Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

3-11-2022 3:00 PM

Effects of spatial and temporal heterogeneity on the genetic diversity of the alpine butterfly Parnassius smintheus

Mel Lucas, *The University of Western Ontario*

Supervisor: Keyghobadi, Nusha, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Biology

© Mel Lucas 2022

Follow this and additional works at: https://ir.lib.uwo.ca/etd

 Part of the Biology Commons

Recommended Citation

Lucas, Mel, "Effects of spatial and temporal heterogeneity on the genetic diversity of the alpine butterfly Parnassius smintheus" (2022). *Electronic Thesis and Dissertation Repository*. 8456.
https://ir.lib.uwo.ca/etd/8456

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

Abstract

Genetic diversity represents a population's evolutionary potential, as well as its demographic and evolutionary history. Advances in DNA sequencing have allowed the development of new and potentially powerful methods to quantify this diversity. However, when using these methods best practices for sampling populations and analyzing data are still being developed. Furthermore, while effects of the landscape on spatial patterns of genetic variation have received considerable attention, we have a poorer understanding of how genetic diversity changes as a result of temporal variation in environmental and demographic variables. Here, I take advantage of advances in DNA sequencing to investigate genetic diversity at single nucleotide polymorphisms (SNPs) across space and time in a model system of the butterfly, *Parnassius smintheus*.

I used double digest restriction site associated DNA sequencing to genotype SNPs in *P. smintheus* from populations in Alberta, Canada. To develop recommendations for analyzing data, I tested the effect of varying the maximum amount of missing data (and therefore the number of SNPs) on common population genetic analyses. Most analyses were robust to varying amounts of missing data, except for population assignment tests where larger datasets (with more missing data) revealed higher-resolution population structure. I also examined the effect of sample size on the same set of analyses, finding that some (e.g., estimation of genetic differentiation) required as few as five individuals per population, while others (e.g., population assignment) required at least 15.

I used the SNP dataset to investigate factors shaping patterns of genetic diversity at different spatial scales and across time. At a larger spatial scale but a single time point, both weather (snow depth and mean minimum temperatures) and land cover (the distance between meadow patches) predicted genetic diversity and differentiation. At a smaller spatial but longer temporal scale, I used a smaller SNP dataset to show that genetic diversity is lost over repeated demographic bottlenecks driven by winter weather, and subsequently recovered through gene flow. My work contributes to understanding how genetic diversity is shaped in natural populations, and points to the importance of both land cover and weather (and specifically, variability in weather) to this process.

Keywords

Bottleneck, Butterfly, Climate, ddRADseq, Genetic differentiation, Genetic diversity, *Parnassius smintheus*, Single nucleotide polymorphism, Weather

Summary for Lay Audience

Genetic diversity describes differences in DNA sequence among individuals of the same population or even between species. In principle, it is this diversity that allows populations to adapt over time when their environment changes. Understanding what factors influence the diversity of natural populations is a central question for both evolutionary biology and for conservation biology (where preserving genetic diversity of endangered populations is a key goal).

I used a new way to measure DNA sequence differences among individuals to assess genetic diversity and to ask what ecological factors are important for maintaining that diversity in the Rocky Mountain Apollo butterfly (*Parnassius smintheus*). This method allows thousands of genetic differences to be identified across an individual's genome. Because this method is new and there are few established guidelines relative to older methods, I tested different ways to process my data. I looked at how to choose which genetic differences to include, as well as how many individuals to sample from each population to get accurate results. I found that compared to some older methods of measuring genetic differences, I was able to sample fewer individuals per population.

I then used the DNA sequence differences I had identified to look at what environmental factors affect genetic diversity in populations from the Rocky Mountains of western Alberta. I found that populations that experience less snow and more extreme winter temperatures have lower genetic diversity. This occurs because these conditions can lead to dramatic reductions in population size, which in turn reduce genetic diversity. Populations surrounded by more forest, as opposed to meadows, also have lower genetic diversity and are more genetically different from other populations. This is because forest limits how easily the butterflies can move among populations.

My work provides real-life evidence of how weather and climate, the physical landscape, and changes in population size are expected to affect a population's genetic diversity. Since climate change will lead to both increased weather extremes and increased forest cover in mountain landscapes, it is likely to result in losses of genetic diversity from populations of the Rocky Mountain Apollo.

Co-Authorship Statement

Chapters 2 and 3 will be published with Gordana Rašić and Ary Hoffmann as co-authors. Dr. Hoffmann provided lab space and lab resources for the development of the double digest restriction site associated DNA library. Dr. Rašić provided supervision and instruction on lab techniques and bioinformatic protocols, and provided the adapter barcoding scheme. I contributed to the study design, prepared ddRADseq libraries for sequencing, analyzed the data, and wrote the chapters.

Chapters 4 and 5 will be published with Jens Roland as a co-author. Chapters 5 will be additionally published with Stephen Matter as a co-author. Dr. Matter and Dr. Roland conducted and supervised the collection of tissue samples.

Chapter 5 will be co-authored with Andrew Chaulk, a PhD student also under the supervision of Dr. Keyghobadi. Andrew and I co-developed the SNP panel I used in Chapter 5.

All chapters will be published with Nusha Keyghobadi as a co-author. Dr. Keyghobadi was involved in developing research questions and study design, assisted in analyzing data, and provided editorial feedback on manuscripts.

