleles present in only a given pair of populations or exclusively shared alleles in a system of populations, then the distribution of allele frequencies of such alleles could be different in the absence and presence of gene flow. The idea can be easily applied to the catalog of loci built using the RADProc graph structure, since the catalog stores all the alleles in all the populations, importantly the populations, in which an allele is present and their corresponding allele frequencies in those populations.

The use of allele frequency data to estimate gene flow can be affected by underestimating alleles present at low frequencies due to sampling effects [41, 40]. The proposed method overcomes this by using exclusive alleles between pairs of populations. The use of exclusive alleles guarantees that underestimating low-frequency alleles would not affect the results as the method uses the allele-frequency differences and their proportions among the exclusive alleles instead of any estimated genetic-differentiation measures such as F_{ST}, G_{ST}, and D.

The method was tested on simulated datasets with different gene flow patterns, migration rates, sample sizes, numbers of loci and population sizes. The tests demonstrate the ability of the proposed method to successfully detect underlying gene flow patterns. The method was able to correctly identify the relative gene flow in case of low ($m = 1e - 06$) and medium ($m = 1e - 05$) level migration rates. The proposed approach did not work well in case of high migration rate ($m = 1e - 04$), likely due to the homogenizing effect of high gene flow, which results in low genetic differentiation between the populations. Though the method was able to correctly find the simulated gene flow for all the different sample sizes, the results showed upward trend in the values of W, because of the improved accuracy in the allele frequency estimations as the sample sizes increased. Unlike varying sample sizes, increasing the number of loci had only minimal impact on the magnitude of W. Applying the method to simulated datasets with uneven population sizes, we observed that the population size impacted the values of W and in some cases the directionality of the gene flow. It will be interesting to see if the results improve by weighing the allele frequencies of populations proportionally to local size [55].

Applying the proposed method to the green-crab dataset, we were able to estimate the relative gene flow among the populations and especially the gene flow to KJI and PLB. Previous studies have observed KJI and PLB to be intermediate between the northern and southern population clusters. The locations KJI and PLB has shown high levels of introgression [59], and the relative gene flow network generated by the proposed method identified asymmetry in the gene flow rates between KJI, PLB and the southern populations.

Though methods such as Migrate [8] and BayesAss [110] can provide estimates of relative gene flow, they are difficult to use correctly, issues with convergence and repeatability of results and these programs are also computationally demanding. Although *divMigrate* could overcome these drawbacks, scenarios such as recent common ancestry could be difficult to resolve using *divMigrate*. The results generated by the proposed method were also compared with the results obtained from *divMigrate*. Unlike *divMigrate*, the proposed method was able to unambiguously distinguish between presence and absence of gene flow and identify relative gene flow in case of both unidirectional and bidirectional gene flow scenarios.

In conclusion, the concept of using exclusively shared alleles between each pair of populations provides a simple and tangible way to estimate relative gene flow among populations. Though the simulations used here do not cover all scenarios in nature, can act as a good tool for evaluating the usefulness of proposed method. The proposed method proves to be a simple and effective tool to understand the gene flow patterns. It would be interesting to examine the effectiveness of the method in the presence of founder effects and when the populations are not in equilibrium. Computing a statistical test to compare the observed allele distributions for r to a theoretical allele distribution predicted under no gene flow could be a possible future work.

Chapter 5

Conclusions

RADSeq is an efficient and cost-effective next-generation sequencing technology for SNP discovery and genotyping and gaining new insights into ecological, evolutionary and conservation-related questions. In order to fully utilize the power of RADSeq techniques, it is important to develop computational methods to efficiently process and extract information from the RADSeq datasets. In this thesis, we have developed graph-based methods to improve data analysis and inference using genome-wide SNP data based on RADSeq short reads. Graphs have been used in biological sequence analysis for a long time, especially in genome assembly and genome alignment for their ability to compactly represent a group of sequences. In this thesis, we have demonstrated different possible uses of representing the short-read data in an undirected graph structure. The methods developed complement different stages of RADSeq data analysis such as de novo locus formation, catalog building, and catalog filtering and population statistics. Specifically, graph-based methods were well-suited to addressing key challenges like distinguishing paralogous sequence variants (PSVs) from true single-nucleotide polymorphisms (SNPs) and accelerating the de novo locus formation process to enable parameter sweeps.

Chapter 2

In Chapter 2, we described the RADProc software package, which can accelerate the *de novo* locus formation and catalog building processes. Restriction-site associated DNA sequencing (RADSeq) is a powerful tool for genotyping of individuals, but the identification of loci and assignment of sequence reads is a crucial and often challenging step. The optimal parameter settings for a given *de novo* RADSeq assembly varies between datasets and can be difficult and computationally expensive to determine. RADProc focuses on the key bottlenecks in accelerating the *de novo* locus formation and catalog-building processes such as redundant and slower sequence-similarity calculations, and processing less abundant and less coverage sequences. RADProc uses a graph data structure to represent all sequence reads and their similarity relationships. Storing sequence-comparison results in a graph eliminated unnecessary and redundant sequence similarity calculations. *De novo* locus formation and catalog building for a given parameter set can be performed on the pre-computed graph, making parameter sweeps far more efficient. RADProc implemented a clustering-based approach for faster sequence similarity calculations. The runtime comparisons using the test datasets showed that RADProc could achieve speeds 10 to 30 times faster than the widely used Stacks software. Comparisons of the de novo loci formed, and catalog built using both the methods demonstrate that RADProc managed to produce 97% to 98% of loci formed by Stacks.

