Signature Page

Network Analysis as a Potential Method Detecting Population Structure

By
Yu Ting (Kiera) Chang

Supervised By
Dr. Timothy Frasier

A Thesis Submitted to
Saint Mary's University, Halifax, Nova Scotia
in Partial Fulfillment of the Requirements for
the Degree of Bachelor of Science.

March 17, 2017, Halifax, Nova Scotia

Approved: Dr. Timothy Frasier
Supervisor
Associated Professor of Biology

Approved: Dr. Zhongmin Dong
Reader
Professor of Biology

Date: April 21, 2017

Network Analysis as a Potential Method Detecting Population Structure

By
Yu Ting (Kiera) Chang

Supervised By
Dr. Timothy Frasier

A Thesis Submitted to
Saint Mary's University, Halifax, Nova Scotia
in Partial Fulfillment of the Requirements for
the Degree of Bachelor of Science.

March 17, 2017, Halifax, Nova Scotia

Approved: Dr. Timothy Frasier
Supervisor
Associated Professor of Biology

Approved: Dr. Zhongmin Dong
Reader
Professor of Biology

Date: April 21, 2017

Network Analysis as a Potential Method Detecting Population Structure

by Yu Ting (Kiera) Chang
supervised by Dr. Timothy Frasier

Abstract

Limited gene flow can divide a species into populations, forming population structure. Population structure is important for evolution and conservation so methods for detecting them were developed. With prior assumptions of population structure, Fst can be used to describe the degree of differentiation between each population. Later, *structure* and Discriminant Analysis and Principle Component (DAPC) overcame those prior assumptions and are widely used contemporarily. One approach called network analysis is widely used in physics and social science to study the patterns in complex relationships using similarities and dissimilarities. Population structure represents patterns in complex relationships, so network analysis should be able to detect population structure. I tested the ability of network analysis to detect population structure by comparing it to *Structure* and DAPC. I used simulated data of 4 populations, each containing 200 individuals' genotypes at 15 loci, with migration rates among them varying from 0.001 to 0.1 migrants per generation. I predicted that network analysis would perform better than the other two methods. Contrary to my expectations, network analysis performed poorly overall, and did not detect any population structure correctly at migration rates greater than or equal to 0.01. At migration rates of 0.05 and 0.1, network analysis detected the correct number of populations just 20% of the time, with individual assignment error rates of 47% and 60%. Thus, network analyses do not appear to be a useful alternative to *Structure* and DAPC. However, only one network clustering method was tested, and therefore future studies could test if other such methods improve the performance of network analyses.

17 March, 2017

TABLE OF CONTENTS

Acknowledgement

First, I would like to thank my supervisor, Dr. Tim Frasier, for all the support on my project and thesis. Nothing would happen without him. I would also like to thank the members from Frasier Lab, Brenna McLeod, Christine Konrad, Leah Springate and Alysha Dupuis, for their ongoing assistant on my project. Moreover, I would like to thank my honor seminar class professors, Dr. Anne Dalziel and Dr. Laura Weir, for all their help with my thesis writing. Finally, special thanks to my honor class 2017. Thank you for the ongoing physical and mental support throughout the year.

Introduction

Importance of Population Structure

Species are often subdivided into multiple populations with limited gene flow or migration among them, resulting in genetically differentiated populations. This type of differentiation among populations within the same species is called population structure (Slatkin, 1987). Such population structure is important from both an evolutionary and conservation perspective. From an evolutionary perspective, when a species is subdivided into multiple populations, the environmental conditions in each location will be at least slightly different, and may result in different selection pressures and evolutionary trajectories for each population. This can lead to ecological and genetic differences over time and may ultimately lead to speciation (Schluter, 2001).

Understanding population structure is essential for successful conservation and management. If a species is subdivided into multiple populations, then by definition each population is acting as an independent demographic and evolutionary unit, and therefore require separate management and/or conservation actions. Two possible scenarios can result from a lack of recognition of population structure. First, any conservation or management actions focused on one population may only have limited impacts on the other populations, leading to ineffective conservation of the species (Bowen *et al*., 1993; Bowen *et al*., 2005). Second, if conservation or management actions are focused on the species as a whole, without consideration of the population structure, then these actions will likely mitigate threats to one population, but not

others, and therefore be ineffective and lead to localized extinction (Moritz, 1994; Palsboll *et al.*, 2006).

