HOW DOES GENOMIC VARIATION EVOLVE IN ASEXUAL POPULATIONS AFTER AN INVASION EVENT?

By

HARRISON ANTHONY

A thesis submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE IN BIOLOGY

WASHINGTON STATE UNIVERSITY
School of Biological Sciences

MAY 2021

To the Faculty of Washington State University:

The members of the Committee appointed to examine the thesis of HARRISON ANTHONY find it satisfactory and recommend that it be accepted.

_____

Mark F. Dybdahl, Ph.D., Chair

_____

Jeremiah Busch, Ph.D.

_____

Omar Cornejo, Ph.D.

HOW DOES GENOMIC VARIATION EVOLVE IN ASEXUAL POPULATIONS AFTER AN INVASION EVENT?

Abstract

by Harrison Anthony, M.S.
Washington State University
May 2021

Chair: Mark F. Dybdahl

Genetic variation is typically associated with the ability to establish a population in a new environment. Most invasion events should be relatively unsuccessful because population bottlenecks reduce genetic variation. This reduction in genetic variation makes new populations susceptible to parasites, disease, and the loss of beneficial variants. Invasive asexual species pose an interesting case as their populations are often presumed to be genetically depauperate and encounter more genetic "challenges". To establish a new population, they must also overcome the accumulation of deleterious mutations and clonal interference. Despite this, many asexual species have become highly invasive. We asked the question; how does genomic variation evolve in asexual populations after invasion? To answer this question, we investigated genomic variation within North American populations of the highly invasive asexual snail, *Potamopygrus antipodarum* (New Zealand Mud Snail). The New Zealand Mud Snail is a parthenogenetic freshwater species that originally invaded Idaho's Snake River in 1987. To determine how their genomes have changed post invasion, we aligned previously sequenced methylated DNA immunoprecipitation reads to a reference transcriptome for two old (river) sites and three new

(lake) sites. We then called single nucleotide polymorphisms (SNPs) to quantify genomic diversity, identify population substructure, and investigate invasion history. To quantify genomic diversity, we calculated population mutation rate (θ) and nucleotide diversity (π). We used hierarchical F statistics to investigate how genetic variation was structured. Lastly, we analyzed allele frequency spectra and calculated Tajima's *D* to investigate invasion history. We found that all populations had a similar amount of nucleotide diversity (.354 -.379) and a similar population mutation rate (.038-.065). Variation was structured at the site level (F_{ST} =.497), and all pairwise F_{ST} comparisons showed genetic differentiation (F_{ST} > .15). Recent bottlenecks were inferred from the allele frequency spectra of all populations. This result was supported by Tajima's D values which were all positive (.941 - 1.40). We found North American populations of *P. antipodarum* to have high levels of genetic variation and were genetically differentiated. We believe this research highlights the importance of testing for genetic diversity in invasive populations before assuming a lack of genetic variation.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER ONE: ANALYSIS OF READS GENERATED FROM POOLED METHYLATED SEQUENCING DATA

Introduction

Pooling the DNA of individuals in a population before sequencing is a cost-effective way to sample many individuals within a population at once and has been shown to be provide accurate estimates of allele frequencies in a population (Futschik & Schlötterer, 2010). Furthermore, pooled sequencing (pool-seq) requires less DNA from each individual and reduces the total work and time needed for sequencing studies (Anand et al., 2016). Despite the many upsides of using pool-seq, there are potential drawbacks.

The most notable drawback to pool-seq is the informational loss of individual genotypes. In turn, this makes any analyses that require estimates of homozygosity and heterozygosity impossible without additional barcoding. To work around this limitation for variant analysis, researchers have turned to using population level allele frequencies in lieu of individual genotypes. However, single nucleotide polymorphism (SNP) based population genetic analyses quickly become complicated. Previous literature has been published to help guide researchers who analyze pool-seq data with SNPs (Anand et al., 2016; Ferretti et al., 2013; Lynch et al., 2014) and there are prebuilt bioinformatic pipelines to identify SNPs and streamline analyses (Kofler et al., 2011; Micheletti & Narum, 2018).

Despite the presence of prebuilt pipelines and guidelines, there are some inconsistencies within these pipelines when there are multiple pools per population or when an atypical sequencing method is used. Our lab has access to a dataset which has both, multiple pools per population, and it has undergone an atypical, methylated sequencing technique. We first aimed to

describe the different ways pooling DNA and methylation could impact SNPs and variant analysis. We then describe the SNP calling pipeline we have developed to use with this type of data. Lastly, we describe ways to account for other errors generated from methylated data using the pooled, methylated data we have access to.

Traditional pooled sequencing pipelines

Generally, all pool-seq pipelines aiming to identify single nucleotide polymorphisms (SNPs) share similar steps. If there are multiple pools comprising each population, then segregating sites are identified after merging their alignment files (F*igure 1*). After finding segregating sites for each population, pipelines will allow the user to enter parameters for filtering options. Typically, this will include several different stages including quality score of the SNP, maximum and minimum read depth, and a minor allele frequency threshold to filter out rare variants (*Figure 1*). After filtering, researchers will then have allele frequencies at each segregating site for each population. Pipelines at this point begin to diverge, and some will calculate population summary statistics for the user (Kofler et al., 2011; Micheletti & Narum, 2018). However, there are some issues still not being addressed particularly with SNP calling and filtering.

Most pool-seq pipelines will use different SNP callers and have different SNP filtering steps. Choice of SNP caller has been thoroughly explored and explained (Huang et al., 2015), but SNP calling has not been explained in the context of having multiple pools per population. When researchers have multiple pools representing each population, it becomes unclear when to combine these pools together. Merging pools before calling SNPs is typically done with other pool-seq pipelines. However, we found that this causes issues with the header of our alignment

file, and the change in the alignment file was generating downstream artifacts which appeared during variant analysis. While this issue could very well be unique to our dataset, it should be clarified within pool-seq pipelines.

Typical MeDIP data

One sequencing method that is not commonly used for population genomics is methylated DNA immunoprecipitation (MeDIP). MeDIP is an affinity enrichment method where DNA fragments that are methylated are non-covalently bound to 5-methylcytosine antibodies (Lam et al., 2005). This technique is typically done to profile genome-wide methylation patterns (Vucic et al., 2009; Staunstrup et al., 2016). While MeDIP has been used to identify epigenetic patterns, there is still a large amount of genomic information being retained.

Studies have demonstrated that some invertebrates have 10-40% of their genome methylated (Tweedie et al., 1997), and some vertebrates have 60-90% of their genome methylated (Head, 2014). Even though MeDIP enrichment results in only capturing a fraction of possible genomic regions, there is still a considerable amount of the genome being retained, depending on the species, which can be used for variant calling.

Another upside of using MeDIP-seq is the use of an immunoprecipitated antibody. The use of an antibody does not require the DNA transformations seen with other 5-methylcystosine targeted sequencing methods. Another technique which also captures methylated regions of the genome, bisulfite sequencing, requires the conversion of unmethylated cytosines to uracil and then to thymine (Lie et al., 2012). MeDIP avoids this and adheres to a more traditional shotgun sequencing technique. Despite the potential upsides of using MeDIP data for genetic analyses, variant calling is currently unexplored.

Potential issues with MeDIP data

Issues can arise with using MeDIP because there will be an inherent methylation bias within the data. One potential issue created by sampling only methylated regions of the genome is finding fragments in some populations but not others due to methylation state. This would cause the fragment to align to a reference in one or more populations where the fragment was methylated. Consequently, there will then be sites identified as polymorphic in some populations but not others even though they likely could be. This scenario could be exacerbated if there are many differentially methylated regions potentially leading to miscalls of SNPs found in only one population.

Another factor that must be taken into consideration when calling SNPs is that MeDIP has a higher enrichment affinity towards hypermethylated reads (Nair et al., 2011). This could affect the genomic data used for variant analysis in two different ways. The first is an overrepresentation of hypermethylated fragments in one or more populations. These hypermethylated regions could artificially inflate the read depth of certain fragments.

The second way is by MeDIP-seq causing an overrepresentation of CpG dense areas. Methylated cytosines are typically found in CpG dinucleotides (Jang et al., 2017) which could potentially result in MeDIP sampling more CpG dense areas than would be sampled with a more common sequencing technique. While it is unclear exactly how sampling more CpG dense sites could affect SNPs being analyzed, there is evidence that guanine to thymine transversions are more prevalent in methylated CpG sequences (Pfeifer, 2006). It has also been shown that methylation could act as a causative force of increased mutation rate and nucleotides nearby CpG sites could also be affected (Kusmartsev et al., 2020).

Solutions

We developed a robust SNP calling pipeline for researchers analyzing pooled DNA data that also includes a method to account for the methylation bias generated from MeDIP enrichment. Our pipeline first finds segregating sites in each pool (*Figure 2*). This is one unique portion of our pipeline that we found prevents artifacts from being generated from a merged alignment file. Our pipeline then filters SNPs based on quality score and read depth before merging pools together. Then our pipeline filters out variants based on minor allele frequency (*Figure 2*).

We have also included a step in our pipeline to account for one of the issues caused by MeDIP enrichment. Sites that have not aligned in all pools are removed (*Figure 2*). This additional filtering stage is not seen in any other SNP calling pipelines and is the most logical way to prevent a false negative (SNP not being found due to methylation state). Performing this step prevents differential methylation from impacting the variants used during downstream analyses.

Examining methylation bias with example data

We used the pipeline presented to call SNPs with a pooled, MeDIP dataset that was aligned to a reference transcriptome published in Wilton et al. (2013). This data was used in a previous study aimed at identifying differentially methylated regions of DNA between North American populations of *Potamopygrus antipodarum* (Thorson et al., 2017). The dataset itself consists of five total populations, and there are three pools for each population. We were

interested in seeing how the potential methylation errors not accounted for in our pipeline could impact variant analysis.

Earlier we posited that methylation bias could impact data by causing an overrepresentation of some mutations or by causing differences in read depth distribution between samples. We expected there could be an excess of guanine to thymine transversions due to CpG dense regions being sampled at a higher frequency. We did not see an overabundance of guanine mutations in the SNP data, and the SNPs found were ~25% for all four nucleotides (*Figure 3*). Expectations for the proportion of mutations could change depending on the CG content of the system being studied. We then investigated how hyper-methylation could affect read depth distribution.

Overrepresentation of hyper-methylated fragments could cause a shift towards a greater average read depth within one or more populations. We found that overall, most populations had low sequencing coverage (*Figure 4*). We also found that each population had a similar distribution of average read depth across all aligned regions (*Figure 4).* Based on this result, it does not appear that there are many hypermethylated regions being sampled as we did not see peaks at greater read depth, and the similarity of distributions demonstrate that it is not likely some populations contain many more hypermethylated fragments than others.

Discussion

There are many factors to take into consideration when handling pooled, MeDIP sequencing data. Variant analysis with this type of data is unexplored, and it is difficult to try to account for all potential methylation errors. However, there is a case to be made that MeDIP is a better option than other methylation-based sequencing techniques. We demonstrated that

potential methylation errors that are not accounted for in our pipeline had little impact on the data which would be retained for variant analysis.

Currently, variant analysis with methylated DNA has only been examined for data generated from bisulfite sequencing based studies (Liu et al., 2012). Researchers have likely shied away from the idea because any genomic data retained will have an inherent methylation bias. Bisulfite data in particular needs nucleotide information restored because of the transformative properties of the bisulfite treatment (unmethylated cytosines are converted to uracil and then to thymine) (Liu et al., 2012). We find MeDIP to have a potential upside for researchers interested in studying differentially methylated regions and variant analysis because it does not alter the nucleotide information of sequenced reads.