Chapter 3

Chapter 3 (PMERGE) addressed the challenge in differentiating paralogous sequence variants (PSVs) from true single-nucleotide polymorphisms (SNPs) during *de novo* locus formation. Due to high similarity between paralogous sequences, they can be wrongly merged into a single locus, causing difficulty in identifying true allelic variations. PMERGE is a simple and effective tool, which builds a network of catalog loci based on their consensus sequence similarity and clusters the loci based on a threshold similarity. PMERGE is applied to the catalog of loci rather than the de novo loci formed in each sample separately because the paralogs may be merged into a single locus in some but not all samples, allowing us to cluster them based on similarity. Catalog loci that are clustered with at least one other catalog locus are flagged as potential paralogs. Results from the Atlantic salmon (*Salmo salar*) and green crab (*Carcinus maenas*) data sets show that PMERGE was able to identify and remove the majority of paralogous loci.

Chapter 4

In chapter 4, we introduced a novel approach to infer gene flow patterns among populations. The method was able to detect relative gene flow even in asymmetric systems and build networks to reflect underlying gene flow patterns. The ability of the method to differentiate genetic structuring from gene flow patterns was demonstrated using both simulated datasets and an empirical dataset from green crab (*Carcinus maenas*). The results from the simulated datasets showed that the method can work even in the case of recent common ancestry. The approach is based on using the allele frequency differences between exclusively shared alleles between each pair of populations in a given group of populations. Using exclusive alleles can eliminate the confounding effects of shared alleles on the gene flow detection and issues of underestimating low frequency alleles.

The large-scale data generated by next-generation sequencing data enables understanding complex genomic structures and making population level inferences. Graph-based structures are a straightforward way to represent these data and the relationship between them in terms of sequence similarity. The graph structures also provide the flexibility of representing different types of sequences and the level of information contained in the graph. By understanding the underlying relationship among the population level short-read data and representing them in well-designed graph structure can help develop sophisticated and effective computational methods to process these data and make new inferences. In RADProc, we used a graph structure where each vertex contained all relevant information about a unique stack and the edges represented the nucleotide distances between unique stacks. By contrast, in the PMERGE graph structure, the vertices were relatively simple and just contained the catalog loci id and the edges represented the nucleotide distances between the catalog loci. In chapter 4, we made use of the population-level information contained in the RADProc graph structure to infer gene flow patterns and relative gene flow rates among the populations.

Future Work

RADProc supports the processing of single-end sequence data from different RAD-Seq protocols and performs parameter sweeps for *de novo* locus formation and catalog building. In single-end sequencing only the one end of the DNA fragment is sequenced, but to obtain longer contigs from the DNA fragments paired-end sequencing is used, which perform sequencing from both the ends of the DNA fragment. A useful extension to the RADProc software package would be to enable paired-end sequence processing.

In chapter 4, an important enhancement would be computing a statistical test to compare the observed allele distributions for r to a theoretical gene distribution predicted under no gene flow. In this thesis, the method was applied only to the major gene flow patterns; applying the method to other empirical datasets and simulations can provide more insights into the ability and limitations of the method.

Using the RADProc graph structure we were able to accelerate the de novo loci formation and enable parameter searches in realistic time. We were also successfully able to extend RADProc to estimate relative gene flow rates among populations. Since the RADProc graph stores all the putative alleles and their relationships such as the similarity between the alleles, populations sharing a given allele, coverage depth and their abundance in each population it is worth exploring other applications of the RADProc graph.

Bibliography

[1] Abadia-cardoso, Alicia, Clemento, Anthony J., and Garza, John Carlos. Discovery and characterization of single-nucleotide polymorphisms in steelhead/rainbow trout, oncorhynchus mykiss. *Molecular Ecology Resources*, 11(s1):31–49, 2011.

[2] Alcala, Nicolas, Goudet, Jérôme, and Vuilleumier, Séverine. On the transition of genetic differentiation from isolation to panmixia: What we can learn from gst and d. *Theoretical Population Biology*, 93:75 – 84, 2014.

[3] Allendorf, Fred W. Genetics and the conservation of natural populations: allozymes to genomes. *Molecular Ecology*, 26(2):420–430, 2017.

[4] Altschul, Stephen F., Gish, Warren, Miller, Webb, Myers, Eugene W., and Lipman, David J. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403 – 410, 1990.