One example of the first scenario is the loggerhead turtle (*Caretta caretta*) in the northwestern Atlantic (Bowen *et al*., 1993; Bowen *et al*., 2005). In this species, females show a migratory pattern between nesting beaches where they lay their eggs, and foraging grounds in the sea. Females tend to show strong fidelity to their nesting grounds, returning to the same site every 2 to 3 years. A major threat to loggerhead turtles is disturbance and subsequent mortality of eggs at nesting sites (Wallace *et al*., 2011). Population genetic analyses showed strong differentiation of mitochondrial DNA between females at nesting sites in Florida and Georgia, indicating that dispersal rates between nest sites are insufficient for replenishment if one becomes extirpated. This work emphasized that to successfully manage the species as a whole, separate management and conservation for each nesting site is required (Bowen *et al*., 1993; Bowen *et al*., 2005).

One example of the second scenario involves tuataras (genus *Spenodon*) in New Zealand (Daugherty *et al*., 1990). This species was described as regionally threatened. It was thought to have only one population and was previously managed as one large unit. The researchers conducted surveys on 24 islands where tuatara lived to record the allozyme and morphological variation. In their survey, they found that rather than existing as one large population, tuataras were separated into multiple distinct populations. Moreover, ignoring population structure had resulted in the extinction of several populations and possibly unknown subspecies (Daugherty *et al,* 1990).

Therefore, for successful conservation and management, populations need to be assigned to management units based on population structure.

Methods for Detecting Population Structure

The importance of population structure for conservation and evolution has been recognized for decades, and therefore numerous approaches have been developed to detect and quantify such structure (Daugherty *et al*., 1990; Hampton *et al*., 2004). Population structure can be closely related with movement patterns. However, movement patterns cannot be the base of analyzing population structure because of two major confounding factors. First, movement of individuals might not be obvious geographically but can be observed in genetic level (Paetkau *et al*., 1995). Second, studies using observation and movement patterns can only provide estimates of contemporary movements, and do not provide information on long-term trends (Rueness *et al.,* 2003). Therefore, to detect population structure, genetic information can reveal gene flow better than observation of movement patterns. As a result, numerous methods for detecting population structure based on genetic data have been developed.

F_{st}

In the late 20th century, biologists typically used F-statistics (F_{st}) to estimate the degree of divergence between one or more populations. F_{st} is how random gametes correlated within populations relative to the gametes in the species (Wright, 1965). F_{st} is

calculated as the reduction of heterozygosity within populations compare to the total species (Frankham *et al.*, 2002). F_{st} has historically been based on a comparison of observed and expected heterozygosity, and this is therefore based on the assumption of Hardy-Weinberg and linkage equilibrium. The value of F_{st} ranges from 0 to 1, with 0 representing no differentiation and 1 representing complete differentiation. Although widely used, one major drawback in the use of F_{st} is that it requires researchers to pre-define which individuals originated from which population. Then F_{st} is used to quantify the differentiation between these pre-defined populations. In other words, in order to calculate F_{st}, the numbers of populations, and individuals within those populations, must be pre-defined. However, if a researcher is studying population structure, then clearly there are questions about the nature of such structure, and therefore these pre-defined assumptions may be incorrect, leading to incorrect analyses of the data, and subsequent conclusions and implications.

Structure

In the twenty-first century, new methods that do not require prior assumptions of population structure were developed and are widely used today. In 2000, Pritchard *et al.* developed a program called *Structure*, which was the first such method, and which revolutionized the analysis of population structure based on genetic data. When given a simulated set of individuals that can be separated into several populations based on allele frequencies, *Structure* calculates the highest probability of a set range of population numbers (e.g. 1-6). The *Structure* program combines Bayesian formulas and

Markov Chain Monte Carlo (MCMC) to find the pattern with the highest probability. Bayesian formulas is used to calculate the possibilities and MCMC is used to tackle as many possibilities as possible. Since the development of this method, pre-assumptions of population numbers and pre-assigning individuals to populations are no longer required. Based on this huge change, method *Structure* eliminates the possibilities of human errors on prior assumptions and this revolutionized how biologists test for population structure. Despite this revolution, *structure* bases its calculations on the assumptions of Hardy-Weinberg and linkage equilibrium, and therefore may be inappropriate for the analysis of many data sets and/or for detecting structuring among populations that are only slightly differentiated (Frantz *et al*., 2009).