Acknowledgments

I would like to begin by expressing my endless gratitude to my supervisor, Dr. Nusha Keyghobadi. Her guidance and expertise are the backbone of my project, and her patience and insight are the reasons I have come this far.

I would like to thank the past and present members of the Keyghobadi lab for their friendship and support: Andrew Chaulk, Benoit Talbot, Helen Chen, Kevin Park, Maryam Jangjoo, and Shayla Kroeze. Andrew has my thanks and my commiseration for our days of shared struggles with lab work and coding. I would like to thank Helen for her years of friendship; my life has been better for having her in it.

I am grateful to the members of my advisory committee, Dr. Graham Thompson and Dr. Kathleen Hill, for their advice and insight on my project. I would like to especially thank Dr. Thompson for his valuable feedback on the final draft of this thesis.

I would like to thank Dr. Gordana Rašić for her mentorship in learning laboratory techniques, and for being a role model in how to mentor my younger colleagues. I also thank Dr. Stephen Matter for his supervision while conducting field work, and for his work along with Dr. Jens Roland to provide many of the samples I used in my work. Likewise, I thank the Biogeoscience Institute of the University of Calgary and the many field work teams that worked there to conduct a long-term mark-recapture study and collect tissue samples used in my work.

I would like to thank my family – my mother Susan, my father David, and my brother Jason – for their unconditional love and support, and their continual interest in my work (although I will clarify one last time for my brother – my work does not involve manipulating dinosaur DNA).

And to all of my friends and family – thank you for your acceptance and your support, as I begin to live as the person I know myself to be. And to Nusha – I am grateful for the acceptance I received from everyone in my life, but you were the first to give me your congratulations. It meant the world to me.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Symbols and Abbreviations

AIC_C	Corrected Akaike information criterion
AFLP	Amplified fragment length polymorphism
bp	Base pair
ddRADseq	Double digest restriction site associated DNA sequencing
H_E	Expected heterozygosity
F_{ST}	Fixation index
HWE	Hardy-Weinberg equilibrium
IBD	Isolation by distance
ML	Maximum likelihood
MLPE	Maximum likelihood population effects
NGS	Next-generation sequencing
H_O	Observed heterozygosity
PDO	Pacific Decadal Oscillation
PGI	Phosphoglucose isomerase
PCR	Polymerase chain reaction
RAPD	Random amplified polymorphic DNA
REML	Restricted maximum likelihood
RFLP	Restriction fragment length polymorphism
RADseq	Restriction site associated DNA sequencing
SNP	Single nucleotide polymorphism
SPRI	Solid-phase reversible immobilization

Chapter 1

1 General introduction

1.1 Genetic diversity and population genetics

Genetic diversity – the total heritable variation within a population – represents the evolutionary potential of that population. Contemporary genetic diversity also reflects a population's demographic and evolutionary history (Epps & Keyghobadi, 2015; Templeton, Routman, & Phillips, 1995; Vandergast, Bohonak, Weissman, & Fisher, 2007). The field of population genetics is concerned with evaluating the amount and distribution of genetic diversity within and among populations, and uncovering the processes that have shaped that genetic diversity (Halliburton, 2004). Population genetic theory is based on the understanding that four fundamental processes – mutation, selection, genetic drift, and gene flow – are responsible for determining how genetic diversity is distributed within and among populations.

Mutations are changes in DNA sequence; examples include insertions, deletions or inversions of sequences, as well as point mutations where a single nucleotide changes state. Mutations are the ultimate source of genetic diversity, and the rate at which mutations occur differs depending on the category of DNA (e.g., mitochondrial vs genomic; Ballard & Whitlock, 2004) as well as across taxa (Britten, 1986). Selection occurs when individuals have differential survival and reproductive success based on heritable traits, and results in certain genotypes being represented at higher or lower frequencies in subsequent generations (Halliburton, 2004). Depending on the type of selective pressure a population experiences, selection can act either to maintain or to decrease genetic diversity. Genetic diversity tends to decrease at loci under stabilizing or directional selection, where a specific phenotype and its underlying genotype consistently have the highest fitness (Lewontin, 1964). In contrast, genetic diversity is maintained under various forms of balancing selection, which can include heterozygote advantage (where heterozygous genotypes have highest fitness), frequency dependent selection

(where genotypes occurring at lower frequencies have higher fitness; Brisson, 2018), as well as temporally or spatially varying selection (Bürger & Gimelfarb, 2002).

Genetic drift refers to changes in allele frequencies that occur between generations due to stochastic sampling of alleles in finite populations (e.g., due to differential mating success among individuals, recombination during meiosis, etc.; Halliburton, 2004). Genetic drift, over time, results in a decrease in genetic diversity as rare alleles may be lost by chance between generations; this effect is more dramatic in smaller populations, where each individual represents a larger proportion of the total population. Gene flow, or the movement of alleles among populations, is the result of dispersal followed by the successful reproduction of the dispersed individual. Gene flow can introduce new genetic variation into populations and has a homogenizing effect on the variation among different populations. Models of gene flow (e.g., Wright's island model and models of isolation by distance) show that the allele frequencies of populations are more similar when they share greater rates of gene flow (Bohonak, 1999; Wright, 1943). Overall, the amount of genetic diversity in a population at any time is determined by the balance among mutation, selection, genetic drift and gene flow (Wright, 1931).