We have highlighted different ways MeDIP enrichment could impact variant analysis. There is still more research needed on how an overrepresentation of hypermethylated regions or CpG dense areas could affect variant analysis, but we have shown in one pooled MeDIP dataset that there is similar read depth distribution across populations and nucleotide polymorphisms are at similar proportions. It would be advisable for researchers interested in using MeDIP for variant analysis to investigate these two parameters to get a sense for how the methylation bias could be affecting their data. Lastly, methylated fragments could be predisposed to have more mutations per window than a non-methylated fragment. Given this information, estimates of population mutation rate could be higher than expected due to the sampling of these methylated regions.

Conclusion

Pool-seq has become more readily used since Futschik & Schlötterer (2010) first laid out the statistical framework necessary for pool-seq population genetics analyses. Throughout the

past decade, there have been new pipelines (Kofler et al., 2011; Micheletti & Narum, 2018) and new analyses (Gautier, 2015; Hivert et al., 2018) to improve variant analysis with pool-seq. We have contributed a pipeline which features a method to handle multiple pools comprising a single population. This feature has not been explicitly stated in other pool-seq pipeline, and we discovered that combining pools early in the pipeline can produce downstream artifacts. Our pipeline is also able to handle datasets which have undergone MeDIP enrichment.

MeDIP data is currently not being used for variant analysis, but we have established a pipeline which can handle one potential bias introduced by MeDIP enrichment and provided descriptions and examples on how to examine for other potential biases generated from hypermethylated regions. Further research is needed to elucidate other impacts a methylated sampling scheme can have on variant calling, but with the framework laid out in our pipeline, other researchers could explore population genetics in previously generated MeDIP datasets or include MeDIP in a future study to examine epigenetic and genetic questions.

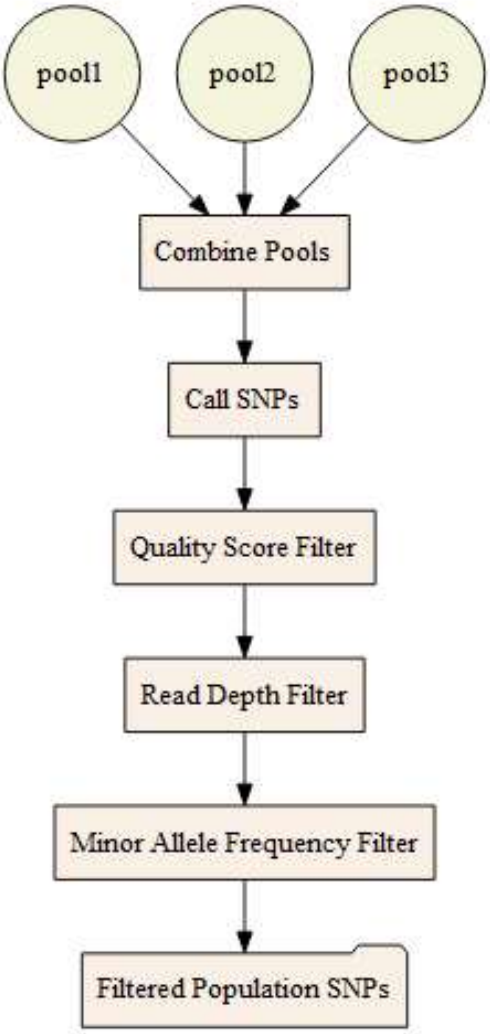

Figure 1 Typical SNP calling pipeline for most pooled-sequencing software.

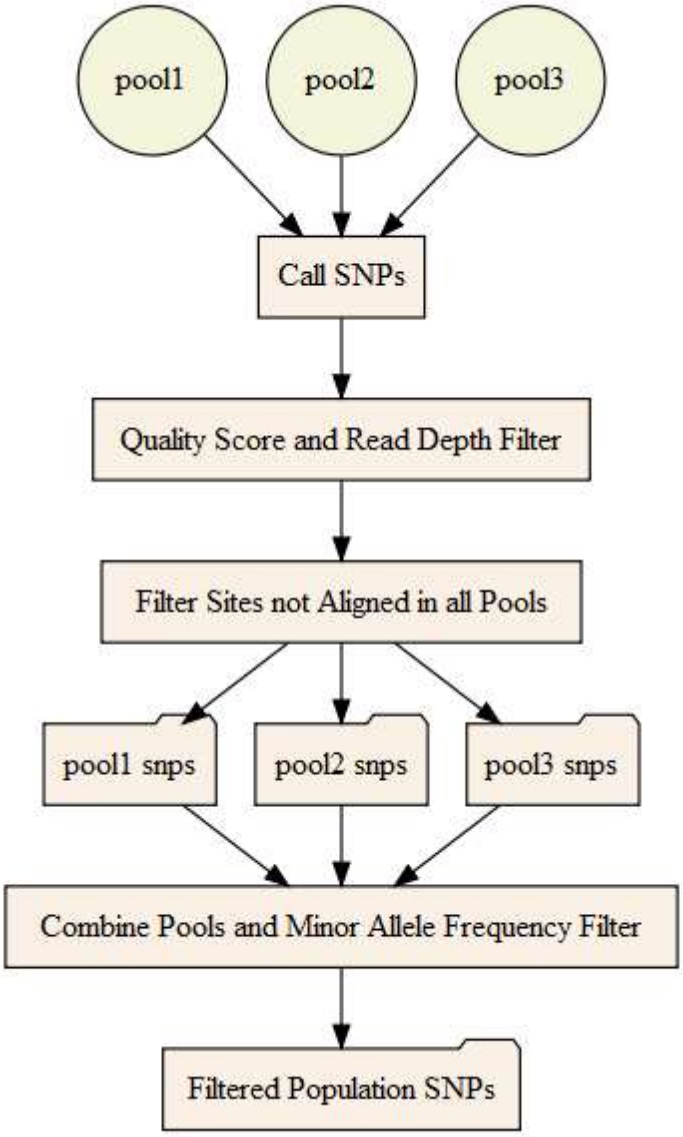

Figure 2 SNP calling pipeline for pooled DNA datasets that also have undergone MeDIP enrichment.

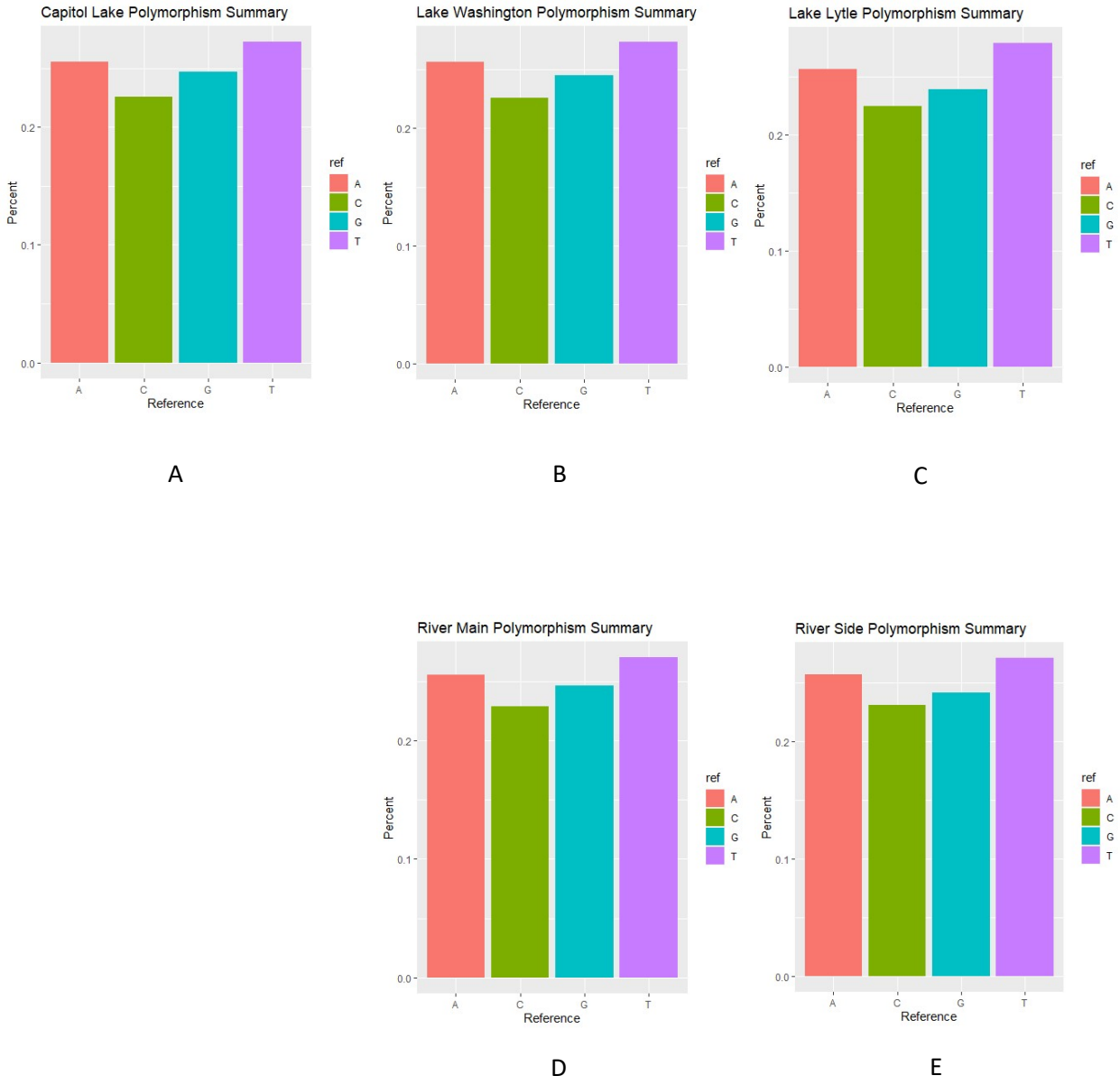

Figure 3 Bar plots of each type of single nucleotide polymorphism type for each example population (A-E).