Discriminant Analysis of Principal Components (DAPC)

As *Structure* continues to be refined (Falush *et al*., 2003; Falush *et al*., 2007; Hubisz *et al*., 2009), other methods have also been developed. Another widely used approach that was recently developed in 2010 is Discriminant Analysis of Principle Components (DAPC). DAPC combines two sets of analyses, which are principle component analysis (PCA) and Discriminant Analysis (DA) (Jombart *et al.,* 2010). First, genetic data are converted to binary data, where each column is an allele and each individual contains a 0, 0.5, or 1 for each allele indicating that they contain 0, 1, or two copies of that allele. PCA analysis is then conducted on these transformed genotypes to reduce them to a few principal components that appropriately represent the variation in the data. A k-means clustering analysis is then conducted on the principal components,

which decomposes the total variance in allele frequencies into between-group and within-group components. This analysis is conducted for a range of hypothesized numbers of groups (populations) (e.g., 1-10), and then Bayesian Information Criterion (BIC) are calculated and compared for each analysis to identify which number of groups is the best fit to the data. Discriminant analysis (DA) is then conducted on the genotypes, using the identified grouping information, to identify and quantify how to partition the data in a way that maximizes the between group variation while minimizing the variation between individuals in a group (Jombart et al. 2010). This method represents another major advancement because compared to *Structure,* DAPC is much faster, can work on much larger data sets, and does not require any prior assumptions regarding Hardy-Weinberg or linkage equilibrium. Therefore, under some circumstances it performs better than *structure* (Jombart *et al.*, 2010).

Network Analysis

Structure and DAPC are currently widely used by biologists to detect population structure (Jombart *et al.*, 2010; Wood *et al*., 2011). However, other cluster-detection methods exist that are used across a range of other scientific fields. The most prominent of these is network analyses, which is used across a range of fields, from physics to analyses of social structure (Borgatti *et al.*, 2009; Estrada, 2013). With network analyses individuals or things are represented as "nodes", and the relationships among them are represented as "vertices" connecting them. Often the length of the vertices between two nodes is indicative of the strength of the connection between them. When data are

arranged in this way, it is possible to quickly identify and quantify several key characteristics, such as the presence and identity of clusters, where within-cluster connections are tighter than between-clusters, as well as the presence of key individuals. In this way, network analyses have been used to study many complex patterns, such as Facebook and the internet (Papacharissi, 2009), as well as how information flows through communities of people (Serrat, 2009).

In terms of population structure, it should be possible to create a network of individuals based on their pairwise relatedness values. If population structure exists, then individuals within each population should be more related to each other than they are to individuals in other populations. This should lead to clusters of related individuals, representing populations, which should be readily apparent and quantifiable based on the analysis of such a network.

Dyer & Nason (2004) were first to apply network analyses to genetic data, and they did so before the implementation of DAPC. However, their analyses and approach only used network analysis to see how pre-defined populations are associated and related to one and other, using the population as the unit (rather than the individual) (Garroway *et al*., 2008). Thus, network analyses have not yet been used in attempts to detect population structure and for assigning individuals to those populations.

Objectives

My research will examine the potential for network analyses to detect population structure. I will test the effectiveness of network analyses at detecting

population structure by comparing it with the performance of *Structure* and DAPC, using F_{st} to quantify the degree of differentiation between populations. To achieve this, I will use simulated microsatellite data under different migration rates from 0.001 to 0.1 migrants/generation. I predict that network analyses will perform better than the other methods at detecting number of populations and assigning individuals to correct populations.