Population genetics began as a primarily theoretical field in the early part of the last century, aimed at incorporating the principles of Mendelian inheritance into the framework of evolution by natural selection (B. Charlesworth & Charlesworth, 2017). The field was initially built on theoretical developments, with empirical data from natural populations available primarily from key systems in which visible morphological traits had a relatively simple heritable basis (B. Charlesworth & Charlesworth, 2017), or from inversion polymorphisms that could be assessed by karyoptyping, for example in *Drosophila pseudoobscura* populations (Sturtevant & Dobzhansky, 1936). The use of molecular approaches to quantify genetic variation in natural populations was not initiated until the 1960s, with the first assessment of variation at allozymes in populations (Hubby & Lewontin, 1966; Lewontin & Hubby, 1966). Since that groundbreaking study, empirical data collected at the molecular level, describing genetic variation within and among natural populations, has increased dramatically (Casillas & Barbadilla, 2017). Methods for sequencing DNA or otherwise assessing variation at the DNA level have

continued to evolve with increasing rapidity, providing population geneticists with increasing power to quantify and study the genetic diversity of natural populations (Levy & Boone, 2019).

1.2 Molecular markers

Even with advances in sequencing technology, it is still difficult and expensive to assess diversity across the entire genome. Instead, molecular markers are used that are assumed to reflect the overall diversity of the genome. Molecular markers are biological molecules that can be interrogated through a variety of methods to reveal the underlying DNA sequence variation. These methods include approaches that differentiate protein allozymes, assess DNA structure, and, at the finest scale, record nucleotide sequences. The most useful markers for assessing population genetic structure are co-dominant, DNA-level markers (Fu, 2000; Milligan & McMurry, 1993). Co-dominant markers provide information about both alleles carried by each individual, whereas dominant markers only indicate the presence or absence of a dominant allele and cannot distinguish between heterozygotes and homozygotes, thereby complicating the process of estimating population allele frequencies.

Some molecular markers reflect differences in the genome that may affect an organism's fitness; this includes mutations that change the level of gene expression (e.g., mutations in a promotor region), as well as mutations in expressed sequences. Such markers are called non-neutral molecular markers, or, when more clearly known to be under selection, adaptive molecular markers. These non-neutral markers include protein allozymes (e.g., in the lizard *Podarcis tiliguerta*; Capula, 1996), and DNA sequences for regions of the genome known or hypothesized to have a function (e.g., the major histocompatibility complex; Campos, Posada, & Morán, 2006; Charlesworth, 2006). In contrast, neutral molecular markers reflect differences in the genome that are not believed to affect an organism's fitness (Holderegger, Kamm, & Gugerli, 2006). As they do not typically result in differences in external phenotype, neutral molecular markers are usually observed through direct interrogation of the DNA sequence. At the population level, variation at neutral molecular markers reflects primarily the effects of drift and

gene flow (Holderegger et al., 2006). Variation at adaptive molecular markers is also influenced by drift and gene flow, but additionally by selection (Kirk & Freeland, 2011).

The use of molecular markers has evolved with the available technology. Early molecular markers, such as protein allozymes (Hubby & Lewontin, 1966), were non-neutral. DNA sequences could be used as molecular markers with the development of the Sanger sequencing method (Sanger, Nicklen, & Coulson, 1977) and the Maxam-Gilbert method (Maxam & Gilbert, 1977). However, manual DNA sequencing was too costly and time consuming to assess multiple loci across the sample sizes required for population genetic studies. Other molecular markers were developed based on using gel electrophoresis to assess differences in the lengths of DNA fragments among individuals, rather than the actual sequence of DNA fragments. The first of these were restriction fragment length polymorphisms (RFLPs; Botstein, White, Skolnick, & Davis, 1980). In this method, DNA is first digested with restriction enzymes. The resulting fragments are run out on an agarose gel and exposed to short probe sequences of DNA, which are labelled either radioactively or fluorescently and allow a subset of the fragments on the gel to be visualized. Mutations in restriction enzyme cut sites produce different sizes of DNA fragments (as restriction enzymes are unable to bind and cleave if the cut site is mutated), and therefore a different pattern of bands among individuals who carry that mutation. Differences in the frequency of individuals carrying these mutations could then be examined across populations (Beaumont & Nichols, 1996). The specificity of RFLPs as a molecular marker was improved with the development of the polymerase chain reaction (PCR) as a method to amplify DNA (Mullis et al., 1986). In PCR-RFLP, rather than digesting all DNA, specific regions of the genome are first amplified and only those regions are digested and analyzed (Maeda et al., 1989).

The development of PCR also allowed entirely novel types of molecular markers to be developed. The random amplified polymorphic DNA (RAPD) method uses a random primer (usually approximately 10 bp in length) to amplify any sections of the genome that are flanked by the complement to the primer sequence (Welsh & McClelland, 1990; Williams, Kubelik, Livak, Rafalski, & Tingey, 1990). Similar to mutations in restriction enzyme cut sites for RFLP markers, mutations in the primer binding sites prevent

amplification and result in a different banding pattern. RAPD markers are relatively cheap and easy to assess for large numbers of individuals; however, they suffer from a lack of reproducibility (Jones et al., 1997) and are a dominant marker, which provides less information per locus than co-dominant RFLPs.