[5] Altshuler, David, Pollara, Victor J., Cowles, Chris R., Van Etten, William J., Baldwin, Jennifer, Linton, Lauren, and Lander, Eric S. An snp map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407(6803):513–516, 2000.

[6] Andrews, Kimberly R., Good, Jeffrey M., Miller, Michael R., Luikart, Gordon, and Hohenlohe, Paul A. Harnessing the power of radseq for ecological and evolutionary genomics. *Nature reviews. Genetics*, 17(2):81 – 92, 2016.

[7] Baird, Nathan A., Etter, Paul D., Atwood, Tressa S., Currey, Mark C., Shiver, Anthony L., Lewis, Zachary A., Selker, Eric U., Cresko, William A., and Johnson, Eric A. Rapid snp discovery and genetic mapping using sequenced rad markers. *PLOS ONE*, 3(10):1–7, 2008.

[8] Beerli, Peter. 3 how to use migrate or why are markov chain monte carlo programs difficult to use?, 2009.

[9] Benestan, Laura, Gosselin, Thierry, Perrier, Charles, Sainte-Marie, Bernard, Rochette, Rémy, and Bernatchez, Louis. Rad genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the american lobster (homarus americanus). *Molecular Ecology*, 24(13):3299–3315, 2015.

[10] Blakeslee, April M. H., McKenzie, Cynthia H., Darling, John A., Byers, James E., Pringle, James M., and Roman, Joe. A hitchhiker's guide to the maritimes: anthropogenic transport facilitates long-distance dispersal of an invasive marine crab to newfoundland. *Diversity and Distributions*, 16(6):879–891, 2010.

[11] Boc, Alix, Diallo, Alpha Boubacar, and Makarenkov, Vladimir. T-rex: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic acids research*, 40(Web Server issue):W573–W579, 2012. 22675075[pmid].

[12] Bolnick, Daniel I. and Nosil, Patrik. Natural selection in populations subject to a migration load. *Evolution*, 61(9):2229–2243, 2007.

[13] Bradbury, Ian R., Hamilton, Lorraine C., Dempson, Brian, Robertson, Martha J., Bourret, Vincent, Bernatchez, Louis, and Verspoor, Eric. Transatlantic secondary contact in atlantic salmon, comparing microsatellites, a single nucleotide polymorphism array and restriction-site associated dna sequencing for the resolution of complex spatial structure. *Molecular Ecology*, 24(20):5130–5144, 2015.

[14] Bradbury, Ian R., Hamilton, Lorraine C., Sheehan, Timothy F., Chaput, Gerald, Robertson, Martha J., Dempson, J. Brian, Reddin, David, Morris, Vicki, King, Timothy, and Bernatchez, Louis. Genetic mixed-stock analysis disentangles spatial and temporal variation in composition of the West Greenland Atlantic Salmon fishery. *ICES Journal of Marine Science*, 73(9):2311–2321, 2016.

[15] Brookes, Anthony J. The essence of snps. *Gene*, 234(2):177–186, 1999.

[16] Catchen, Julian, Bassham, Susan, Wilson, Taylor, Currey, Mark, O'Brien, Conor, Yeates, Quick, and Cresko, William A. The population structure and recent colonization history of oregon threespine stickleback determined using restriction-site associated dna-sequencing. *Molecular Ecology*, 22(11):2864–2883, 2013.

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

[18] Catchen, Julian M., Amores, Angel, Hohenlohe, Paul, Cresko, William, and Postlethwait, John H. Stacks: building and genotyping loci de novo from short-read sequences. *G3 (Bethesda, Md.)*, 1(3):171–182, 2011. 22384329[pmid].

[19] Chen, Hua. Population genetic studies in the genomic sequencing era. *Dong wu xue yan jiu = Zoological research*, 36(4):223–232, 2015. 26228473[pmid].

[20] Choudhuri, Supratim. *Bioinformatics for beginners : genes, genomes, molecular evolution, databases and analytical tools*. London, London, 2014.

[21] Christensen, Kris A., Brunelli, Joseph P., Lambert, Matthew J., DeKoning, Jenefer, Phillips, Ruth B., and Thorgaard, Gary H. Identification of single nucleotide polymorphisms from the transcriptome of an organism with a whole genome duplication. *BMC bioinformatics*, 14:325–325, 2013. 24237905[pmid].

[22] Cortázar-Chinarro, Maria, Lattenkamp, Ella Z., Meyer-Lucht, Yvonne, Luquet, Emilien, Laurila, Anssi, and Höglund, Jacob. Drift, selection, or migration?: Processes affecting genetic differentiation and variation along a latitudinal gradient in an amphibian. *BMC Evolutionary Biology*, 17(1):189, 2017.

[23] Crawford, Nicholas G. smogd: software for the measurement of genetic diversity. *Molecular Ecology Resources*, 10(3):556–557, 2010.