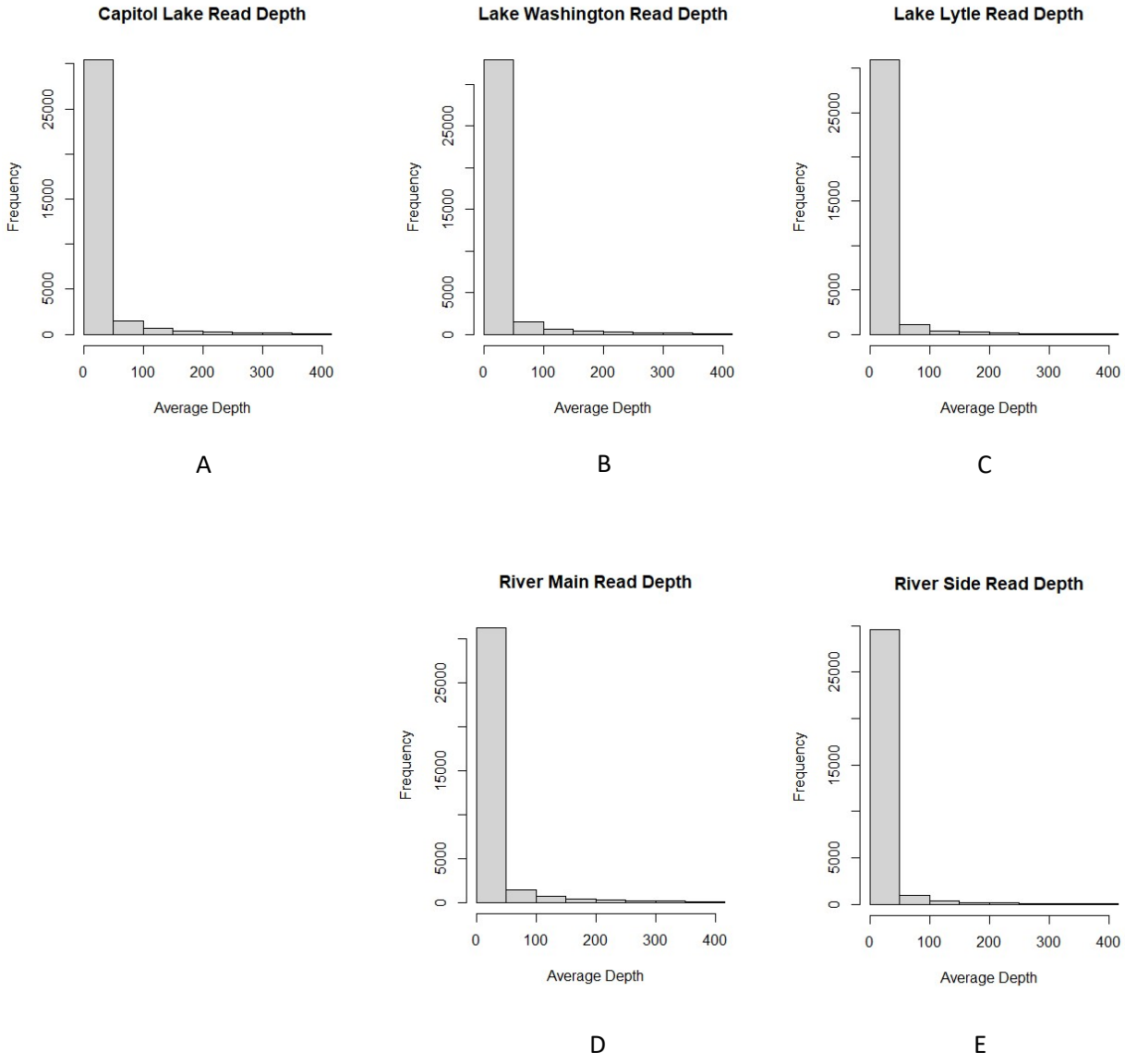

Figure 4 Distribution of average read depth across all mapped contigs for all example populations (A-E).

Literature Cited

Anand, Santosh, Mangano, Eleonora, Barizzone, Nadia, Bordoni, Roberta, Sorosina, Melissa, Clarelli, Ferdinando, Corrado, Lucia, Martinelli Boneschi, Filippo, D'Alfonso, Sandra, & De Bellis, Gianluca. (2016). Next Generation Sequencing of Pooled Samples: Guideline for Variants' Filtering. Scientific Reports, *6*(1), 33735–33735.

Ferretti, Luca, Ramos-Onsins, Sebastián E, & Pérez-Enciso, Miguel. (2013). Population genomics from pool sequencing. *Molecular Ecology*, *22*(22), 5561–5576.

Futschik, Andreas, & Schlötterer, Christian. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics (Austin), 186(1),* 207–218.

Gautier, Mathieu. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. Genetics (Austin). 2015;201(4):1555-1579.

Head, Jessica A. (2014). Patterns of DNA Methylation in Animals. Integrative and Comparative Biology, 54(1), 77–86.

Hivert, Valentin, Leblois, Raphaël, Petit, Eric J, Gautier, Mathieu, Vitalis, Renaud. Measuring Genetic Differentiation from Pool-seq Data. Genetics (Austin). 2018;210(1):315-330.

Huang, Howard W, Mullikin, James C, & Hansen, Nancy F. (2015). Evaluation of variant detection software for pooled next-generation sequence data. BMC Bioinformatics, 16(1), 235–235.

Jang, Hyun Sik, Shin, Woo Jung, Lee, Jeong Eon, & Do, Jeong Tae. (2017). CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function. Genes, *8*(6), 148.

Kofler, Robert, Orozco-terWengel, Pablo, De Maio, Nicola, Pandey, Ram Vinay, Nolte, Viola, Futschik, Andreas, Kosiol, Carolin, & Schlötterer, Christian. (2011). PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PloS One, 6(1), e15925–e15925.

Kusmartsev, Vassili, Drożdż, Magdalena, Schuster-Böckler, Benjamin, & Warnecke, Tobias. (2020). Cytosine Methylation Affects the Mutability of Neighboring Nucleotides in Germline and Soma. Genetics (Austin), 214(4), 809–823.

Lam, Wan L, Haase, Michael, Schübeler, Dirk, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nature genetics. 2005;37(8):853-862.

Liu, Y., Siegmund, K.D., Laird, P.W. *et al.* Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. Genome Biol **13,** R61 (2012).

Lynch, Michael, Bost, Darius, Wilson, Sade, Maruki, Takahiro, & Harrison, Scott. (2014). Population-Genetic Inference from Pooled-Sequencing Data. Genome Biology and Evolution, *6*(5), 1210–1218.

Micheletti, Steven J, & Narum, Shawn R. (2018). Utility of pooled sequencing for association mapping in nonmodel organisms. Molecular Ecology Resources, 18(4), 825–837.

Nair, Shalima S, Coolen, Marcel W, Stirzaker, Clare, Song, Jenny Z, Statham, Aaron L, Strbenac, Dario, Robinson, Mark D, & Clark, Susan J. (2011). Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. Epigenetics, 6(1), 34–44.

Pfeifer, G. P. (2006). Mutagenesis at Methylated CpG Sequences. DNA Methylation: Basic Mechanisms, 301, 259–281.

Staunstrup, Nicklas H, Starnawska, Anna, Nyegaard, Mette, Christiansen, Lene, Nielsen, Anders L, Børglum, Anders, & Mors, Ole. (2016). Genome-wide DNA methylation profiling with MeDIP-seq using archived dried blood spots. Clinical Epigenetics, *8*(1), 81–81.

Strichman-Almashanu, Liora Z, Lee, Richard S, Onyango, Patrick O, Perlman, Elizabeth, Flam, Folke, Frieman, Matthew B, & Feinberg, Andrew P. (2002). A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. Genome Research, *12*(4), 543–554.

S Tweedie, J Charlton, V Clark, & A Bird. (1997). Methylation of genomes and genes at the invertebrate-vertebrate boundary. Molecular and Cellular Biology, 17(3), 1469–1475.

Thorson, Jennifer L M, Smithson, Mark, Beck, Daniel, Sadler-Riggleman, Ingrid, Nilsson, Eric, Dybdahl, Mark, & Skinner, Michael K. (2017). Epigenetics and adaptive phenotypic variation between habitats in an asexual snail. Scientific Reports, 7(1), 14139–11.

Vucic, Emily A, Wilson, Ian M, Campbell, Jennifer M, & Lam, Wan L. (2009). Methylation Analysis by DNA Immunoprecipitation (MeDIP). Microarray Analysis of the Physical Genome, *556*, 141–153.

Wilton, Peter R, Sloan, Daniel B, Logsdon Jr, John M, Doddapaneni, Harshavardhan, & Neiman, Maurine. (2013). Characterization of transcriptomes from sexual and asexual lineages of a New Zealand snail (Potamopyrgus antipodarum). Molecular Ecology Resources, 13(2), 289–294.

CHAPTER TWO: ASEXUAL GENOME EVOLUTION POST INVASION

Introduction

When invasive species colonize and establish themselves in new environments, they typically undergo a population bottleneck. The bottleneck can result in a loss of genetic variation due to genetic drift (Dlugosch and Parker, 2008). The loss of genetic variation makes founding populations more susceptible to disease, inbreeding depression, and the loss of beneficial mutations (Estoup et al., 2016). Despite this, many invasive species expand and grow their range of invasion (Uller & Leimu, 2011). While invasion bottlenecks may lead to issues associated with low levels of genetic variation and fitness, these issues are more acute in asexual species.

Asexual populations face additional challenges during invasion. Colonists might have sampled few multi-locus genotypes from the ancestral range (Barbuti et al., 2012; Grapputo et al., 2005), and they do not sexually recombine genetic material. New variation in asexual populations would then require the accumulation of mutations. The accumulation of new mutations can be slow, and many of these mutations may be deleterious (Jiang et al., 2011). Deleterious mutations accumulate in asexual populations in an irreversible manner and are inherited together, without the ability of sexual recombination to break apart their linkage (Mueller, 1964).

Another side effect of asexual reproduction involves the inability of clones to swap beneficial mutations, such that clones with different beneficial mutations might compete against one another (clonal interference) (Mueller, 1932). This makes selection less effective at fixing beneficial mutations in asexual populations. Less effective selection on beneficial mutations is an issue because selective response leads to adaptation and might contribute to population

persistence (Gonzalez et al., 2013). Despite having to overcome the genetic consequences of a bottleneck event and challenges unique to asexual organisms, some asexual species are highly invasive (Dong et al., 2006; Hall et al., 2003).

We were specifically interested in the dynamics of genomic variation in an asexual invasive snail. The New Zealand Mud Snail (*Potamopyrgus antipodarum*) is a freshwater snail native to New Zealand. In its native range, *P. antipodarum* consists of sexually reproducing diploids and parthenogenetic female triploids (Wallace, 1992). The invasive populations are considered to be exclusively clonal triploids (Hauser et al., 1992; Hughes, 1996; Alonso & Castro-Diez, 2008). Populations in North America in particular have low genotypic diversity (Dybdahl & Drown, 2011). In the western United States, populations are interesting because these snails originally colonized rivers in inland regions (1987), but they have more recently colonized lakes in coastal regions (2009-2011) (Benson et al., 2021). These populations are interesting because they all share a single multi-locus genotype (Dybdahl & Drown 2011) and have been shown to exhibit divergent, adaptive phenotypes (Kistner & Dybdahl, 2013,2014; Thorson et al., 2017). The New Zealand Mud Snail and other asexual invaders must overcome population bottlenecks and other challenges associated with asexuality in order to colonize a new environment and expand their range of invasion. What is unknown is the dynamics of genomic variation in this species, including what restored their adaptive potential after the original bottleneck.

Rapid adaptation generally stems from standing genetic variation (Barrett, 2008; Estoup et al., 2016). Existing beneficial variants are more likely to fix than novel mutations (Barrett 2008; Estoup et al., 2016). While asexuals are typically considered to be genetically depauperate,

there are different ways genomic variation can evolve after invasion, leading to the restoration of adaptive potential in asexual invaders.

If asexuals are also polyploid, then each allelic copy will accumulate its own mutations providing a greater capacity for rapid adaptation (Selmecki et al., 2015). While *de novo* mutations accumulate slowly, they can be common enough to contribute variation for adaptation (Osswoki et al., 2010). They will also accumulate more rapidly when generation times are short (Duglosch et al., 2015) and when growth rates are high (Otto & Whitlock, 1997). It is then possible that novel mutations are likely to have accumulated in *P. antipodarum* because they have short generation times (3-6 months) and they have experienced rapid population growth and expansion during their 35-year invasion history in the western US.

In order to determine how genomic variation has evolved in North American populations of *P. antipodarum* after their original invasion event, we investigated levels of genomic variation within and among five populations. The sampled populations were three lakes and two rivers that contain snails expressing divergent phenotypes and were from the same multi-locus genotype. These populations are comprised of older (rivers) and newer (lakes) sites of invasion. We identified a total of 36,239 SNPs across all five populations and used them to exam the following questions about how their genomes have evolved post invasion.

First, we asked whether there is more genomic variation in older or newer populations. Older populations (rivers) would have more time to accumulate mutations, and newer populations (lakes) would have more recently lost rare variants via the founder effect. On the other hand, if mutations are accumulating rapidly, then the newer populations will have variants that are not present in the older populations.