Amplified fragment length polymorphisms (AFLPs) were developed as a more reliable alternative to RAPD markers, while still not requiring any prior information about the sequence of the genome (e.g., as would be required for PCR-RFLP or microsatellites, see below; Vos et al., 1995). Like RFLPs, AFLPs are partially the result of mutations in restriction enzyme cut sites that produce different banding patterns when visualized on a gel. By using two restriction enzymes that leave "sticky" ends after cutting, adapter DNA sequences with complementary sticky ends can be ligated and only fragments with a different adapter sequence at each end will be amplified. To reduce the number of restriction fragments, during PCR amplification the primers are designed to be complementary to the adapter sequences plus an additional one to three random bases, so that only fragments with the restriction enzyme cut site plus an additional short nucleotide sequence are amplified (Vos et al., 1995). This provides the benefits of PCR-RFLP, where only a small amount of starting DNA is required, without needing to know the flanking sequences to develop the primers. The major downside of AFLPs is that they are a dominant marker – the band for a particular locus will be present if at least one allele has the unmutated restriction enzyme cut site.

Microsatellites – regions of the genome where a short sequence motif of nucleotides (1-6 bp) is repeated in tandem – were proposed as a novel molecular marker after they were shown to be hypervariable in the fruit fly *Drosophila melanogaster* and in humans (Tautz, 1989). Most microsatellites do not have a function and are presumed to reflect neutral genetic processes, although there are some exceptions (e.g., a small number of microsatellites resulting in diseases in humans; Brouwer, Willemsen, & Oostra, 2009) and it is possible for microsatellites to be physically linked to areas of the genome under selection. In contrast, the location of the mutations underlying RFLP, RAPD, and AFLP markers are typically not known, although most likely fall outside of expressed regions of the genome. Microsatellite discovery is more time consuming and expensive than

optimizing an AFLP protocol (i.e., choosing appropriate restriction enzymes) because it involves first finding microsatellites in the genome and then sequencing flanking regions to design PCR primers. However, after microsatellites are developed it is easier to prepare samples for genotyping than when using AFLPs. Microsatellite preparation requires only a PCR reaction, as compared to the restriction enzyme digestion, adapter ligation, and PCR of AFLP preparation. Microsatellites also have the advantage both of being a co-dominant marker (unlike dominant AFLPs), having high reproducibility, and being hypervariable. Their hypervariability (a result of slippage during DNA replication, resulting in alleles of different lengths) means that once a microsatellite region is discovered in the genome it is likely to have multiple alleles; as a result, each microsatellite marker is more informative than other (e.g., ALFP) markers. Microsatellites were the most commonly used molecular marker in population genetic studies after techniques for their discovery and genotyping were developed in the 1990s (Hodel et al., 2016). While they still remain a popular choice, single nucleotide polymorphisms have emerged as an alternative marker.

DNA sequence data were theoretically able to be used as molecular markers since the development of Sanger and Maxam-Gilbert sequencing, but in practice required both the development of PCR as well as automations to the Sanger sequencing method to be readily applied to generate data on genetic variation at the population level. DNA sequence data from natural populations can be used to explore the process of selection, for example by sequencing and comparing the haplotypes of candidate loci (i.e., loci hypothesized to be under selection) among populations. Sequences of highly conserved genes (e.g., mitochondrial DNA) are useful in exploring evolutionary relationships among species and populations, as well as historical patterns of population size and connectivity (Avise et al., 1987; Ramos-Onsins & Rozas, 2002; Templeton et al., 1995). When DNA sequences are used as molecular markers, the entire haplotype of the sequenced locus is typically used. Alternatively, point mutations – or single nucleotide polymorphisms (SNPs) – can be extracted from sequence data and used as individual molecular markers. Outside of the whole genome sequencing of model species, SNPs outside of expressed sequences historically had to be discovered through fairly laborious processes. For example, a common method was exon primed intron crossing, where PCR

primers were developed based on conserved exon sequences and used to sequence and search for SNPs in neighbouring introns. SNP panels developed using these methods and used for population genetic studies often include less than 20 SNPs (e.g., Ledoux et al., 2012; Tay, Behere, Heckel, Lee, & Batterham, 2008; White, Endersby, Chan, Hoffmann, & Weeks, 2015). The power of these small SNP panels to detect population genetic patterns are often compared to that of microsatellites. While each individual SNP, with typically only two observed alleles, is less informative than a highly variable microsatellite locus (Schopen, Bovenhuis, Visker, & Arendonk, 2008), small SNP panels are still suitable for estimating population parameters including genetic differentiation and diversity (Coates et al., 2009). Nonetheless, such small SNP panels were not frequently used in population genetic studies relative to microsatellites. However, in taxa such as the Lepidoptera where microsatellite discovery is made difficult by low microsatellite frequency and the repetition of flanking sequences among different microsatellite loci (Zhang, 2004), AFLPs and small SNP panels are sometimes used as alternatives to microsatellites (e.g., Collier et al., 2010; Fountain et al., 2016; Gradish, Keyghobadi, & Otis, 2015). Until the mid- to late 2000's, the size of SNP panels was traditionally limited by the cost of developing and sequencing each SNP locus, but the development of next-generation sequencing methods vastly increased the size of SNP panels that could be developed and genotyped for population genetic studies in non-model organisms.