[24] Darling, John A., Tsai, Yi-Hsin Erica, Blakeslee, April M. H., and Roman, Joe. Are genes faster than crabs?: Mitochondrial introgression exceeds larval dispersal during population expansion of the invasive crab carcinus maenas. *Royal Society open science*, 1(2):140202–140202, 2014. 26064543[pmid].

[25] Davey, John W., Hohenlohe, Paul A., Etter, Paul D., Boone, Jason Q., Catchen, Julian M., and Blaxter, Mark L. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12:499 EP –, 2011. Review Article.

[26] David, Lior, Rosenberg, Noah A., Lavi, Uri, Feldman, Marcus W., and Hillel, Jossi. Genetic diversity and population structure inferred from the partially duplicated genome of domesticated carp, cyprinus carpio l. *Genetics Selection Evolution*, 39(3):319, 2007.

[27] Davidson, William S., Koop, Ben F., Jones, Steven J. M., Iturra, Patricia, Vidal, Rodrigo, Maass, Alejandro, Jonassen, Inge, Lien, Sigbjorn, and Omholt, Stig W. Sequencing the genome of the atlantic salmon (salmo salar). *Genome biology*, 11(9):403–403, 2010. 20887641[pmid].

[28] De Mita, Stéphane and Siol, Mathieu. Egglib: processing, analysis and simulation tools for population genetics and genomics. *BMC genetics*, 13:27–27, 2012. 22494792[pmid].

[29] Dias, Paula C. Sources and sinks in population biology. *Trends in Ecology & Evolution*, 11(8):326–330, 1996.

[30] DiBattista, Joseph D., Saenz-Agudelo, Pablo, Piatek, Marek J., Wang, Xin, Aranda, Manuel, and Berumen, Michael L. Using a butterflyfish genome as a general tool for rad-seq studies in specialized reef fish. *Molecular Ecology Resources*, 17(6):1330–1341, 2017.

[31] Dou, Jinzhuang, Zhao, Xiqiang, Fu, Xiaoteng, Jiao, Wenqian, Wang, Nannan, Zhang, Lingling, Hu, Xiaoli, Wang, Shi, and Bao, Zhenmin. Reference-free snp calling: improved accuracy by preventing incorrect calls from repetitive genomic regions. *Biology Direct*, 7(1):17, 2012.

[32] Dyer, Rodney J. and Nason, John D. Population graphs: the graph theoretic shape of genetic structure. *Molecular Ecology*, 13(7):1713–1727, 2004.

[33] Eaton, Deren A. R. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses . *Bioinformatics*, 30(13):1844–1849, 2014.

[34] Eaton, Deren A. R. and Ree, Richard H. Inferring phylogeny and introgression using radseq data: an example from flowering plants (pedicularis: Orobanchaceae). *Systematic biology*, 62(5):689–706, 2013. 23652346[pmid].

[35] Edgar, Robert C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 2010.

[36] Elshire, Robert J., Glaubitz, Jeffrey C., Sun, Qi, Poland, Jesse A., Kawamoto, Ken, Buckler, Edward S., and Mitchell, Sharon E. A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. *PLOS ONE*, 6(5):1–10, 2011.

[37] Epps, Clinton W. and Keyghobadi, Nusha. Landscape genetics in a changing world: disentangling historical and contemporary influences and inferring change. *Molecular Ecology*, 24(24):6021–6040, 2015.

[38] Faubet, Pierre, Waples, Robin S., and Gaggiotti, Oscar E. Evaluating the performance of a multilocus bayesian method for the estimation of migration rates. *Molecular Ecology*, 16(6):1149–1166, 2007.

[39] Fumagalli, Matteo, Vieira, Filipe G., Korneliussen, Thorfinn Sand, Linderoth, Tyler, Huerta-Sánchez, Emilia, Albrechtsen, Anders, and Nielsen, Rasmus. Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, 195(3):979–992, 2013.

[40] Fung, Tak and Keenan, Kevin. Confidence intervals for population allele frequencies: the general case of sampling from a finite diploid population of any size. *PLOS ONE*, 9(1):e85925–e85925, 2014. 24465792[pmid].

[41] Gautier, Mathieu, Foucaud, Julien, Gharbi, Karim, Cézard, Timothée, Galan, Maxime, Loiseau, Anne, Thomson, Marian, Pudlo, Pierre, Kerdelhué, Carole, and Estoup, Arnaud. Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology*, 22(14):3766–3779, 2013.

[42] Gnerre, Sante, MacCallum, Iain, Przybylski, Dariusz, Ribeiro, Filipe J., Burton, Joshua N., Walker, Bruce J., Sharpe, Ted, Hall, Giles, Shea, Terrance P., Sykes, Sean, Berlin, Aaron M., Aird, Daniel, Costello, Maura, Daza, Riza, Williams, Louise, Nicol, Robert, Gnirke, Andreas, Nusbaum, Chad, Lander, Eric S., and Jaffe, David B. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, 108(4):1513–1518, 2011.