Second, we assessed how the variation was structured post invasion. Evolutionary forces have likely shaped population substructure throughout the invasion history. Within their short invasion history, demographic events such as migration and bottlenecks would have the largest effect on the underlying substructure. A lack of gene flow between populations and repeated founder effects would create greater structure, but if migrations are frequent and populations are large, then structure will be low.

Lastly, we used our estimates of genetic diversity to investigate alternative scenarios of invasion history. We derived two possible outcomes. A single-origin invasion could have taken place where the population from the original river site colonizes all other sampled populations. We would expect in this case to see all lake populations share many segregating sites with one or more river sites, but they would have more recently experienced the founder effect. Another possible scenario is a stepping-stone invasion from river to a lake and from lake to lake. We would expect in this case for lakes to share more polymorphic sites with one another than with river sites. There would also be an increasing loss of diversity with distance from the original site of invasion. We could also expect that a single origin pattern of invasion would lead to a star phylogeny, while a steppingstone invasion would lead to a simple progression phylogenetic pattern.

Methods

Sample Collection

Potamopygrus antipodarum individuals in this research were previously sampled for another study investigating DNA methylation and shell shape (Thorson et. al, 2017). Snails were collected from regions that were known to exhibit divergent phenotypes and consist of one

single, multi-locus genotype. Sampled regions were also from the oldest recorded invasion site (Snake River, 1987) and more recent invasion sites (3 coastal lakes 2009-2011) (nas.er.usgs.gov). Snails were sampled from three lakes and two rivers. The lake snails were collected from two lakes in Washington and one lake in Oregon. The lakes are Lake 1: Capitol Lake (47.0405° N, 122.9101° W) in Olympia, WA, Lake 2: Lake Washington in Seattle, WA (47.6971°N, 122.2711°W), and Lake 3: Lake Lytle in Rockaway Beach, OR (45.6272°N, 123.9392°W). The rivers are River 1: the main channel of the Snake river (42.7439°N, 114.8416°W) near Wendell, ID, and River 2: a subset of the Snake River at Ritter Island (42.7439°N, 114.8420°W). The samples were kept cool on wet paper towels and transported back to Washington State University in Pullman, WA. A full description of sampling methods and locations is described in Thorson et al. (2017).

Tissue sampling and DNA extraction

Genomic DNA was extracted from thin slices of foot pad tissue and that had been stored in Nanopure™ water and frozen down. Foot pad slices were pooled and DNA was isolated from each of these pools. Pools made from lake individuals required 10 individuals for enough DNA, while river populations required 20 individuals for enough DNA yield. DNA yield was similar across all pools and sampling sites. Complete DNA extraction and tissue sampling protocol is described in a previous study (Thorson et al., 2017). DNA concentration was measured using the Nanodrop (Thermo Fisher) to ensure consistency across pools.

MeDIP (Methylated DNA Immunoprecipitation) and DNA sequencing

Genomic DNA was prepared for sequencing using a Methylated DNA Immunoprecipitation (MeDIP) protocol. Approximately 6 μg of DNA for each pooled sample was used for MeDIP preparation. Magnetic beads (Dynabeads M-280 Sheep anti-Mouse IgG; 11201D) were used in combination with a metal rack to pull down methylated portions of DNA. DNA concentration was measured in Qubit with an ssDNA kit. DNA concentration was measured in Qubit with ssDNA kit. This MeDIP protocol was previously described in Thorson et al 2017.

DNA libraries were prepared from the snail foot pad MeDIP pools using NEBNext® Ultra™ RNA Library Prep Kit for Illumina®. The manufacturer's protocol was followed from step 1.4 and was used to generate double stranded DNA. Each pool was identified by a separate index primer. Next generation sequencing (NGS) was performed at the WSU Spokane Genomic Core laboratory using the Illumina® HiSeq 2500 with a PE50 application. Paired end data was generated and read size was approximately 50 bp in length. Each pool consisted of approximately 100 million reads.

Data transformation and SNP calling

Data was trimmed using TrimGalore (Martin, 2011). The program was run with default parameters, but further analysis of the quality scores required extra flags to be used. The flag –paired was used to specify paired end data allowing two paired end reads to be trimmed at once. TrimGalore then validates paired reads after trimming to determine if one read has become much shorter than the other. After comparing pre-trim and post-trim fastqc (Andrews, 2010) reports, reads were also trimmed from their 5' and 3' end to remove overrepresentation of bases. Adapters were trimmed with a stringency value of 6bp.

The data was aligned to a reference transcriptome (Wilton et al., 2013). We indexed the reference and aligned our reads with Bowtie2 (Langmead & Salzberg, 2012.) The -build flag was used to pre-process the transcriptome and increase query speed when aligning. The only flags used were to specify pair end data (-1, -2) and a SAM file output (-s). The program Samtools (Li et al., 2009) was then used to filter out ambiguously mapped reads. Reads were filtered to have a quality of at least 20 (-q). The reads were converted to a binary format (-bS). The sort function in Samtools was used to sorted reads by their left-most coordinates.

SNPs were called using bcftools (Li, 2011). The mpileup function is used to transform sorted bam files into variant call format files. Bcftools was also used for snp filtering. The results were stored in a VCF format. To prevent methylation bias in the methylated-DNA dataset, only regions that were aligned in all populations were used. The next stage of filtering was done to filter out SNPs with a minimum quality score less than 30. Potential paralogs (>2 alleles per position) were also filtered out. Loci of filtered variants were stored in separate text files and used to filter the previously generated mpileup files. Popoolation2 (Kofler et al., 2011) was used to convert the filtered mpileup into a unique Popoolation2 file (sync) containing allelic counts at each locus. The sync file was imported into R. Inhouse R scripts were then used to filter SNPs to have a minimum read depth of 20, and a maximum read depth of 300. Allelic counts for each pool were combined to get allele frequencies for each population. We then filtered based on a minor allele frequency of .05

Population Genomic Analyses:

Quantifying genomic diversity in old and new populations

Because the first invasion site was recorded approximately 30 years ago, enough mutations could have accumulated to fuel adaptive evolution. To quantify the genomic diversity is present in North American populations of *P. antipodarum*, we calculated nucleotide diversity (π) (Nei & Li, 1979) and population mutation rate (θ) (Watterson, 1975) for each sampled population. This was done to describe genetic diversity for each population and compare levels of diversity between populations. Nucleotide diversity was calculated for every segregating site where k is the number of the major allele and n is the sample size (major and minor allele).

$$\pi = \frac{(2k\ (n-k))}{(n\ (n-1))}$$

The sum of all π values was averaged to provide population level genetic diversity estimates. We expect populations further away from the original site of invasion to have lower values of nucleotide diversity because they would have more recently lost their rare (less diverse) variants via the founder effect.

Population mutation rate was calculated for each aligned contig. K is the number of segregating sites and n is the number of unique haploid samples in the i'th contig. Because the samples were pooled, we estimated the most probabilistic value of n based on each contig's average read depth. The average of all values was taken to provide a population level estimate of θ. We expect older populations to have a higher value of theta than newer populations as they would have had more time to accumulate mutations and would consequently have more segregating sites.

$$\theta = \frac{K}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

Identifying bottlenecks and patterns of invasion history

Populations that have undergone selection or recent bottleneck events will have genomic patterns that deviate from what is expected under mutation-drift equilibrium. This is because populations that deviate from mutation-drift equilibrium will either end up with more segregating sites than haplotypes or more haplotypes than segregating sites. An abundance of rare alleles (more segregating sites than haplotypes) is indicative of a recent population expansion after a bottleneck or diversifying selection. When rare alleles are scarce (greater number of haplotypes than segregating sites), it is likely due to a recent bottleneck or balancing selection. Tajima's D (Tajima, 1989) is a test that highlights the difference between number of haplotypes and number of segregating sites.

$$D = \frac{\pi - \theta}{\sqrt{var(\pi - \theta)}}$$

Tajima's *D* was calculated for each population. To make sure our estimates of π and θ were on the same scale for *D*, we estimated them using the site frequency spectrum as described in Fay & Wu, 2000. Tajima's D was used to see how much each population deviated from mutation-drift equilibrium. We compared values of *D* between all populations. Newer sites would have a positive *D* value possibly indicating a recent population contraction or diversifying selection.

To further investigate invasion history, we investigated the allele frequency spectrum. Bottlenecks are detectable by comparing the different allele frequency classes in the spectrum. Bottlenecks will cause alleles at the lowest frequency (between 0 and .1 allele frequency) to become less abundant than the intermediate allele frequency classes (Luikart et al., 1998).

Finding populations with a larger reduction in the lowest allele frequency class indicates they likely have undergone a more recent bottleneck or a more extreme bottleneck.

We created a neighbor-end joining tree (Saitou & Nei, 1987) using pairwise F_{ST} values as distance metrics. This was done using the unrooted, nj function within the R package, APE (Popescu et al., 2012). If there was a distinguishable pattern of invasion in our sampled populations, we would expect to see a phylogenetic tree to that reflects a single-origin invasion (all populations have small branch length with River Main) or a stepping-stone invasion (small branch lengths with populations that founded one another).

Characterizing population substructure and differentiation

Individuals from sampled populations occupy diverse habitats and exhibit divergent phenotypes. Based on this information we hypothesized that genetic variation could be structured at the site or region (river v lake) level. To determine how genetic variation has been structured post invasion, *F* statistics (Hudson et al., 1992) were calculated. This was done to demonstrate how the diversity present was structured between and within sampled populations.

$$F_{ST} = \frac{\pi_{total} - \pi_{withi}}{\pi_{total}}$$

F statistics were calculated using different values of π. π_{total} represents the average number of pairwise nucleotide differences for loci between all populations. π_{within} represents the average number of pairwise nucleotide differences for loci within a population. By changing which populations contribute to which k and n in π we are able to calculate different *F* statistics (F_{ST}, F_{RT}). If environment and selection influence SNP genetic variation, then we would predict high global averages of F_{RT} alongside high values of regional pairwise F_{ST} comparisons (lakes v

rivers). While populations used in this study are genotypically the same, they express divergent, adaptive phenotypes based on their region.

We also conducted a PCA analysis to further investigate how populations were related. This was done to determine similarities in genetic variation between populations. PCA analysis was done in R using the prcomp function. PCA plots of minor allele frequencies were generated for the first three principal components for all populations. Principal component 1 was plotted against principal component 2, and we also plotted principal component 2 against principal component 3. Clustering of populations and their pools could imply similar levels of genetic variance and population relatedness.

Our last two methods to determine population differentiation were running Kolmogorov-Smirnov tests on the distribution of allele frequencies between two populations and visualizing SNP overlap between two populations. We ran these tests for all pairwise comparisons.

Results:

Is there more genetic diversity in newer or older populations?

All populations had a similar number of SNPs (5,533-8,831) (*Table 1*). Populations did differ in the number of population-specific SNPs. Capitol Lake had the most population-specific SNPs (1,782), and Lake Lytle had the fewest (379) (*Table 1*). All other populations had a similar number of population-specific SNPs (between 907 and 1,166) (*Table 1*). All populations had a similar amount of nucleotide diversity (between .354 and .379) (*Table 1*).

What can our estimates of genetic diversity tell us about invasion history?

Tajima's D values for all populations were all positive (0.94 – 1.40) (*Table 1*). This would imply all populations are evolving outside of mutation drift equilibrium. The positive values indicate populations have either recently undergone a population bottleneck or are experiencing balancing selection.