1.3 Next-generation sequencing and molecular markers

First-generation DNA sequencing methods – Sanger and Maxam-Gilbert sequencing – are typically used to sequence small numbers of relatively long (up to 1000 bp) DNA sequences. Second-generation, or next-generation (NGS), sequencing refers to several techniques developed in the early 2000s and implemented on different sequencing platforms (pyrosequencing: Roche 454, ligation sequencing: SOLiD, sequencing by synthesis: Illumina HiSeq) (Liu et al., 2012). Next-generation sequencing differs from first-generation sequencing in that many short (initially up to 150 bp) sequences are sequenced simultaneously, and at a much smaller cost per base pair. These shorter sequences make next-generation sequencing arguably less suitable for characterizing

entire candidate loci compared to Sanger sequencing, but are very suitable for genotyping many independent SNP loci. One application of NGS is to genotype known SNP loci (e.g., in known expressed sequences or, as previously, developed using exon priming intron crossing) for many more individuals than would be possible at the same cost and effort as using Sanger sequencing. Another is to use NGS to discover and genotype previously unknown SNPs across the genome. While it is possible to use NGS platforms to sequence the entire genome of multiple individuals (in order to assess variation at all possible sites in the genome) this is still both costly and time consuming when processing the tens or hundreds of individuals typical of population genetic studies. To address this issue, over the past decade several methods have been developed to isolate a replicable subsample of the genome from which SNPs can be identified (e.g., Baird et al., 2008; N. R. Campbell, Harmon, & Narum, 2015; Elshire et al., 2011; Peterson, Weber, Kay, Fisher, & Hoekstra, 2012). This approach is called "reduced representation" sequencing, referring to using this isolated subsample of the genome to represent the rest of the unsequenced portions of the genome.

One of the earliest of these reduced representation sequencing approaches is restriction site associated DNA sequencing, or RADseq (Baird et al., 2008). In the original RADseq methodology, a single restriction enzyme is used to digest genomic DNA, which is then ligated with flanking "adapter" DNA sequences recognized by the Illumina sequencing platform. To further shorten the restricted DNA sequences, the original protocol uses a sonicator to mechanically shear the fragments. Alternatively, a modified protocol called double digest RAD sequencing (ddRADseq) cuts the DNA with a second restriction enzyme with a different recognition site, producing a more consistent fragment length than the random mechanical shearing of the original protocol (Peterson et al., 2012). In both cases, the reduced representation occurs by retaining only a certain size range of DNA fragments for sequencing (typically between 100-500 bp) and discarding fragments outside of this size range.

RADseq and other reduced representation sequencing approaches allow for the genotyping of thousands or, for some approaches, tens of thousands of SNPs. In some ways these SNP panels are similar to, but much larger than, the SNP panels previously

used for population genetic studies, being co-dominant, biallelic markers that can be either neutral or non-neutral. However, the reduced representation approach leads to several key differences compared to SNP panels designed using Sanger sequencing. With Sanger sequencing, the polymorphic site is either known in advance, or can be easily identified by comparing otherwise identical (in sequence and length) DNA sequences. When using reduced representation sequencing, especially for species without a reference genome, polymorphic sites are initially unknown. SNPs must be identified *in silico* using bioinformatics pipelines (e.g., Stacks, PyRAD) that align the millions of DNA sequences produced into likely copies of the same locus, then search for polymorphic sites among alleles of each putative locus. This approach has several implications for the genotyped SNPs. Overall, much less is known about the characteristics of each SNP, including their potential function, putative neutrality, and linkage status.

SNPs genotyped through reduced representation sequencing also have considerably higher rates of missing data, where individuals are by chance not genotyped at some SNP loci and are therefore "missing" genotype data at those loci (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Crotti, Barratt, Loader, Gower, & Streicher, 2019; Gautier et al., 2013). The approach used to identify SNPs from raw NGS output impacts the characteristics of the resulting SNP panel. Intuitively, it seems safest to use the most stringent parameters at all stages of SNP genotyping, from the initial stages of SNP identification (e.g., requiring more identical reads to identify an allele or allowing fewer differences among alleles) to later stages of filtering to determine which SNPs will be retained for analysis (e.g., setting a maximum amount of missing data per locus). However, both too stringent and too lenient parameters can lead to biases and errors in the ensuing SNP panel (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013; Mastretta-Yanes et al., 2015). For example, using too stringent parameters early in SNP identification can lead to a locus being identified as distinct, when in reality it is actually an allele of another identified locus (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013; Mastretta-Yanes et al., 2015). The same parameters set too leniently can erroneously call multiple distinct loci as alleles of a single locus (Catchen et al., 2013; Mastretta-Yanes et al., 2015). After SNPs are called, parameters are chosen to determine which SNPs are retained for analysis; this process is referred to as filtering, where SNPs

that do not meet the chosen parameters are removed from the dataset. At this stage, filtering too leniently results in SNP panels with high amounts of missing data, which may adversely affect some population genetic analyses (e.g., Shafer et al. 2017). On the other hand, filtering too stringently both reduces the number of SNPs available for analysis and disproportionately removes SNPs at loci that experience higher mutation rates (Huang & Knowles, 2016). In many cases decisions around filtering and parameter selection are not transparent, perhaps based on the assumption that the quantity of markers will outweigh any bias or error introduced at any specific marker (Mastretta-Yanes et al., 2015). Based on a small number of studies that examine the effects of filtering SNP panels on population genetic analyses, some analyses (including population assignment and isolation by distance) are sensitive to filtering parameters such as the maximum amount of missing data per locus while others, including estimates of $F_{ST,}$ are not (Chattopadhyay, Garg, & Ramakrishnan, 2014; Shafer et al., 2017).