[43] Gonen, Serap, Lowe, Natalie R., Cezard, Timothé, Gharbi, Karim, Bishop, Stephen C., and Houston, Ross D. Linkage maps of the atlantic salmon (salmo salar) genome derived from rad sequencing. *BMC genomics*, 15:166–166, 2014. 24571138[pmid].

[44] Guo, Yu, Yuan, Hui, Fang, Dongming, Song, Lianbo, Liu, Yan, Liu, Yong, Wu, Lu, Yu, Jianping, Li, Zichao, Xu, Xun, and Zhang, Hongliang. An improved 2b-rad approach (i2b-rad) offering genotyping tested by a rice (oryza sativa l.) f2 population. *BMC Genomics*, 15(1):956, 2014.

[45] Harvey, Michael G., Judy, Caroline Duffie, Seeholzer, Glenn F., Maley, James M., Graves, Gary R., and Brumfield, Robb T. Similarity thresholds used in dna sequence assembly from short reads can reduce the comparability of population histories across species. *PeerJ*, 3:e895, 2015.

[46] Hedrick, Philip W. A standardized genetic differentiation measure. *Evolution*, 59(8):1633–1638, 2005.

[47] Hein, Jotun, Jiang, Tao, Wang, Lusheng, and Zhang, Kaizhong. On the complexity of comparing evolutionary trees. *Discrete Applied Mathematics*, 71(1):153 – 169, 1996.

[48] Hernandez, David, François, Patrice, Farinelli, Laurent, Osterås, Magne, and Schrenzel, Jacques. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome research*, 18(5):802–809, 2008. 18332092[pmid].

[49] Ho, Meng-Ru, Tsai, Kuo-Wang, Chen, Chun-houh, and Lin, Wen-chang. dbdnv: a resource of duplicated gene nucleotide variants in human genome. *Nucleic acids research*, 39(Database issue):D920–D925, 2011. 21097891[pmid].

[50] Hohenlohe, Paul A., Amish, Stephen J., Catchen, Julian M., Allendorf, Fred W., and Luikart, Gordon. Next-generation rad sequencing identifies thousands of snps for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, 11(s1):117–122, 2011.

[51] Hohenlohe, Paul A., Bassham, Susan, Etter, Paul D., Stiffler, Nicholas, Johnson, Eric A., and Cresko, William A. Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. *PLOS Genetics*, 6(2):1–23, 2010.

[52] Holsinger, Kent E. and Weir, Bruce S. Genetics in geographically structured populations: defining, estimating and interpreting fst. *Nature Reviews Genetics*, 10:639 EP –, 2009. Review Article.

[53] Hurles, Matthew. Gene duplication: The genomic trade in spare parts. *PLOS Biology*, 2(7), 2004.

[54] Hyten, David L., Cannon, Steven B., Song, Qijian, Weeks, Nathan, Fickus, Edward W., Shoemaker, Randy C., Specht, James E., Farmer, Andrew D., May, Gregory D., and Cregan, Perry B. High-throughput snp discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics*, 11(1):38, 2010.

[55] Hössjer, Ola, Olsson, Fredrik, Laikre, Linda, and Ryman, Nils. A new general analytical approach for modeling patterns of genetic differentiation and effective size of subdivided populations over time. *Mathematical Biosciences*, 258:113 – 133, 2014.

[56] Ilut, Daniel C., Nydam, Marie L., and Hare, Matthew P. Defining Loci in Restriction-Based Reduced Representation Genomic Data from Nonmodel Species: Sources of Bias and Diagnostics for Optimal Clustering. *BioMed Research International*, 2014:9, 2014.

[57] Jahnke, Marlene, Jonsson, Per R., Moksnes, Per-Olav, Loo, Lars-Ove, Nilsson Jacobi, Martin, and Olsen, Jeanine L. Seascape genetics and biophysical connectivity modelling support conservation of the seagrass zostera marina in the skagerrak-kattegat region of the eastern north sea. *Evolutionary Applications*, 11(5):645–661, 2018.

[58] Jeffery, Nicholas W., DiBacco, Claudio, Van Wyngaarden, Mallory, Hamilton, Lorraine C., Stanley, Ryan R. E., Bernier, Renée, FitzGerald, Jennifer, Matheson, K., McKenzie, C. H., Nadukkalam Ravindran, Praveen, Beiko, Robert, and Bradbury, Ian R. Rad sequencing reveals genomewide divergence between independent invasions of the european green crab (carcinus maenas) in the northwest atlantic. *Ecology and Evolution*, 7(8):2513–2524, 2017.

[59] Jeffery, Nicholas W., DiBacco, Claudio, Wringe, Brendan F., Stanley, Ryan R. E., Hamilton, Lorraine C., Ravindran, Praveen N., and Bradbury, Ian R. Genomic evidence of hybridization between two independent invasions of european green crab (carcinus maenas) in the northwest atlantic. *Heredity*, 119(3):154–165, 2017.