Methods

Generation of Population Genetic Data for Method Comparisons

To generate data with known patterns of structure for comparing the performance of each method, data were simulated using the program Easypop (Balloux, 2001). Four populations were simulated, each containing 200 individuals genotyped at 15 microsatellite loci. These populations were organized based on the "continuous stepping stone" model, where individuals could only migrate to the population directly adjacent to them (**Figure 1**). The same migration rate was used across all populations within a simulation, but was allowed to vary across simulations. Four different migration rates were tested: 0.001, 0.005, 0.01, and 0.05 migrants per generation, and 10 iterations were run under each condition. The simulations were run for 10,000 generations to allow the population to reach equilibrium, at which time fifty individuals were sampled from each. These 200 samples (50 from each of the 4 populations) represented the test data for each simulation, used to test the performance of each method. The generated data were analyzed by the three methods: *Structure,* DAPC and

network analysis; and Fst was calculated for all pairwise comparisons to quantify the degree of differentiation.

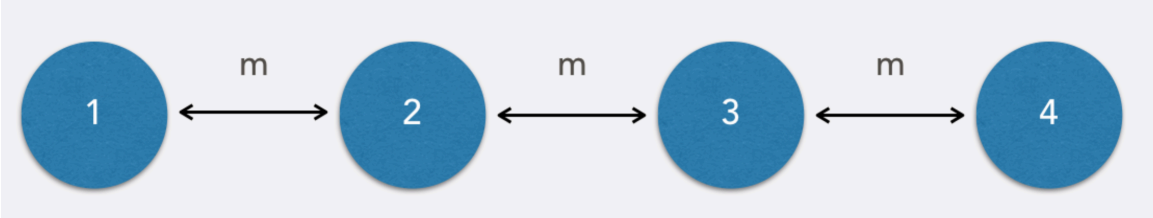

Figure 1. Population model: Continuous stepping stone model used in stimulating genetic data. This model includes four populations exchanging genetic information with only the population beside them. The exchange of genetic information is defined as the migration rate (m).

Comparison of the Different Methods in Different Programs

Running Structure

The generated genetic data was analyzed with the program STRUCTURE to test this method's ability to detect population numbers and assign individuals to correct populations (Falush *et al*., 2003; Falush *et al*., 2007; Hubisz *et al*., 2009; Pritchard *et al*., 2000). The parameters for all the iterations under all migration rates included a burn-in period of 50,000 steps and 500,000 MCMC repeats recorded after burn-in, allowing admixture and a correlation of allele frequencies between populations. Each iteration had the population number (K) set to range from 1 to 6, with 4 iterations at each K. The average probability for each K (across the 4 iterations) was taken as the probability for that K.

DAPC Analysis

DAPC analyses were conducted using the *adegenet* package in R (Jombart & Ahmed, 2011; Rstudio team, 2016). Data were first read into R, and then the number of clusters was estimated using the find.clusters function, setting the maximum number of clusters to 10. This function first converts the data to principal components, and then conducts a *k*-means clustering analysis on the data, using the number of clusters from 1 to the maximum specified. Bayesian information criterion (BIC) are then used to identify which model (number of clusters) best fits the data. From these data, discriminant analyses were conducted using the dapc function. One issue with discriminant analyses is that they can "over-fit" the data if they are based on too many principal components, meaning that they can discriminate between any cluster, not just those that are biologically meaningful. To account for this, we used the a.score function to identify at what number of included principal components the performance of the model started to plateau. The inflection point of such an analysis represents the optimal number of principal components to retain. Subsequent DAPC analyses were based on this identified optimal number of principal components.

Network Analysis

Genetic data were read into R, and relatedness was estimated, using the *related* package (Pew *et al*., 2015; Rstudio team, 2016). The coancestry function was used to estimate pairwise relatedness based on Wang's (2002) estimator. A pairwise matrix of relatedness was created, and then truncated so that all values less than zero were

reported as zero because vertices cannot have a negative length in networks. This matrix was converted to a graph (network) using the graph.adjacency function of the *igraph* package (Csardi & Nepusz, 2006). Cluster analyses were then conducted on the graphs using the walktrap.community function (Pons & Latapy, 2006). The number of clusters identified, and the assignment of individuals to those clusters, were saved for subsequent analyses.

Fst

Simulated data was read into R and package *gstudio* was used for F_{st} (Rodney, 2014; Rstudio team, 2016). The structure parameter from Nei (1978) and genetic.structure function were used for calculating the F_{st} value. Nei's (1978) structure parameter was chosen as it was the original parameter developed for F_{st} calculation. F_{st} values were calculated to act as a reference to quantify the differentiation among populations. The values between each population were saved for further analysis.