RADseq, alongside genotyping-by-sequencing (Elshire et al., 2011), is one of the original approaches to reduced representation sequencing for simultaneous SNP discovery and genotyping; subsequent protocols are often modified versions of RADseq and follow a similar naming system (E. O. Campbell, Brunet, Dupuis, & Sperling, 2018). The trade-off among these various protocols is often between the number of markers returned, the cost and labour of sequencing, and the reliability that all the markers will be sequenced for each individual (Scheben, Batley, & Edwards, 2017). For example, double-digest RADseq returns fewer SNPs on average than single-digest RADseq (Flanagan & Jones, 2018), but by using two enzymes to fragment the genomic DNA, rather than random shearing, the genome is more reliably subsampled and there is less variation in the number of reads among individuals (Peterson et al., 2012).

1.4 *Parnassius smintheus* as a model system

Studying the amount and distribution of genetic diversity in natural populations is a fundamental component of population genetics research (Halliburton, 2004). Population genetic theory can be applied to understand the demographic histories of populations (e.g., inferring recent bottlenecks using contemporary allele frequencies; Luikart, Allendorf, Cornuet, & Sherwin, 1998), and natural populations can be used to test and

validate population genetic theory (e.g., observing a bottleneck, and looking for the expected changes in allele frequencies in subsequent generations, e.g., Cammen et al., 2018; Suárez, Betancor, Fregel, Rodríguez, & Pestano, 2012).

There are many considerations when selecting a study system for population genetic research, particularly when working on animals. For example, studying a species at risk may provide important, species-specific data for conservation work, but also means working with small population sizes and limitations to sampling (e.g., non-lethal or non-invasive sampling). Alpine insect populations are particularly interesting study systems, both for being insects and occupying an alpine habitat. Insects, with their often short generation times and large populations relative to vertebrates, allow researchers to potentially collect large sample sizes. Alpine species provide potential models for responses to climate change, and are particularly vulnerable to climate change effects (Grabherr, Gottfried, & Pauli, 2010). They face similar habitat range shifts as non-alpine species, with their preferred climatic envelopes shifting not only poleward, but also upward in elevation and altitude. However, while it is potentially feasible for some non-alpine (i.e., lower elevation) species to track their shifting habitat along altitudinal gradients, alpine species are limited in the extent to which their habitable range can shift in elevation; eventually, they simply run out of space. In the absence of adapting locally to changing climatic conditions, such species may instead rely on long-distance dispersal events to neighbouring alpine habitat along a latitudinal gradient to avoid extinction under climate change scenarios (Brooker, Travis, Clark, & Dytham, 2007).

One alpine insect that has been well studied in both populations genetic and ecological contexts is the butterfly *Parnassius smintheus*, a species found throughout the Rocky Mountains (Guppy & Shepard, 2001). This species possesses several traits that makes it a good candidate for population genetic research. It is a relatively common butterfly, occurring in many locations where there is suitable habitat and often at high abundance, and is easily identified. Individuals are also relatively large (3-5 cm wingspan) and slow fliers, able to be caught readily using hand nets. *Parnassius smintheus* has very specific habitat requirements, specifically alpine meadows containing the larval hostplant *Sedum lanceolatum*, so habitat patches can be easily identified and delineated.

A set of populations of *P.* smintheus occurring on Jumpingpound Ridge, and other nearby ridges, in Alberta has, in particular, been the subject of intensive study for the past 25 years (Caplins et al., 2014; Jangjoo, Matter, Roland, & Keyghobadi, 2016; Keyghobadi, Roland, & Strobeck, 1999; Matter, Keyghobadi, & Roland, 2014; Roland, Keyghobadi, & Fownes, 2000). Jumpingpound Ridge occurs within a mixed-use area (with both public hiking trails and commercial natural gas extraction) close to the University of Calgary Barrier Lake Field Station. The ridge includes a network of Sedum-containing meadows separated by forest. Many of these meadows contain *P. smintheus* populations that are connected by dispersing individuals (Roland et al., 2000). The Jumpingpound *P. smintheus* populations were surveyed initially in 1995. Individuals in each population were captured with hand nets, marked with unique three letter codes, released and re-captured over a three-week period (Roland et al., 2000). Tissue samples were taken from the wings of a subset of captured individuals and used for microsatellite genotyping (Keyghobadi et al., 1999). In every year since 1995, mark-recapture data has been collected from these populations (Matter et al., 2014; Roland & Matter, 2016). In many but not all of these years (Fig. 1) wing clips have also been collected (Caplins et al., 2014; Jangjoo et al., 2016; Keyghobadi et al., 1999). The initial analyses stemming from the 1995 and 1996 surveys used both dispersal rates from mark-recapture (Roland et al., 2000) and genetic distances calculated from the genotyped wing clips (Keyghobadi et al., 1999) to identify landscape features that influence dispersal and the resulting genetic structure. These studies identified forest as a key determinant of dispersal in *P. smintheus*: individuals move approximately half as frequently through forest than through meadow (Roland et al., 2000). Populations display both patterns of isolation by distance and isolation by resistance: more distant populations are more genetically different, and populations separated by forest are more differentiated, for a given geographic distance, than populations separated by open, non-forested land cover (Keyghobadi et al., 1999).