[60] Jones, Julia C., Fan, Shaohua, Franchini, Paolo, Schartl, Manfred, and Meyer, Axel. The evolutionary history of xiphophorus fish and their sexually selected sword: a genome-wide approach using restriction site-associated dna sequencing. *Molecular Ecology*, 22(11):2986–3001, 2013.

[61] Jost, Lou. Gst and its relatives do not measure differentiation. *Molecular Ecology*, 17(18):4015–4026, 2008.

[62] Karki, Roshan, Pandya, Deep, Elston, Robert C., and Ferlini, Cristiano. Defining ”mutation” and ”polymorphism” in the era of personal genomics. *BMC medical genomics*, 8:37–37, 2015. 26173390[pmid].

[63] Keenan, Kevin, McGinnity, Philip, Cross, Tom F., Crozier, Walter W., and Prodöhl, Paulo A. diversity: An r package for the estimation and exploration of population genetics parameters and their associated errors. *Methods in Ecology and Evolution*, 4(8):782–788, 2013.

[64] Kurtz, Stefan, Narechania, Apurva, Stein, Joshua C., and Ware, Doreen. A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics*, 9(1):517, 2008.

[65] Langmead, Ben, Trapnell, Cole, Pop, Mihai, and Salzberg, Steven L. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[66] Leaché, Adam D., Chavez, Andreas S., Jones, Leonard N., Grummer, Jared A., Gottscho, Andrew D., and Linkem, Charles W. Phylogenomics of Phrynosomatid Lizards: Conflicting Signals from Sequence Capture versus Restriction Site Associated DNA Sequencing. *Genome Biology and Evolution*, 7(3):706–719, 2015.

[67] Li, Heng and Durbin, Richard. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760, 2009.

[68] Lien, Sigbjørn, Koop, Ben F., Sandve, Simen R., Miller, Jason R., Kent, Matthew P., Nome, Torfinn, Hvidsten, Torgeir R., Leong, Jong S., Minkley, David R., Zimin, Aleksey, Grammes, Fabian, Grove, Harald, Gjuvsland, Arne, Walenz, Brian, Hermansen, Russell A., von Schalburg, Kris, Rondeau, Eric B., Di Genova, Alex, Samy, Jeevan K. A., Olav Vik, Jon, Vigeland, Magnus D., Caler, Lis, Grimholt, Unni, Jentoft, Sissel, Inge Våge, Dag, de Jong, Pieter, Moen, Thomas, Baranski, Matthew, Palti, Yniv, Smith, Douglas R., Yorke, James A., Nederbragt, Alexander J., Tooming-Klunderud, Ave, Jakobsen, Kjetill S., Jiang, Xuanting, Fan, Dingding, Hu, Yan, Liberles, David A., Vidal, Rodrigo, Iturra, Patricia, Jones, Steven J. M., Jonassen, Inge, Maass, Alejandro, Omholt, Stig W., and Davidson, William S. The atlantic salmon genome provides insights into rediploidization. *Nature*, 533:200 EP –, 2016. Article.

[69] Limborg, Morten Tønsberg, Seeb, Lisa W., and Seeb, James E. Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. *Molecular Ecology*, 25(10):2117–2129, 2016.

[70] Lu, Fei, Lipka, Alexander E., Glaubitz, Jeff, Elshire, Rob, Cherney, Jerome H., Casler, Michael D., Buckler, Edward S., and Costich, Denise E. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based snp discovery protocol. *PLOS Genetics*, 9(1):1–14, 2013.

[71] Macqueen, Daniel J. and Johnston, Ian A. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proceedings of the Royal Society B: Biological Sciences*, 281(1778):20132881, 2014.

[72] Marçais, Guillaume and Kingsford, Carl. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011.

[73] Massip, Florian, Sheinman, Michael, Schbath, Sophie, and Arndt, Peter F. How evolution of genomes is reflected in exact dna sequence match statistics. *Molecular Biology and Evolution*, 32(2):524–535, 2014.

[74] Mastretta-Yanes, Alicia, Arrigo, Nils, Alvarez, Nadir, Jorgensen, Tove H., Piñero, Daniel, and Emerson, Brent C. Restriction site-associated dna sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, 15(1):28–41, 2015.

[75] Mastretta-Yanes, Alicia, Zamudio, Sergio, Jorgensen, Tove H., Arrigo, Nils, Alvarez, Nadir, Piñero, Daniel, and Emerson, Brent C. Gene duplication, population genomics, and species-level differentiation within a tropical mountain shrub. *Genome biology and evolution*, 6(10):2611–2624, 2014. 25223767[pmid].

[76] McKinney, Garrett J., Waples, Ryan K., Seeb, Lisa W., and Seeb, James E. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, 17(4):656–669, 2017.

[77] Meirmans, Patrick G. Nonconvergence in bayesian estimation of migration rates. *Molecular Ecology Resources*, 14(4):726–733, 2014.