Analyzing the result

Out of every ten iterations under each migration rate (0.001, 0.005, 0.01, 0.05 & 0.1), the error rates of the methods (*Structure*, DAPC and network) on detecting the correct number of populations were calculated. For the iterations with the correct number of populations detected, the average error rates for assigning individuals to correct populations were calculated as well. The error rates were then compared separately under each migration rate among methods.

Results

Detecting Number of Populations

All three structure detection methods (DAPC, *structure*, and network analyses) identified the correct number of populations (4) at the lowest simulated migration rate (m = 0.001 migrants/generation). However, at higher migration rates (m = 0.005, 0.01, 0.05 & 0.1) differences in the performance of the methods became evident. Specifically, the network analysis method performed poorly at all other migration rates, even those that were still quite low, with an error rate of 50% at a migration rate of 0.005 (**Figure 2**). The error rates for the other methods (*structure* and DAPC) remained low at migration rates of 0.005 and 0.01, but then greatly increased at a migration rate of 0.05 (**Figure 2**), showing that the critical degree of differentiation for estimator performance occurs between migration rates of 0.01 and 0.05 migrants/generation. The performance of *Structure* and DAPC were similar at migration rates of 0.005 and 0.01, with DAPC performing slightly better at the lower migration rate and *Structure* performing slightly better at the higher rate (**Figure 2**). Interestingly, even in those situations (m= 0.05 & 0.1) where DAPC did not detect 4 populations, it tended to detect 3 populations, corresponding to the merging of 2 populations before population structure vanished. The network method, on the other hand, did not estimate a biologically reasonable number of populations in such case (m = 0.05 & 0.1), with the number of estimated populations ranging from 2 to 8.

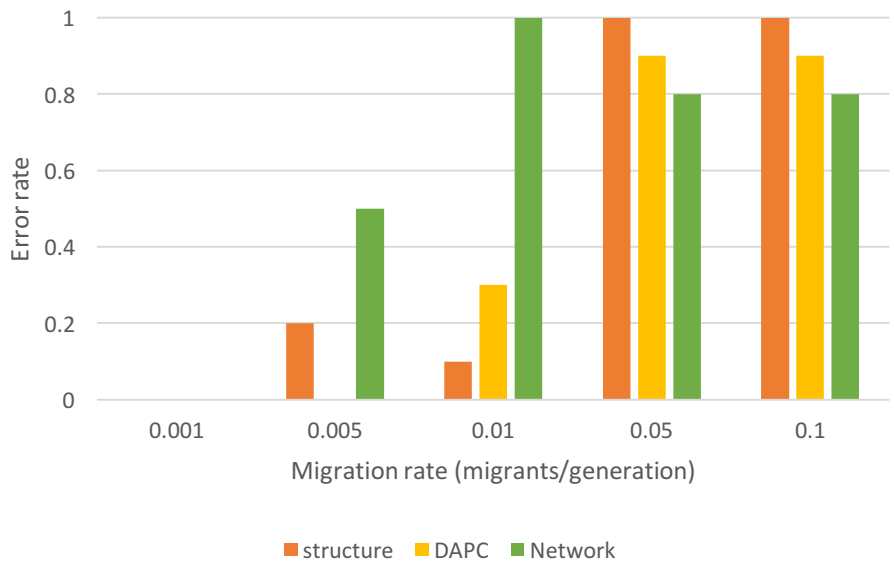

Figure 2. The error rate of each method with respect to detecting the correct number of populations (4) out of the 10 iterations performed under each migration rate.

Population Assignment

Population assignment (the assignment of individuals to the correct population) was only assessed for those iterations where the correct number of populations was detected. Note that this excludes many of the iterations for all methods at higher migration rates (m = 0.05 & 0.1), and for the network analyses even at the lower rates (m = 0.01). All three methods showed a general increase in error rate as migration rates increased. However, at migration rates of 0.05 and 0.1, the one iteration of each in which DAPC detected the correct population number, the error rate was lower than it was at migration rates of 0.005 and 0.01 (**Figure 3**). This contrasts with network analysis, where the error rate in population assignment was very high (0.47 and 0.60) at the higher migration rates even when the correct number of populations was detected

(**Figure 4**). This suggests that the identification of the correct number of populations at these rates is not indicative of the method performing well, but rather may be by chance.