Our allele frequency spectrum results indicate the positive D values are likely because of bottleneck effects. We found all populations have undergone a recent population bottleneck demonstrated by a reduction in the lowest allele frequency class in the distribution (*Figure 5, Figure 6*). Capitol Lake and Lake Lytle (*Figure 5*) appear to have a similar reduction in the lowest allele frequency class as River Side (*Figure 6*). They likely have undergone a more recent or more extreme population bottleneck than River Main (*Figure 6*) or Lake Washington (*Figure 5*). Lake Washington (*Figure 5*) also has a greater reduction in the lowest allele frequency class than River Main (*Figure 6*).

After running two-sided Kolmogorov-Smirnov tests we found the difference in allele frequency distributions to be non-significant for Lake Washington and Capitol (two-sided Kolmogorov-Smirnov Test, $p = .585$), but all other pairwise comparisons were significantly different (two-sided Kolmogorov-Smirnov Test, $p < 2e^{-16}$).

Our visualization of SNP overlap (*Figure 10*) shows limited overlap between populations (all pairwise comparisons < 51% overlap). River Main and Lake Washington had the highest SNP overlap (50.7%). River Side had the lowest SNP overlap with all other populations (between 33.1% and 33.9%).

How is variation structured among and between populations?

We found a global F_{ST} of .497 and an F_{RT}. value of .502. This implies clear genetic subdivision and a large amount of genetic differentiation between sites. The high F_{ST} value is due to the population-specific SNPs causing a modal shift in the distribution at .8 (*Figure 7*). While excluding population-specific SNPs would help normalize the distribution, it would also remove key differences that are useful for determining population differentiation.

We also investigated pairwise F_{ST} values to determine how genetically different populations are from one another. River side has the most divergent F_{ST} values (.27-.32) (*Table 2, Figure 8*). All other pairwise comparisons had similar F_{ST} values (between .24 and .26). Finding an F_{ST} value between .15 and .25 implies great genetic differentiation (Hartl & Clark, 1997).

After performing PCA we found each population's three pools clustered together (*Figure 11, Figure 12*). This result was mainly seen when plotting PC1 against PC2. Plotting PC2 against PC3 was less structured than plotting PC1 against PC2 and showed less clear clustering of each population's pools. Capitol Lake had each one of its pools spread out further than the other populations (*Figure 10*). This is likely because Capitol Lake had the most SNPs and the most population-specific SNPs.

Table 1: Population Summary Statistics

	Population	# of SNPs	# of population-specific SNPs	pi	theta	Tajima's D
1	Capitol Lake	8831	1782	0.377	0.0653	1.401398
2	Lake Lytle	5803	379	0.377	0.0442	1.367672
3	Lake Washington	7913	1107	0.354	0.0546	1.06783
4	River Side	5533	907	0.379	0.038	1.40207
5	River Main	8159	1166	0.354	0.0506	0.941342

Table 1 Summary statistics for each population**.** Reported π is averaged per site and reported θ is averaged per contig.

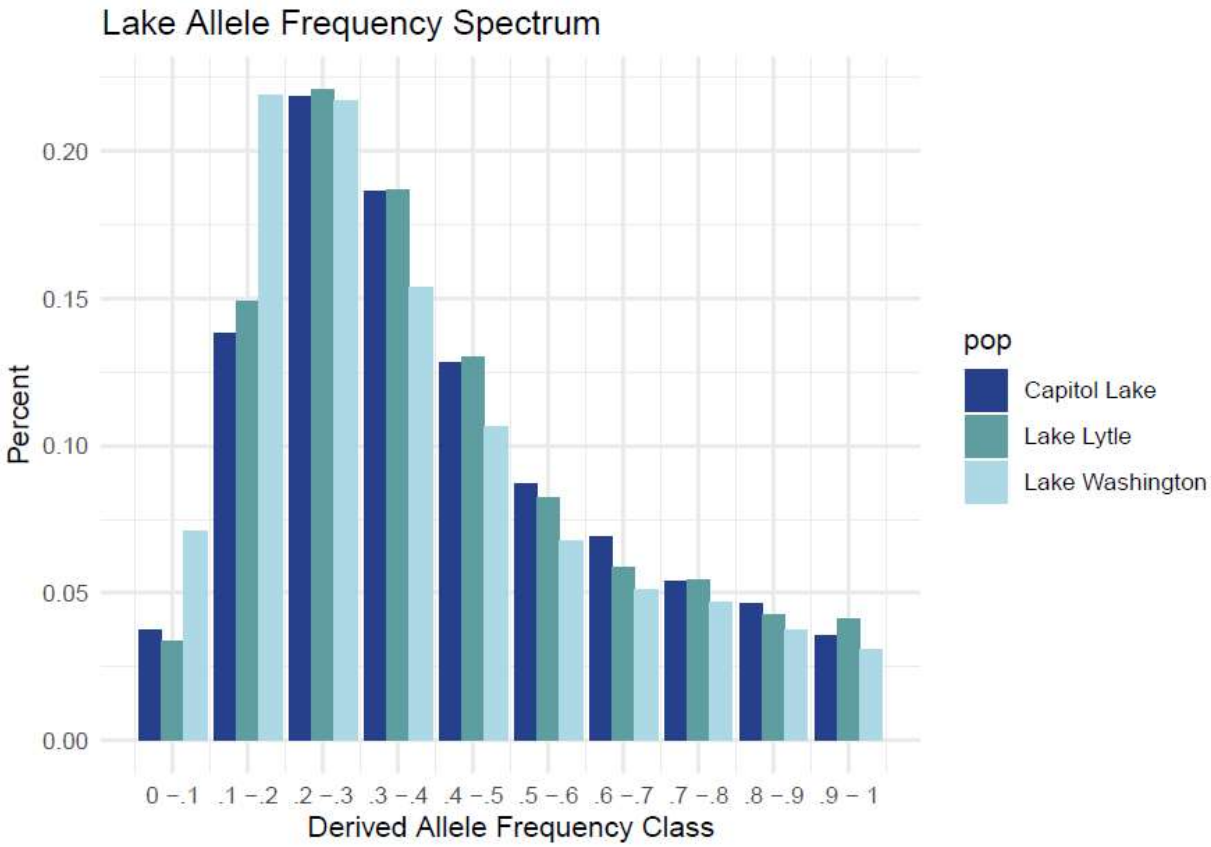

Figure 5 Allele frequency spectrum for all lake populations using the reference allele as the ancestral allele.

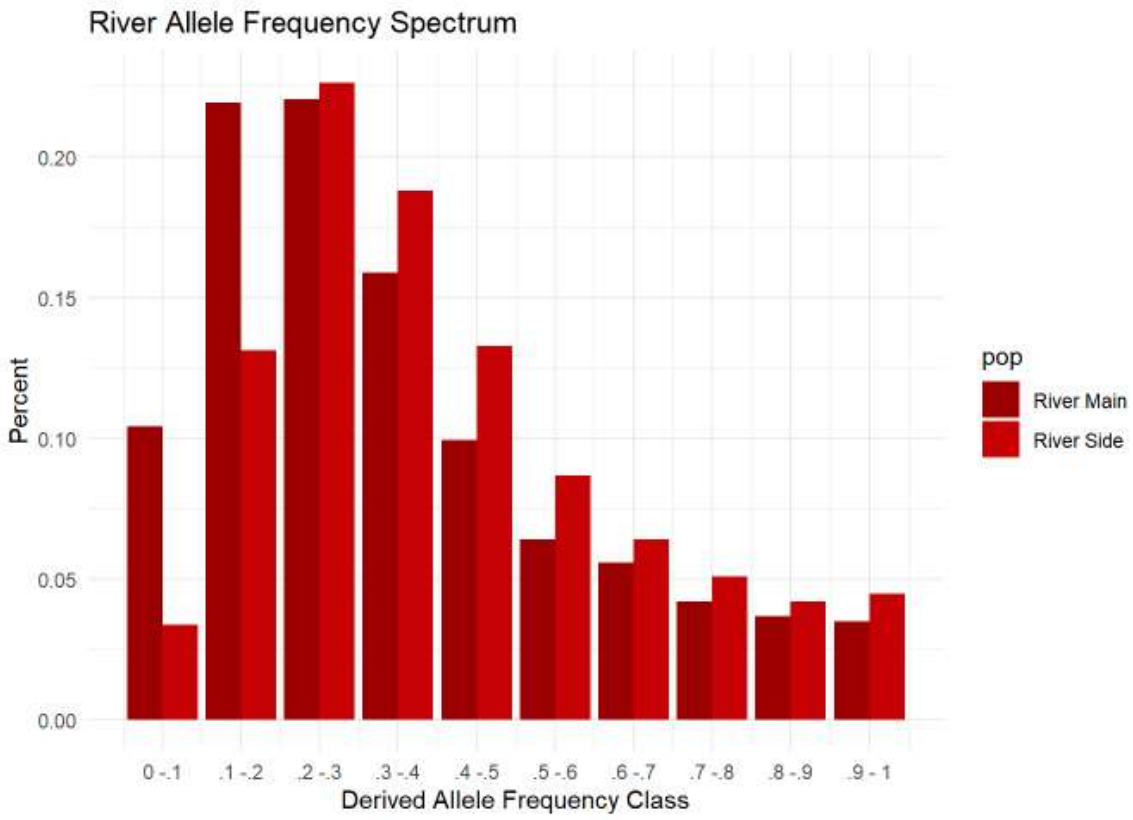

Figure 6 Allele frequency spectrum for all river populations using the reference allele as the ancestral allele.

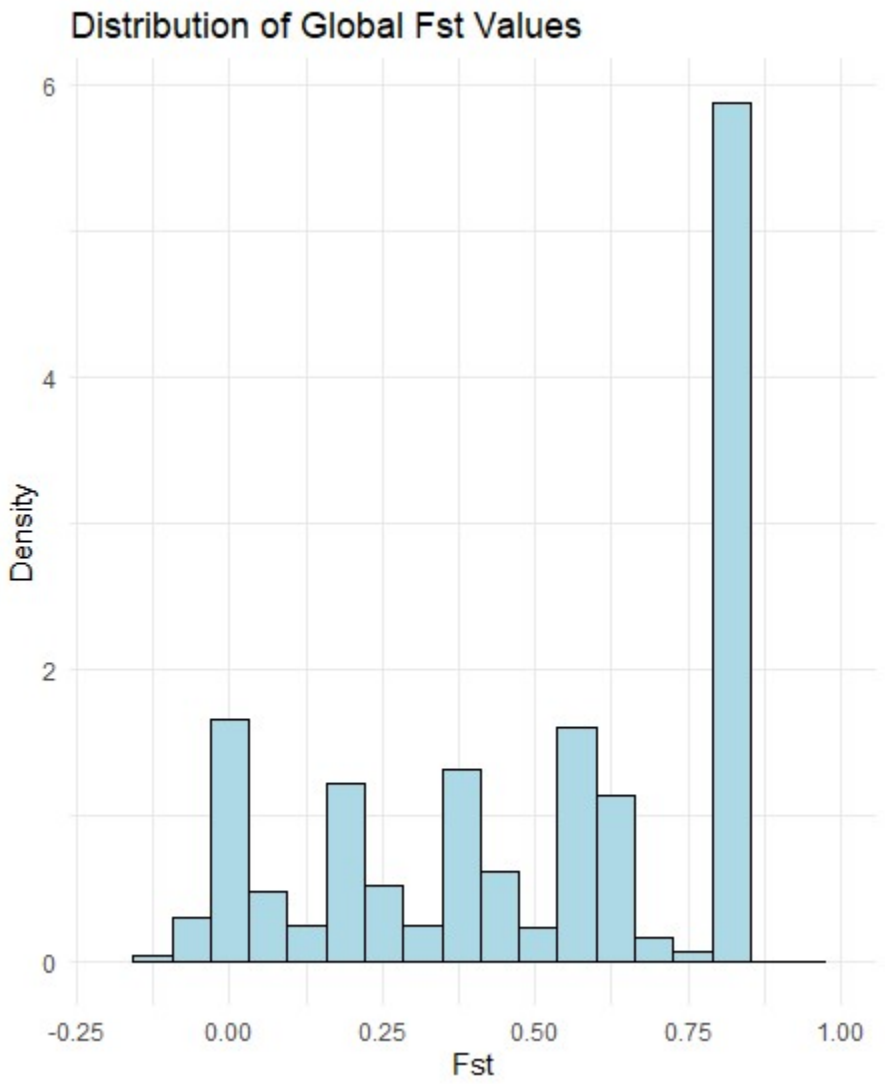

Figure 7 Distribution of global F_{ST} values.