[78] Melsted, Páll and Pritchard, Jonathan K. Efficient counting of k-mers in dna sequences using a bloom filter. *BMC Bioinformatics*, 12(1):333, 2011.

[79] Miller, Michael R., Dunham, Joseph P., Amores, Angel, Cresko, William A., and Johnson, Eric A. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated dna (rad) markers. *Genome research*, 17(2):240–248, 2007. 17189378[pmid].

[80] Nei, Masatoshi. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, 70(12):3321–3323, 1973.

[81] Nielsen, Rasmus, Korneliussen, Thorfinn, Albrechtsen, Anders, Li, Yingrui, and Wang, Jun. Snp calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLOS ONE*, 7(7):1–10, 2012.

[82] Paris, Josephine R., Stevens, Jamie R., and Catchen, Julian Michael. Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution*, 8(10):1360–1373, 2017.

[83] Parks, Donovan H., Mankowski, Timothy, Zangooei, Somayyeh, Porter, Michael S., Armanini, David G., Baird, Donald J., Langille, Morgan G. I., and Beiko, Robert G. Gengis 2: Geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework. *PLOS ONE*, 8(7):1–10, 2013.

[84] Peterson, Brant K., Weber, Jesse N., Kay, Emily H., Fisher, Heidi S., and Hoekstra, Hopi E. Double digest radseq: An inexpensive method for de novo snp discovery and genotyping in model and non-model species. *PLOS ONE*, 7(5):1–11, 2012.

[85] Pichler, Verena, Kotsakiozi, Panayiota, Caputo, Beniamino, Serini, Paola, Caccone, Adalgisa, and della Torre, Alessandra. Complex interplay of evolutionary forces shaping population genomic structure of invasive aedes albopictus in southern europe. *PLOS Neglected Tropical Diseases*, 13(8):1–24, 2019.

[86] Pringle, James M., Blakeslee, April M. H., Byers, James E., and Roman, Joe. Asymmetric dispersal allows an upstream region to control population structure throughout a species' range. *Proceedings of the National Academy of Sciences*, 108(37):15288–15293, 2011.

[87] Pukk, Lilian, Ahmad, Freed, Hasan, Shihab, Kisand, Veljo, Gross, Riho, and Vasemägi, Anti. Less is more: extreme genome complexity reduction with ddrad using ion torrent semiconductor technology. *Molecular Ecology Resources*, 15(5):1145–1152, 2015.

[88] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[89] Richards, Paul M., Liu, M. Maureen, Lowe, Natalie, Davey, John W., Blaxter, Mark L., and Davison, Angus. Rad-seq derived markers flank the shell colour and banding loci of the cepaea nemoralis supergene. *Molecular Ecology*, 22(11):3077–3089, 2013.

[90] Robinson, D.F. and Foulds, L.R. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131 – 147, 1981.

[91] Rodríguez-Ezpeleta, Naiara, Bradbury, Ian R., Mendibil, Iñaki, Álvarez, Paula, Cotano, Unai, and Irigoien, Xabier. Population structure of atlantic mackerel inferred from rad-seq-derived snp markers: effects of sequence clustering parameters and hierarchical snp selection. *Molecular Ecology Resources*, 16(4):991–1001, 2016.

[92] Schilling, Martin P., Gompert, Zachariah, Li, Fay-Wei, Windham, Michael D., and Wolf, Paul G. Admixture, evolution, and variation in reproductive isolation in the boechera puberula clade. *BMC evolutionary biology*, 18(1):61–61, 2018. 29699502[pmid].

[93] Shen, Peidong, Wang, Wenyi, Chi, Aung-Kyaw, Fan, Yu, Davis, Ronald W., and Scharfe, Curt. Multiplex target capture with double-stranded dna probes. *Genome Medicine*, 5(5):50, 2013.

[94] Shendure, Jay and Ji, Hanlee. Next-generation dna sequencing. *Nature Biotechnology*, 26:1135 EP –, 2008.

[95] Simpson, Jared T., Wong, Kim, Jackman, Shaun D., Schein, Jacqueline E., Jones, Steven J. M., and Birol, Inanç. Abyss: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–1123, 2009. 19251739[pmid].

[96] Slatkin, Montgomery. Rare alleles as indicators of gene flow. *Evolution*, 39(1):53–65, 1985.

[97] Sovic, Michael G., Fries, Anthony C., and Gibbs, H. Lisle. Aftrrad: a pipeline for accurate and efficient de novo assembly of radseq data. *Molecular Ecology Resources*, 15(5):1163–1171, 2015.

[98] Sundqvist, Lisa, Keenan, Kevin, Zackrisson, Martin, Prodöhl, Paulo, and Kleinhans, David. Directional genetic differentiation and relative migration. *Ecology and evolution*, 6(11):3461–3475, 2016. 27127613[pmid].