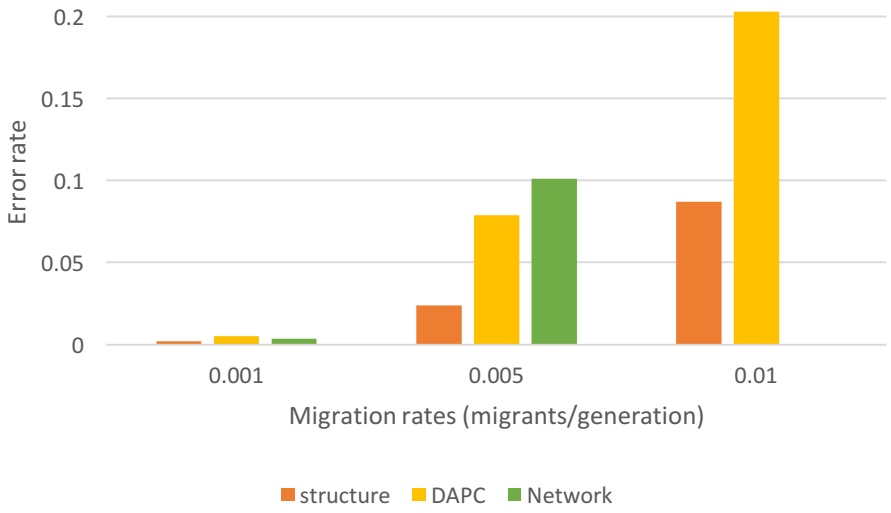

Figure 3. The error rates of assigning individuals to the correct populations for each method when the correct population number (4) was detected at 3 relatively low migration rates (0.001, 0.005 & 0.01 migrants/generation).

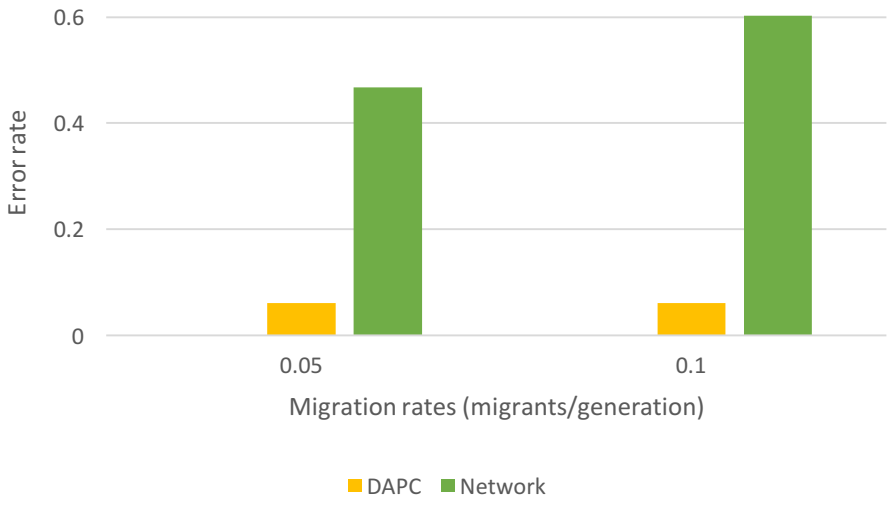

Figure 4. The error rates of assigning individuals to the correct populations for each method when the correct population number (4) was detected at 2 relatively high migration rates (0.05 & 0.1 migrants/generation).

Discussion

As tested, network analysis does not appear to be a powerful tool for detecting population structure nor assigning individuals to their correct populations (**Figure 2**). Although network analysis was expected to work as well as, if not better than, the other methods, with even moderate levels of gene flow its ability to detect population structure and assign individuals correctly decreased dramatically (**Figure 2, 3, 4**). Therefore, network analysis does not seem like a viable alternative for detecting population structure and assigning individuals to correct populations, even when migration rates are relatively low (ie. m >= 0.05).