In 1999, a broader genetic survey of 27 *P. smintheus* populations was conducted in the Rocky Mountain foothills and front ranges (including the Banff and the Kananaskis regions in Alberta) to examine how landscape features at a larger spatial scale than a single ridgeline affect population structure (Keyghobadi, Roland, & Strobeck, 2005). The study area was divided into three regions: East Kananaskis, where Jumpingpound Ridge

is located, is a region of greater forest cover and more fragmented alpine meadow habitat, whereas both West Kananaskis and Banff have larger and more connected alpine meadows. The patterns of genetic differentiation observed at this scale, and the differences in genetic diversity and structure among the three regions, were consistent with the importance of meadow connectivity to *P. smintheus* dispersal that was observed on Jumpingpound Ridge. Patterns of isolation by distance, lower differentiation among populations overall, and higher within-population diversity were present in Banff and West Kananaskis as compared to East Kananaskis; in East Kananaskis the greater amounts of forest cover on the landscape appear to limit movement and gene flow, leading to these patterns.

In 2003, populations of *P. smintheus* at Jumpingpound Ridge rapidly collapsed in size. On average, populations declined by 86% in the estimated adult population size compared to the previous year (Caplins et al., 2014; Jangjoo, Matter, Roland, & Keyghobadi, 2020). Population sizes recovered to pre-collapse levels by 2007. In 2010 a second demographic bottleneck was observed, with estimated population sizes declining by 75% on average from the previous year (Jangjoo et al., 2020). While the decline in average estimated population size was similar between the two bottlenecks, the second bottleneck lasted longer with population sizes remaining at a similarly low level through 2011 before beginning to increase in 2012. Most recently the populations appear to have collapsed in 2019 and remained low in the summer of 2020.

Annual population size change in *P. smintheus* is associated with weather and climate conditions. Specifically, years with higher population growth are associated with moderate values of the Pacific Decadal Oscillation (PDO) index, while extreme values of winter PDO (indicating either wet/cold or warm/dry years) are associated with population declines (Roland & Matter, 2013). As winter PDO had a stronger relationship with *P. smintheus* population growth than PDO overall, poor over-wintering conditions were hypothesized to drive population declines, including the extreme demographic bottleneck events (Roland & Matter, 2013). For example, at one extreme of PDO warm and dry winters may result in too little snow cover to insulate eggs, while at the other cold and wet winters may result in egg freezing during extreme cold snaps or before sufficient

snow has accumulated. However, the demographic bottlenecks remained difficult to predict *a priori* based solely on PDO conditions. Specifically, a bottleneck in the Jumpingpound Ridge populations was predicted for the summer of 2015 based on the high PDO values observed in the winter of 2014 (Matter & Roland, 2015); however, no bottleneck occurred that year. As a result, additional potential weather predictors of population growth were considered. By using data from a local snow pillow and weather station, a finer scale model considering weather at Jumpingpound Ridge found a stronger relationship between November temperature and snowfall and population growth than a model using PDO as a predictor (Roland & Matter, 2016).

Demographic bottlenecks are expected to have potentially strong genetic consequences for a population. Typically, a bottleneck results in a loss of genetic diversity and a breakdown of patterns of genetic differentiation (e.g., isolation by distance) as alleles are lost by chance in each population (Allendorf, 1986; Chakraborty & Nei, 1977). The repeated bottlenecks in the Jumpingpound Ridge populations and their subsequent recoveries, combined with the ongoing mark-recapture and tissue sampling of these populations, provide the opportunity to test these theoretical genetic consequences. By measuring genetic diversity within and among the *P. smintheus* populations on Jumpingpound Ridge before, during and after the bottlenecks, a pattern of a loss of genetic diversity through genetic drift, followed by a recovery of genetic diversity through gene flow, has emerged (Caplins et al., 2014; Jangjoo et al., 2016, 2020).

1.5 Overview of Thesis

Here, I use a next-generation sequencing approach to develop new SNP datasets to examine genetic variation in *Parnassius smintheus* at two spatial scales, as well as over time. I demonstrate the importance of landscape and weather in shaping genetic variation at a broader spatial scale, and I show how genetic drift, gene flow, and potentially selection shape genetic variation over repeated bottlenecks at a smaller spatial scale. By developing these SNP datasets and evaluating their strengths and weaknesses when applied to this system, I also contribute a new method of quantifying genetic variation in the ongoing study of this unique and informative study system.

In Chapters 2 and 3, I focus on the development, limitations, and strengths of a reduced representation sequenced SNP dataset. In Chapter 2, I evaluate how filtering SNP datasets by changing the maximum permitted amount of missing data affects common population genetic analyses. In Chapter 3, I explore the hypothesis that the large number of SNPs produced by reduced representation sequencing can allow for few individuals to be sampled per population (while retaining the power to detect population genetic patterns).