[99] Templeton, Alan R. Chapter 6 - gene flow and subdivided populations. In Templeton, Alan R., editor, *Human Population Genetics and Genomics*, pages 155 – 193. Academic Press, San Diego, 2019.

[100] Tepolt, Carolyn K. and Palumbi, Stephen R. Transcriptome sequencing reveals both neutral and adaptive genome dynamics in a marine invader. *Molecular Ecology*, 24(16):4145–4158, 2015.

[101] Toonen, Robert J., Puritz, Jonathan B., Forsman, Zac H., Whitney, Jonathan L., Fernandez-Silva, Iria, Andrews, Kimberly R., and Bird, Christopher E. ezrad: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, 1:e203–e203, 2013. 24282669[pmid].

[102] Torkamaneh, Davoud, Laroche, Jérôme, and Belzile, François. Genome-wide snp calling from genotyping by sequencing (gbs) data: A comparison of seven pipelines and two sequencing technologies. *PLOS ONE*, 11(8):1–14, 08 2016.

[103] van Orsouw, Nathalie J., Hogers, René C. J., Janssen, Antoine, Yalcin, Feyruz, Snoeijers, Sandor, Verstege, Esther, Schneiders, Harrie, van der Poel, Hein, van Oeveren, Jan, Verstegen, Harold, and van Eijk, Michiel J. T. Complexity reduction of polymorphic sequences (crops™): A novel approach for large-scale polymorphism discovery in complex genomes. *PLOS ONE*, 2(11):1–10, 2007.

[104] Wang, Shi, Meyer, Eli, McKay, John K., and Matz, Mikhail V. 2b-rad: a simple and flexible method for genome-wide genotyping. *Nature Methods*, 9:808 EP –, 2012.

[105] Wang, Xueqin, Ye, Xiaying, Zhao, Lei, Li, Dezhu, Guo, Zhenhua, and Zhuang, Huifu. Genome-wide rad sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (poaceae: Bambusoideae). *Scientific reports*, 7(1):11546–11546, 2017. 28912480[pmid].

[106] Waples, Ryan K., Seeb, Lisa W., and Seeb, James E. Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (oncorhynchus keta). *Molecular Ecology Resources*, 16(1):17–28, 2016.

[107] Weir, Bruce S. and Cockerham, Columbus C. Estimating f-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370, 1984.

[108] Whidden, Christopher, Beiko, Robert G., and Zeh, Norbert. Fixed-parameter algorithms for maximum agreement forests. *SIAM Journal on Computing*, 42(4):1431–1466, 2013.

[109] Willis, Stuart C., Hollenbeck, Christopher M., Puritz, Jonathan B., Gold, John R., and Portnoy, David S. Haplotyping rad loci: an efficient method to filter paralogs and account for physical linkage. *Molecular Ecology Resources*, 17(5):955–965, 2017.

[110] Wilson, Gregory A. and Rannala, Bruce. Bayesian inference of recent migration rates using multilocus genotypes. *Genetics*, 163(3):1177–1191, 2003. 12663554[pmid].

[111] Wright, Sewall. Evolution in mendelian populations. *Genetics*, 16(2):97–159, 1931. 17246615[pmid].

[112] Wright, Sewall. Isolation by distance. *Genetics*, 28(2):114, 1943.

[113] Wright, Sewall. The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354, 1949.

[114] Xuereb, Amanda, Benestan, Laura, Normandeau, Éric, Daigle, Rémi M., Curtis, Janelle M. R., Bernatchez, Louis, and Fortin, Marie-Josée. Asymmetric oceanographic processes mediate connectivity and population genetic structure, as revealed by radseq, in a highly dispersive marine invertebrate (parastichopus californicus). *Molecular Ecology*, 27(10):2347–2364, 2018.

[115] Zerbino, Daniel R. and Birney, Ewan. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829, 2008. 18349386[pmid].

Appendix A

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Feb 03, 2020

This Agreement between Praveen Nadukkalam Ravindran ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4761340724895
License date	Feb 03, 2020
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Molecular Ecology Resources
Licensed Content Title	RADProc: A computationally efficient de novo locus assembler for population studies using RADseq data
Licensed Content Author	Praveen Nadukkalam Ravindran, Paul Bentzen, Ian R. Bradbury, et al
Licensed Content Date	Dec 21, 2018
Licensed Content Volume	19
Licensed Content Issue	1

Licensed Content Pages	11
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Title of your thesis / dissertation	COMPUTATIONAL METHODS FOR EFFICIENT PROCESSING AND ANALYSIS OF SHORT-READ NEXT-GENERATION DNA SEQUENCING DATA
Expected completion date	Mar 2020
Expected size (number of pages)	142
Requestor Location	Praveen Nadukkalam Ravindran Apt# 709 41 Cowie Hill Rd HALIFAX, NS B3P 2M7 Canada Attn: Praveen Nadukkalam Ravindran
Publisher Tax ID	EU826007151
Total	0.00 CAD

Terms and Conditions

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library
http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.