Limitation for Network Analysis and Future Analysis

There are two clear limitations of this study as it relates to the performance of network analyses. First, due to time limitations I only tested 5 migration rates (m = 0.001, 0.005, 0.01, 0.05 & 0.1 migrants/generation). After the rate of 0.005, tested migration rates differed by large amounts, resulting in unknown performance of all three methods between those migration rate gaps. Therefore, future research could focus on testing more migration rates with smaller intervals, developing a better understanding of the performance of network analysis at detecting population structure. Furthermore, clustering methods in network analysis might play an essential role in detecting correct number of populations. In this project, only one clustering method (walktrap community) was tested. However, other methods exist, such as spinglass, leading eigenvector, fastQ and edge betweenness (Rodriguez & Pepe, 2008; Steinhaeuser & Chawla, 2010).

Therefore, other clustering methods might show different performance other than walktrap community.

At first, walktrap community clustering method was chosen based on the preliminary analysis on how different clustering methods perform at detecting population structure. However, walktrap community did not perform as well as others based on the results. Therefore, due to the worst performance walktrap community showed during this study, additional researches were done. One study has compared the performance of several clustering methods on identifying community structure, and found that the walktrap community detection method tended to find clusters at a relatively higher complexity than the other methods. They ranked the methods on different criteria and in general walktrap had the worst performance in most criteria (Steinhaeuser & Chawla, 2010). Therefore, walktrap community might not be the best clustering method for my simulated simple stepping stone model. Future research could, first, test walktrap clustering method on models with higher complexity and, second, test other clustering methods for clustering the populations in network analysis.

Comparison between *Structure* and DAPC

Although network analysis did not show great performance in detecting population structure, the results allowed us to compare the relative performance of DAPC and *Structure*. Previous studies testing the performance of *Structure* found that it could accurately identify populations when F_{st} is as low as 0.03 (Latch *et al.*, 2006). My results showed a similar pattern. At migration rate of 0.05, F_{st} values between

populations are all lower than 0.01. This is also the migration rate that *Structure* failed to detect any population structure (**Figure 2**).

On the other hand, DPAC performed the best at detecting the correct number of populations at lower migration rates (m = 0.001 & 0.005) as well as maintaining relatively low error rate at migration rate of 0.01. Even at high migration rates (m = 0.05 & 0.1), DAPC tended to detect population numbers that were close to the correct population number (4). A study by Jombart *et al*. (2010) also saw this pattern. However, they found that in stepping stone model, DPAC usually perform better than *Structure* when our study showed similar performance for both methods at low migration rates (m = 0.001 & 0.05) (Figure 2; Jombart *et al.*, 2010).

In addition, at lower migration rates (m = 0.001, 0.005 & 0.01), comparing to DAPC, *Structure* performed the best at assigning individuals to the correct population and this also agreed with a previous study (**Figure 3**; Latch *et al.*, 2006).

Therefore, based on my results, at lower migration rates (m = 0.001, 0.005 & 0.01), a combination of DAPC's number of populations and the *Structure's* assignment should be used to reach the best results. When migration rate reached 0.05, the results from neither method are reliable anymore.

Reference

Balloux F. (2001). Easypop (version 1.7): a computer program for population genetics simulations. *Journal of Heredity*, 92:301-2.

Borgatti SP, Mehra A, Brass DJ & Labianca G. (2009). Network analysis in the social science. *Science,* 323:892-5.

Bowen B, Avise JC, Richardson JI, Meylan AB, Margaritoulis D & Hopkins-Murphu SR. (1993). Population structure of loggerhead turtles (*Caretta caretta*) in the Northwestern Atlantic Ocean and Mediterranean Sea. *Conservation Biology,* 7: 834-844.

Bowen BW, Bass AL, Soares L & Toonen RJ. (2005). Conservation implications of complex population structure: lessons from the loggerhead turtle (*Caretta caretta*). *Molecular Ecology,* 14: 2389-2402.

Csardi G & Nepusz T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Dyer RJ & Nason JD. (2004). Population graphs: the graph theoretic shape of genetic structure. *Molecular Ecology,* 13:1713-27.

Daugherty CH, Cree A, Hay JM & Thompson MB. (1990). Neglected taxonomy and continuing extinctions of tuatara (*Sphenodon*). *Nature,* 347:177-9.

Estrada E. (2013). Graph and network theory in physics. University of Strathclyde, Department of Mathematics and Statistics, Glasgow

Falush D, Stephens M & Pritchard JK. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164:1567-87.