Table 2: Pairwise Fst Table

	Capitol Lake	Lake Lytle	Lake Washington	River Main	River Side
Capitol Lake	0	0.212	0.191	0.186	0.237
Lake Lytle	0.212	0	0.188	0.196	0.205
Lake Washington	0.191	0.188	0	0.176	0.235
River Main	0.186	0.196	0.176	0	0.220
River Side	0.237	0.205	0.235	0.220	0

Table 2 Matrix of all pairwise F_{ST} comparisons.

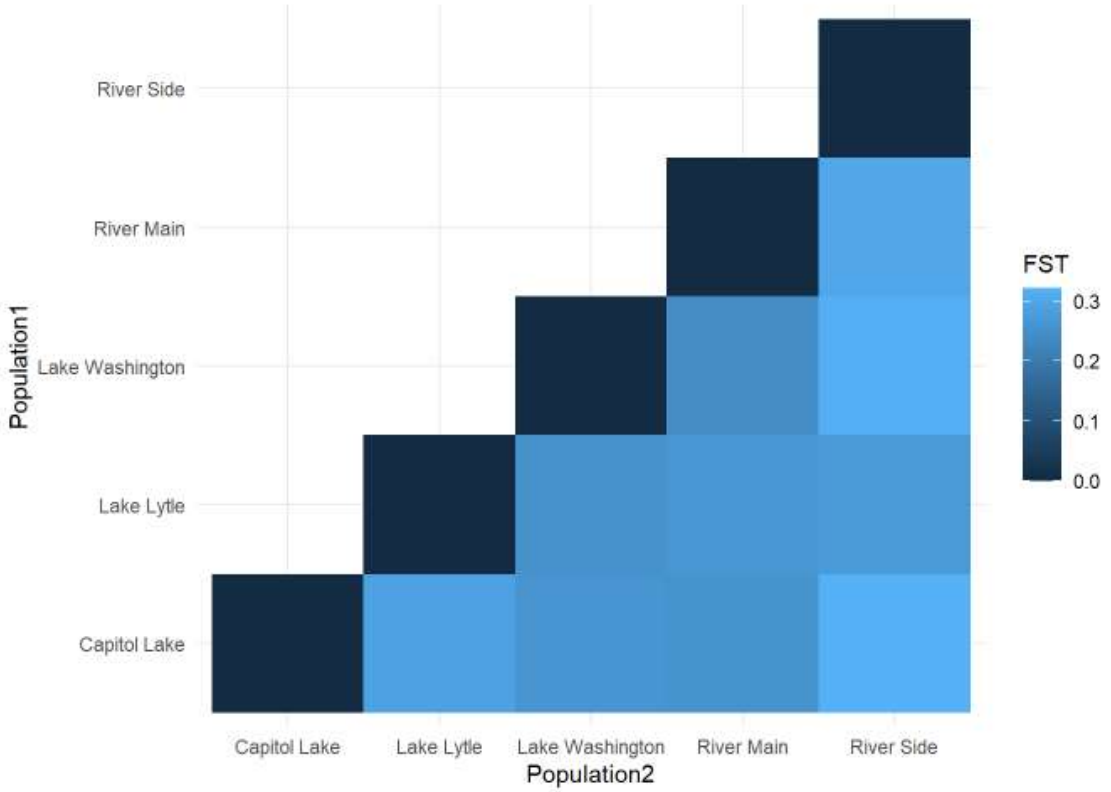

Figure 8 Heatmap of pairwise F_{ST} values.

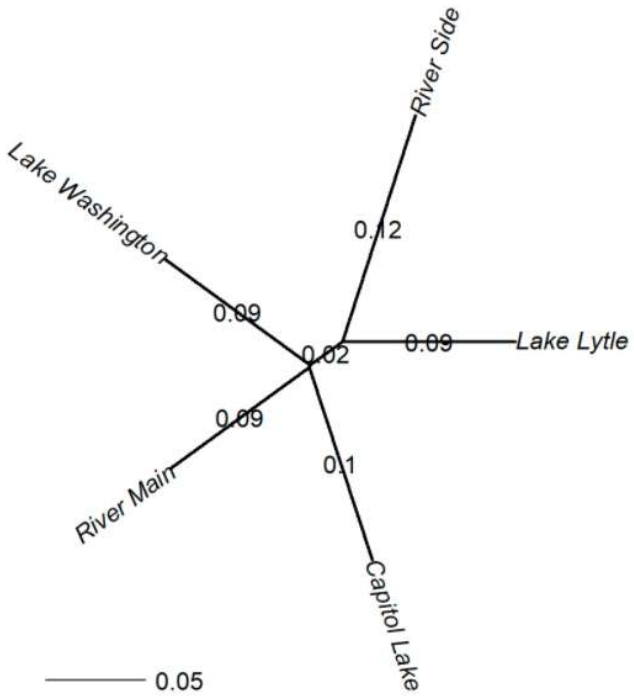

Figure 9 Neighbor-end joining tree created using pairwise F_{ST} values as distance metrics.

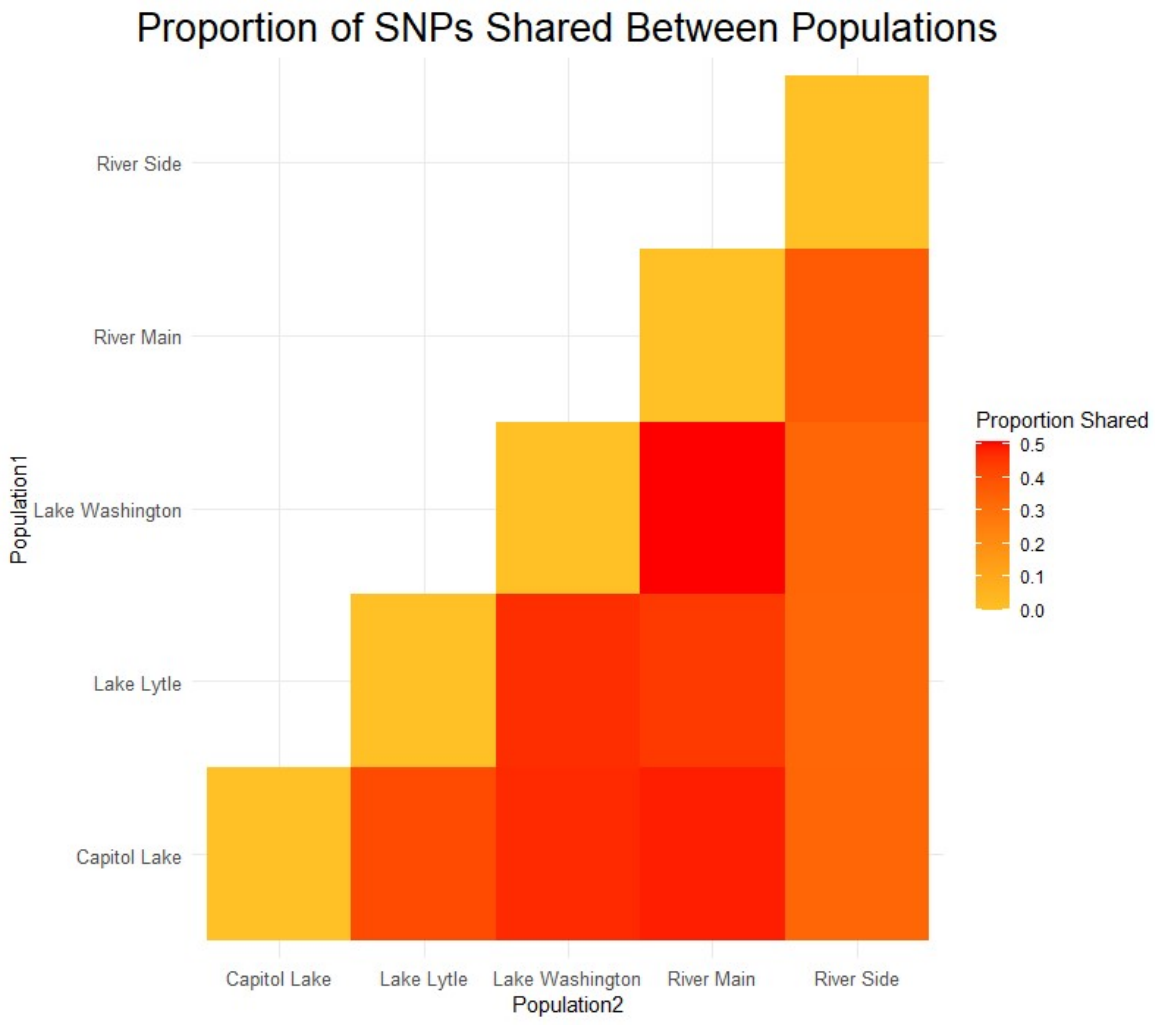

Figure 10 Heatmap showing pairwise SNP sharing between populations. The proportion is the number of loci shared out of the total unique loci found across two populations.