Falush D, Stephens M & Pritchard JK. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes,* 7(4):574-8.

Frankham R, Ballou JD & Briscoe DA. (2002). Introduction to conservation genetics. Cambridge University Press, Cambridge.

Frantz AC, Cellina S, Krier A, Schley L & Burke T. (2009). Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? *Journal of Applied Ecology,* 46:493-505.

Garroway CJ, Bowman J, Carr D & Wilson PJ. (2008). Applications of graph theory to landscape genetics. *Evolutionary Applications,* 620-30.

Hampton JO, Spencer PBS, Alpers D, Twigg LE, Woolnough AP, Doust J, …, Pluske J. (2004). Molecular techniques, wildlife management and the importance of genetic population structure and dispersal: a case study with feral pigs. *Journal of Applied Ecology,* 41:735-43.

Hubisz MJ, Falush D, Stephens M & Prtichard JK. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources,* 9:1322-32.

Jombart T, Devillart S & Balloux F. (2010). Discriminant analysis of principle components: a new method for the analysis of genetically structured populations. *BMC Genetics,* 11:94.

Jombart T & Ahmed I. (2011). adegenet 1.3-1: new tools for the analysis of genome wide SNP data. Bioinformatics.

Latch EK, Dharmarajan G, Glaubitz JC & Rhodes Jr OE. (2006). Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, 7:295-302.

Moritz C. (1994). Defining 'evolutionarily significant units' for conservation. *TREE*, 9(10):373-5.

Nagylaki T. (1998). Fixation indices in subdivided population. *Genetics,* 148:1325-1332.

Nei M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, 76:379-90.

Paetkau D, Calvert W, Stirling I & Strobeck C. (1995). Microsatellite analysis of population structure in Canadian polar bear. *Molecular Ecology*, 4:347-54.

Palsboll PJ, Berube M & Allendorf FD. (2006). Identification of management units using population genetic data. *Ecology and Evolution*, 22(1):11-6.

Papacharissi Z. (2009). The virtual geographies of social networks: a comparative analysis of Facebook, Linkedln and ASmallWorld. *New Media & Society,* 11(1&2):199-220.

Pew J, Wang J, Muir P & Frasier T. (2015). related: related: an R package for analyzing pairwise relatedness data based on codominant molecular markers. R package version 1.0.

Pons P & Latapy M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms Application*, 10(2):191-218.

Pritchard JK, Stephens M & Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetics,* 155:945-59.

Rodney JD. (2014). gstudio: Spatial utility functions from the Dyer laboratory. R package version 1.2.

Rodriguez MA & Pepe A. (2008). On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics,* 2:195-201.

RStudio Team. (2016). RStudio: Integrated Development for R. *RStudio*, Inc., Boston, MA.

Rueness EK, Jorde PE, Hellborg L, Stenseth NC, Ellegren H, Jakobsen KS. (2003). Cryptic population structure in a large, mobile mammalian predator: the Scandinavian lynx. *Molecular Ecology*, 12:2623-33.

Schluter D. (2001). Ecology and the origin of species. *Ecology & Evolution*, 16(7):372-80.

Serrat O. (2009). Social network analysis. *Knowledge Solution,* 1-4.

Slatkin M. (1987). Gene flow and the geographic structure of natural population. *Science,* 236(4803):787-92.

Steinhaeuser K & Chawla NV. (2010). Identifying and evaluating community structure in complex network. *Pattern Recognition Letters*, 31:413-21.

Wallace BP, DiMatteo AD, Bolten AB, Chaloupka MY, Hutchinson BJ, Abreu-Grobois FA, …, Mast RB. (2011). Global conservation priorities for marine turtles. *Plos ONE*, 6(9):1-14.

Wang J. (2002). An estimator of pairwise relatedness using molecular markers. *Genetics,*

160:1203-15.
Wood SA, Frasier TR, McLeod BA, Gilbert JR, White BN, Bowen WD, …, Brault S. (2011). The genetics of recolonization: an analysis of the stock structure of grey seals (*Halichoerus grypus*) in the northwest Atlantic. *Canadian Journal of Zoology*, 89:409-7.
Wright S. (1965). The interpretation of population structure by F-statistics with special regard to system of mating. *Evolution,* 19(3):395-420.