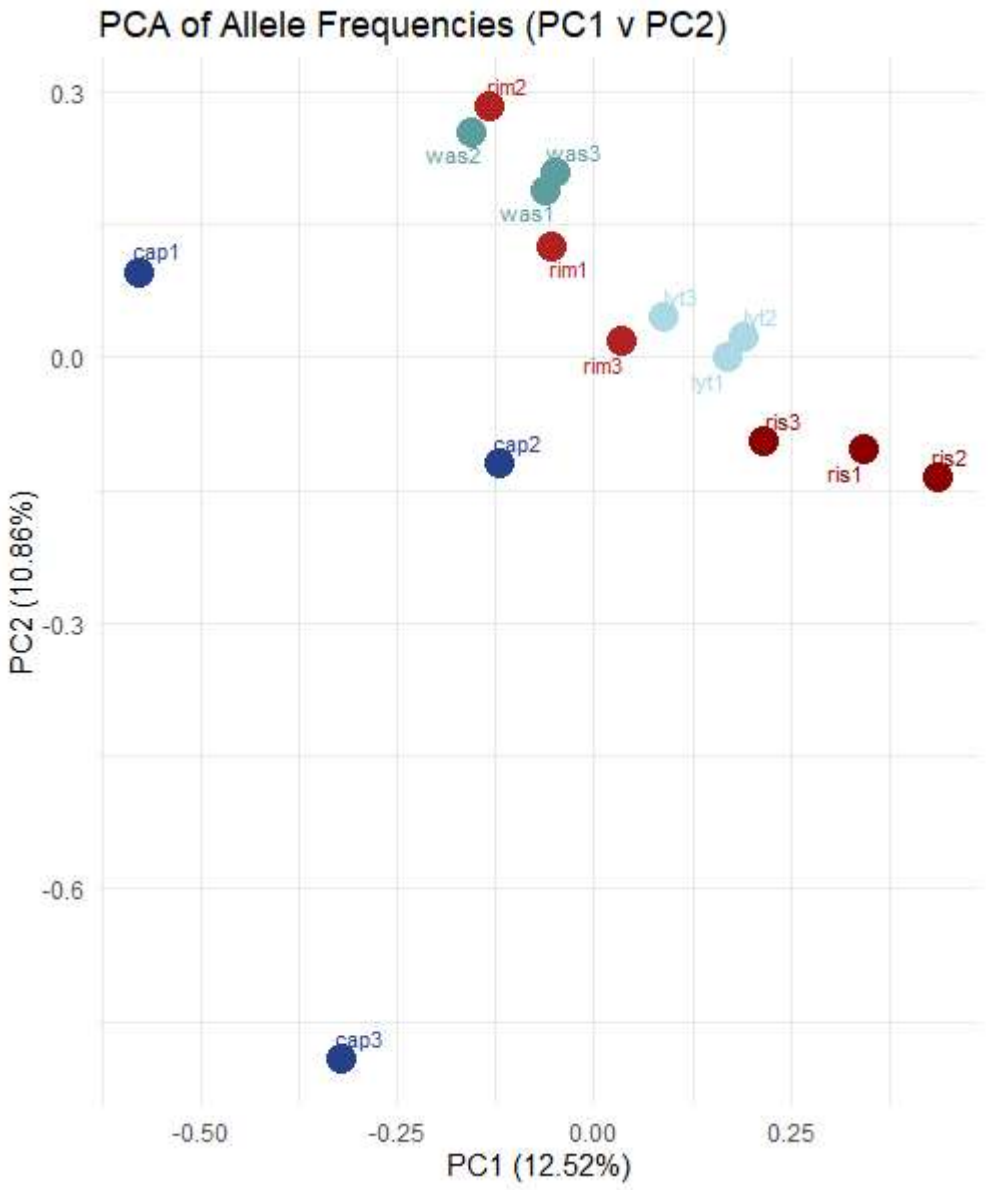

Figure 11 PC1 plotted against PC2 using minor allele frequencies found within each pool in each population.

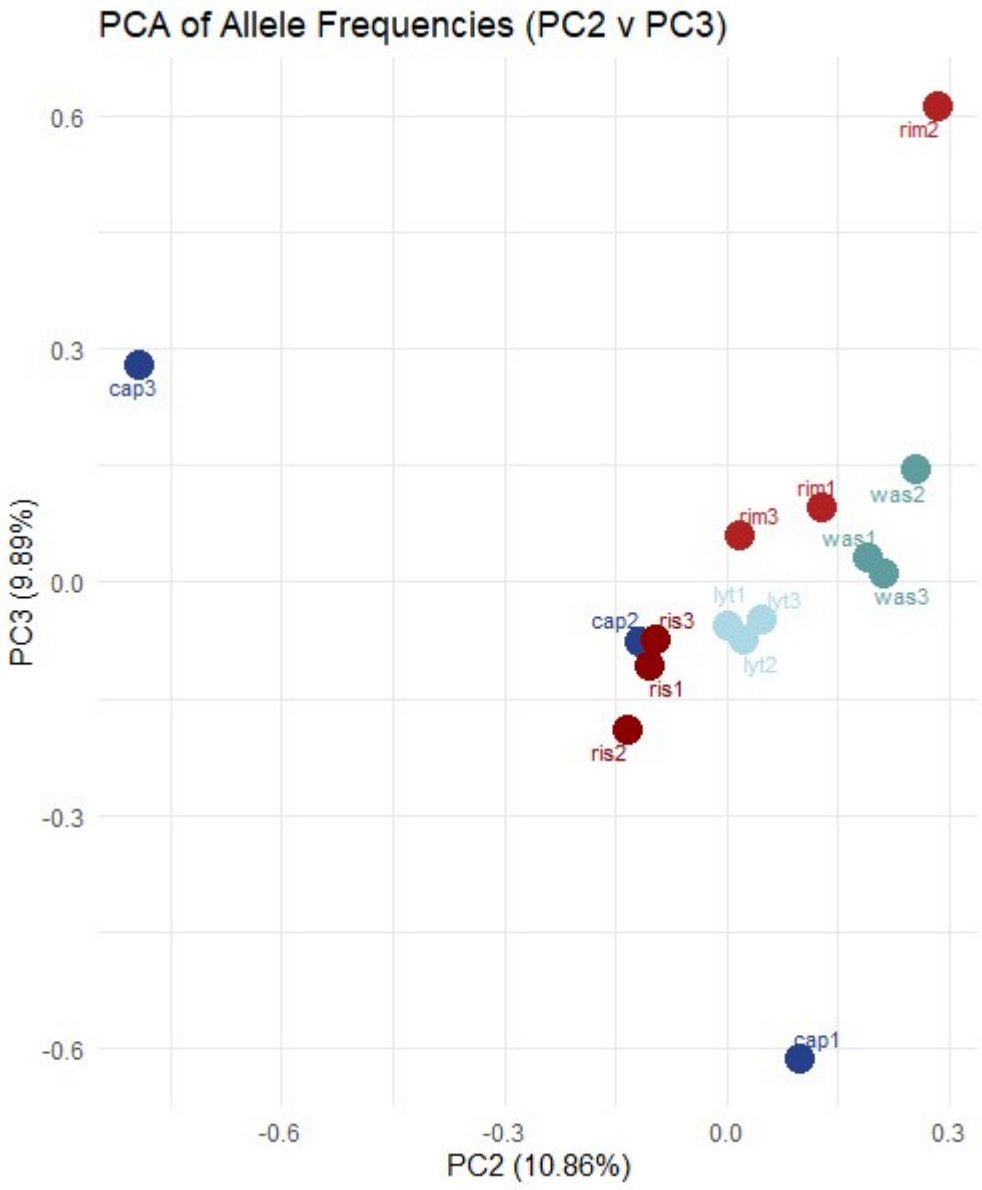

Figure 12 PC2 plotted against PC3 using minor allele frequencies found within each pool in each population.

Discussion

Demographic events and other evolutionary processes determine how much genomic variation is present in invasive populations and how that variation evolves over time. Ultimately, the fate of these new populations is tied to this variation because the restoration of their adaptive potential will come from standing genetic variants that surfed over from their native range or *de novo* variants which have accumulated post invasion. Recent bottlenecks will reduce the number of rare variants found within a population, and gene flow will dictate how many variants are shared as the range of invasion expands.

Our use of segregating sites has led to the discovery that there is high within-locus diversity present in North American populations of *P. antipodarum*, and that there has been a reduction in the number of rare variants in all populations. We also found that the variation present is structured at the site level, implying genetic differentiation. Furthermore, this differentiation appears to be driven by a lack of gene flow between populations indicated by an abundance of population-specific SNPs. Finally, the lack of gene flow, site level structure, and reduction of rare variants demonstrates that out of the two invasion history scenarios, a single-origin invasion is most likely.

Low population mutation rate and high nucleotide diversity

Genomic variation is shaped over time by evolutionary forces. As the range of invasion expands, we would expect founder effects and drift to likely trim SNP variation by removing more rare variants in newer populations. This would lead older sites to be more diverse. Instead we found that nucleotide diversity and population mutation rate were similar for all populations. We did find a large estimate of π and low estimate of θ. The difference between the two is likely

inflated by the original bottleneck. Population mutation rate is likely much smaller than nucleotide diversity due to the original bottleneck reducing the total number of mutations in each population. Nucleotide diversity is likely larger because the original bottleneck has reduced the number of less diverse (rare) segregating sites in all populations. Another recent study has also seen high estimates of genetic diversity in clonal lines of *P. antipodarum* (Million et al., 2020).

Limited gene flow between populations

Evolutionary forces and the size of each diem dictate the size of F_{ST} estimates. How each evolutionary force and population size affects F_{ST} is described by Sewall Wright (1931). One scenario discussed is how F_{ST} will be large if populations are small and migration is infrequent, and F_{ST} will be small when populations are large, and migration is frequent (Wright, 1931). Our F-statistics were large and suggest genetic variation is structured at the site level and populations are genetically differentiated

Even though we do not have information on how large these populations are, our estimate of F_{ST} is likely large due to limited migration between populations. This idea is reinforced by our abundance of population-specific SNPs and limited SNP overlap between populations. We also know that our population-specific SNPs are influencing our estimate of F_{ST}. While it is known how rare variants and choice of estimator can affect F_{ST} (Bhatia et al., 2013), it is somewhat unknown exactly how population-specific SNPs can impact F_{ST}. We have shown that they cause a modal peak in the distribution of F_{ST} values (*Figure 7*). Similarly, another study found a high estimate of F_{ST} when using data that contained population-specific alleles (Savage et al., 2008).

While our F_{ST} values do not shed light on whether the invasion was single origin or stepping-stone scenario, the amount of population substructure indicates there are limited

migration events and consequently limited gene flow between populations. The idea that there is limited gene flow is also indicated by populations having many population-specific SNPs and limited SNP overlap. Large estimates of F_{ST} are seen in other invasive species that have limited dispersal ability (Grapputo et al., 2005; Darling & Reitzel, 2009).

Lack of gene flow drives differentiation

Demographic history can be a driving factor of differentiation between populations. Multiple bottlenecks and migration rate will impact the number of variants lost and the number of variants shared resulting in population-specific SNPs. While population-specific SNPs are understudied (Choudhury et al., 2014), they have been used to understand demographic histories in humans (Lohmueller et al., 2010, Parra et al. 1998).

Two scenarios have been previously outlined (Choudhury et al., 2014) which describe possible origins of population-specific SNPs. One scenario where a SNP might arise *de novo* and stay restricted to that population or another scenario where a standing variant is lost in all but one population (Choudhury et al., 2014). That same study concluded that if at least 80% of common population-specific alleles were found to be the derived allele then they were likely to have arisen after populations diverged from one another (Choudhury et al. 2014). While we have limited information about the ancestral state of each mutation, we are assuming our reference is the ancestral state and the SNPs we have found are the derived allele.

Demographic history (population bottlenecks in particular) has been described in another study as a driving factor of SNPs being retained in one population but lost in others (Harding et al., 2000). We have found evidence of recent population bottlenecks in all our sampled populations, and these demographic events in tandem with *de novo* mutations likely have

resulted in our observed abundance of population-specific sites. Taken together, this information tells us that many standing variants were likely carried over from the very first invasion event back in 1987 and as the invasive range expanded throughout time, populations would very rarely swap migrants or come back into contact with one another.

Invasion History

We were interested in seeing if there was a clear pattern of invasion history in our sampled populations. Based on our results, it is difficult to discern whether the path of invasion was a single-origin scenario or a stepping-stone scenario. The traditional stepping-stone model (Kimura & Weiss, 1964) typically shows a decrease in genetic relatedness with distance from the original site of invasion (a simple-progression pattern). We found even the closest populations (River Main and River Side) to have evidence of genetic differentiation (*Figure 8, Table 2*). Our phylogenetic tree shows what appears to be a starburst pattern (*Figure 9*) and does not provide a clear relation between any two populations. This notion is seen again by limited pairwise SNP sharing between all populations (*Figure 10*).

Out of the two scenarios we have posited we are most likely seeing a single-origin invasion scenario. This would be characterized by an original river site going on to establish all other populations. We would have expected to see all populations have a higher proportion of SNP sharing with one or more river sites. The seemingly equal amount of genetic divergence from one population to another (*Figure 9*) to and each population having similar population summary statistics (*Table 1*) is likely due to demographic history. The repeated founder effects could account for a loss of rare alleles in all populations alongside a reduced estimate of θ. To fully estimate a path of invasion we would likely need to sample more populations consisting of

the same multi-locus genotype and incorporate a stronger simulation-based approach like the ones discussed in Estoup & Guillemaud (2010).

Conclusion

We sought out to learn more about the dynamics of genomic variation within the genotypically depauperate, invasive populations of *P. antipodarum* in North America. We have found evidence that populations of *P. antipodarum* in North America not only have high estimates of nucleotide diversity, but they are also genetically divergent from one another despite belonging to the same multi-locus genotype. Population bottlenecks are typically associated with a reduction in genetic diversity, but we found that even the newer populations have retained much diversity by only losing their rarest sites. These populations act as a good example of an asexual species that has few multi-locus genotypes, but within a single multi-locus genotype, have substantial within locus diversity even after repeated bottleneck events.

Bottlenecks and founder effects are typically thought to reduce a new population's adaptive potential because of the decrease in population size, We found in our study that these repeated bottleneck events did not result in a significant loss of variation or these populations have been accumulating new mutations and not sharing migrants. It is likely that both conditions could be the case and provide a path for newly founded populations of these invasive asexuals to restore their adaptive potential. We believe this research highlights the importance of testing for genetic diversity in invasive populations before assuming a lack of genetic variation.

Literature Cited

Alonso, A, & Castro-Diez, P. (2008). What explains the invading success of the aquatic mud snail Potamopyrgus antipodarum (Hydrobiidae, Mollusca)? Hydrobiologia, 614(1), 107–116.

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].

Barbuti, Roberto, Mautner, Selma, Carnevale, Giorgio, Milazzo, Paolo, Rama, Aureliano, & Sturmbauer, Christian. (2012). Population dynamics with a mixed type of sexual and asexual reproduction in a fluctuating environment. BMC Evolutionary Biology, 12(1), 49–49.

Barrett, Rowan D.H, & Schluter, Dolph. (2008). Adaptation from standing genetic variation. Trends in Ecology & Evolution (Amsterdam), 23(1), 38–44.

Benson, A.J., R.M. Kipp, J. Larson, and A. Fusaro, 2021, Potamopyrgus antipodarum (J.E. Gray, 1853): U.S. Geological Survey, Nonindigenous Aquatic Species Database, Gainesville, FL.

Bhatia, Gaurav, Patterson, Nick, Sankararaman, Sriram, & Price, Alkes L. (2013). Estimating and interpreting FST: the impact of rare variants. Genome Research, 23(9), 1514–1521.

Choudhury, Ananyo, Hazelhurst, Scott, Meintjes, Ayton, Achinike-Oduaran, Ovokeraye, Aron, Shaun, Gamieldien, Junaid, Jalali Sefid Dashti, Mahjoubeh, Mulder, Nicola, Tiffin, Nicki, & Ramsay, Michèle. (2014). Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. BMC Genomics, 15(1), 437–437.

Darling, JA, Kuenzi, A, & Reitzel, AM. (2009). Human-mediated transport determines the non-native distribution of the anemone Nematostella vectensis, a dispersal-limited estuarine invertebrate. Marine Ecology. Progress Series (Halstenbek), 380, 137–146.

Dlugosch, K. M, & Parker, I. M. (2008). Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. Molecular Ecology, 17(1), 431–449.

Dlugosch, Katrina M, Anderson, Samantha R, Braasch, Joseph, Cang, F. Alice, & Gillette, Heather D. (2015). The devil is in the details: genetic variation in introduced populations and its contributions to invasion. Molecular Ecology, 24(9), 2095–2111.

Dong, Mei, Lu, Bao-Rong, Zhang, Hong-Bo, Chen, Jia-Kuan, & Li, Bo. (2006). Role of sexual reproduction in the spread of an invasive clonal plant Solidago canadensis revealed using intersimple sequence repeat markers. Plant Species Biology, 21(1), 13–18.

Dybdahl, Mark F, Drown, Devin M, & Drown, Devin M. (2011). The absence of genotypic diversity in a successful parthenogenetic invader. Biological Invasions, 13(7), 1663–1672.

Estoup, Arnaud, Ravigné, Virginie, Hufbauer, Ruth, Vitalis, Renaud, Gautier, Mathieu, & Facon, Benoit. (2016). Is There a Genetic Paradox of Biological Invasion? Annual Review of Ecology, Evolution, and Systematics, 47(1), 51–72.

Estoup, Arnaud, & Guillemaud, Thomas. (2010). Reconstructing routes of invasion using genetic data: why, how and so what? Molecular Ecology, 19(19), 4113–4130.

Fay, Justin C, & Wu, Chung-I. (2000). Hitchhiking Under Positive Darwinian Selection. Genetics (Austin), 155(3), 1405–1413.

Gonzalez, Andrew, Ronce, Ophélie, Ferriere, Regis, & Hochberg, Michael E. (2013). Evolutionary rescue: an emerging focus at the intersection between ecology and evolution. Philosophical Transactions. Biological Sciences, 368(1610), 20120404–20120404.

Grapputo, Alessandro, Kumpulainen, Tomi, Mappes, Johanna, & Parri, Silja. (2005). Genetic diversity in populations of asexual and sexual bag worm moths (Lepidoptera: Psychidae). BMC Ecology, 5(1), 5–5.

Harding, Rosalind M, Healy, Eugene, Ray, Amanda J, Ellis, Nichola S, Flanagan, Niamh, Todd, Carol, Dixon, Craig, Sajantila, Antti, Jackson, Ian J, Birch-Machin, Mark A, & Rees, Jonathan L. (2000). Evidence for Variable Selective Pressures at MC1R. American Journal of Human Genetics, 66(4), 1351–1361.

Hartl DL, Clark AG (1997) Principles of Population Genetics, 3nd edn. Sinauer Associates, Inc, Sunderland, MA.

Hauser L; Carvalho GR; Hughes RN; Carter RE, 1992. Clonal Structure of the introduced freshwater snail Potamopyrgus antipodarum (Prosobranchia, Hydrobiidae), as revealed by DNA fingerprinting. Proceedings of the Royal Society B-Biological Sciences, 249:19-25.

H.J. Muller. (1932). Some Genetic Aspects of Sex. The American Naturalist, 66(703), 118–138.

H.J. Muller. (1964). The Relation of Recombination to Mutational Advance. Mutation Research, 106, 2.

Hudson, R. R, Slatkin, M, & Maddison, W. P. (1992). Estimation of Levels of Gene Flow From DNA Sequence Data. Genetics (Austin), 132(2), 583–589.

Hughes, R. N. 1996. Evolutionary ecology of parthenogenetic strains of prosobranch snail, Potamopyrgus antipodarum (Gray) 5 P. jenkinsi (Smith). Malac. Rev. Suppl. 6:101–113

Jiang, Xiaoqian, Xu, Zhao, Li, Jingjing, Shi, Youyi, Wu, Wenwu, & Tao, Shiheng. (2011). The influence of deleterious mutations on adaptation in asexual populations. PloS One, 6(11), e27757–e27757.

Kimura, M, & Weiss, G H. (1964). The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. Genetics (Austin), 49(4), 561–576.

Kistner, Erica J, & Dybdahl, Mark F. (2013). Adaptive responses and invasion: the role of plasticity and evolution in snail shell morphology. Ecology and Evolution, 3(2), 424–436.

Kistner, Erica J, & Dybdahl, Mark F. (2014). Parallel variation among populations in the shell morphology between sympatric native and invasive aquatic snails. Biological Invasions, 16(12), 2615–2626.

Kofler, Robert, Pandey, Ram Vinay, & Schlötterer, Christian. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics, 27(24), 3435–3436.

Langmead, Ben, & Salzberg, Steven L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359.

Li, Heng. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27(21), 2987–2993.

Li, Heng, Handsaker, Bob, Wysoker, Alec, Fennell, Tim, Ruan, Jue, Homer, Nils, Marth, Gabor, Abecasis, Goncalo, & Durbin, Richard. (2009). The Sequence Alignment Map format and SAMtools. Bioinformatics, 25(16), 2078–2079.

Lohmueller, Kirk E, Bustamante, Carlos D, & Clark, Andrew G. (2010). The effect of recent admixture on inference of ancient human population history. Genetics (Austin), 185(2), 611–622.

Luikart, G, Allendorf, F W, Cornuet, J M, & Sherwin, W B. (1998). Distortion of allele frequency distributions provides a test for recent population bottlenecks. The Journal of Heredity, 89(3), 238–247.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1), pp. 10-12.

Masatoshi Nei, & Wen-Hsiung Li. (1979). Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases. Proceedings of the National Academy of Sciences - PNAS, 76(10), 5269–5273.

Million KM, Bhattacharya A, Dinges ZM, Montgomery S, Smith E, Lively CM. DNA Content Variation and SNP Diversity Within a Single Population of Asexual Snails. J Hered. 2021 Mar 12;112(1):58-66.

Otto, S. P, & Whitlock, M. C. (1997). The Probability of Fixation in Populations of Changing Size. Genetics (Austin), 146(2), 723–733.

Parra, Esteban J, Marcini, Amy, Akey, Joshua, Martinson, Jeremy, Batzer, Mark A, Cooper, Richard, Forrester, Terrence, Allison, David B, Deka, Ranjan, Ferrell, Robert E, & Shriver, Mark D. (1998). Estimating African American Admixture Proportions by Use of Population-Specific Alleles. American Journal of Human Genetics, 63(6), 1839–1851.

Popescu, Andrei-Alin, Huber, Katharina T, & Paradis, Emmanuel. (2012). ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. Bioinformatics (Oxford, England), 28(11), 1536–1537.

Robert O. Hall, Jennifer L. Tank, & Mark F. Dybdahl. (2003). Exotic Snails Dominate Nitrogen and Carbon Cycling in a Highly Productive Stream. Frontiers in Ecology and the Environment, 1(8), 407–411.

Saitou, N, & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4(4), 406–425.

Selmecki, Anna M, Maruvka, Yosef E, Richmond, Phillip A, Guillet, Marie, Shoresh, Noam, Sorenson, Amber L, De, Subhajyoti, Kishony, Roy, Michor, Franziska, Dowell, Robin, & Pellman, David. (2015). Polyploidy can drive rapid adaptation in yeast. Nature (London), 519(7543), 349–352.

Stephan Ossowski, Korbinian Schneeberger, José Ignacio Lucas-Lledó, Norman Warthmann, Richard M. Clark, Ruth G. Shaw, Detlef Weigel, & Michael Lynch. (2010). The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis thaliana. Science (American Association for the Advancement of Science), 327(5961), 92–94.

Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics (Austin), 123(3), 585–595.

Thorson, Jennifer L M, Smithson, Mark, Beck, Daniel, Sadler-Riggleman, Ingrid, Nilsson, Eric, Dybdahl, Mark, & Skinner, Michael K. (2017). Epigenetics and adaptive phenotypic variation between habitats in an asexual snail. Scientific Reports, 7(1), 14139–11.

Uller, Tobias, & Leimu, Roosa. (2011). Founder events predict changes in genetic diversity during human-mediated range expansions. Global Change Biology, 17(11), 3478–3485.

Wallace, C. (1992). Parthenogenesis, sex and chromosomes in Potamopyrgus. Journal of Molluscan Studies, 58(2), 93–107.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theoretical Population Biology, 7(2), 256–276.

Wilton, Peter R, Sloan, Daniel B, Logsdon Jr, John M, Doddapaneni, Harshavardhan, & Neiman, Maurine. (2013). Characterization of transcriptomes from sexual and asexual lineages of

a New Zealand snail (Potamopyrgus antipodarum). Molecular Ecology Resources, 13(2), 289–294.