In Chapters 4 and 5, I use the SNP dataset developed and tested in Chapters 2 and 3 to examine the factors underlying spatial and temporal variation in genetic diversity in *P. smintheus* populations. In Chapter 4, I assess the contributions of weather and landscape to patterns of genetic diversity across *P. smintheus* populations separated by tens to hundreds of kilometers. Specifically, I test site-specific metrics of landscape (e.g., patch size) and weather (e.g., average temperature in November) as predictors of either genetic distance or genetic diversity. In Chapter 5, I examine how genetic diversity and differentiation change over two demographic bottleneck events in the Jumpingpound Ridge populations. Here, I derive a smaller SNP dataset, from the larger dataset used in Chapters 2-4, that is suitable for genotyping the small quantities of DNA extracted from the wing clips collected at this site. By using a SNP panel, rather than the previously used microsatellites, to characterize the populations on Jumpingpound Ridge, I can examine changes at both putatively neutral and expressed loci to assess how genetic diversity changes over repeated bottlenecks as a result of genetic drift, gene flow, and potentially selection.

1.6 Literature cited

Allendorf, F. W. (1986). Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology*, *5*(2), 181–190.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81–92.

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., … Saunders, N. C. (1987). Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, *18*, 489–522.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced rad markers. *PLoS ONE*, *3*(10), e3376.

Ballard, J. W. O., & Whitlock, M. C. (2004). The incomplete natural history of mitochondria. Molecular Ecology, 13(4), 729–744.

Beaumont, M. A., & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London B: Biological Sciences*, *263*(1377), 1619–1626.

Bohonak, A. J. (1999). Dispersal, gene flow, and population structure. *The Quarterly Review of Biology*, *74*(1), 21–45.

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, *32*(3), 314–331.

Brisson, D. (2018). Negative frequency-dependent selection is frequently confounding. *Frontiers in Ecology and Evolution*, *6*(10).

Britten, R. J. (1986). Rates of DNA sequence evolution differ between taxonomic groups. Science, 231(4744), 1393–1398.

Brooker, R. W., Travis, J. M. J., Clark, E. J., & Dytham, C. (2007). Modelling species' range shifts in a changing climate: The impacts of biotic interactions, dispersal distance and the rate of climate change. *Journal of Theoretical Biology*, *245*(1), 59–65.

Bürger, R., & Gimelfarb, A. (2002). Fluctuating environments and the role of mutation in maintaining quantitative genetic variation. *Genetics Research*, *80*(01), 31–46.

Cammen, K. M., Schultz, T. F., Bowen, W. D., Hammill, M. O., Puryear, W. B., Runstadler, J., … Kinnison, M. (2018). Genomic signatures of population

bottleneck and recovery in Northwest Atlantic pinnipeds. *Ecology and Evolution*, *8*(13), 6599–6614.

Campbell, E. O., Brunet, B. M. T., Dupuis, J. R., & Sperling, F. A. H. (2018). Would an RRS by any other name sound as RAD? *Methods in Ecology and Evolution*, *9*(9), 1920–1927.

Campbell, N. R., Harmon, S. A., & Narum, S. R. (2015). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, *15*(4), 855–867.

Campos, J. L., Posada, D., & Morán, P. (2006). Genetic variation at MHC, mitochondrial and microsatellite loci in isolated populations of Brown trout (*Salmo trutta*). *Conservation Genetics*, *7*(4), 515–530.

Caplins, S. A., Gilbert, K. J., Ciotir, C., Roland, J., Matter, S. F., & Keyghobadi, N. (2014). Landscape structure and the genetic effects of a population collapse. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1796), 20141798.

Capula, M. (1996). Evolutionary genetics of the insular lacertid lizard *Podarcis tiliguerta*: Genetic structure and population heterogeneity in a geographically fragmented species. *Heredity*, *77*(5), 518–529.

Casillas, S., & Barbadilla, A. (2017). Molecular population genetics. *Genetics*, *205*(3), 1003–1035.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140.

Chakraborty, R., & Nei, M. (1977). Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. *Evolution*, *31*(2), 347–356.

Charlesworth, B., & Charlesworth, D. (2017). Population genetics from 1966 to 2016. *Heredity*, *118*(1), 2–9.

Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, *2*(4), e64.

Chattopadhyay, B., Garg, K. M., & Ramakrishnan, U. (2014). Effect of diversity and missing data on genetic assignment with RAD-Seq markers. *BMC Research Notes*, *7*, 841.

Collier, N., Gardner, M., Adams, M., McMahon, C. R., Benkendorff, K., & Mackay, D. A. (2010). Contemporary habitat loss reduces genetic diversity in an ecologically specialized butterfly. *Journal of Biogeography*, *37*(7), 1277–1287.

Crotti, M., Barratt, C. D., Loader, S. P., Gower, D. J., & Streicher, J. W. (2019). Causes and analytical impacts of missing data in RADseq phylogenetics: Insights from an African frog (Afrixalus). *Zoologica Scripta*, *48*(2), 157–167.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLOS ONE*, *6*(5), e19379.

Epps, C. W., & Keyghobadi, N. (2015). Landscape genetics in a changing world: Disentangling historical and contemporary influences and inferring change. *Molecular Ecology*, *24*(24), 6021–6040.

Flanagan, S. P., & Jones, A. G. (2018). Substantial differences in bias between single-digest and double-digest RAD-seq libraries: A case study. *Molecular Ecology Resources*, *18*(2), 264–280.

Fountain, T., Nieminen, M., Sirén, J., Wong, S. C., Lehtonen, R., & Hanski, I. (2016). Predictable allele frequency changes due to habitat fragmentation in the Glanville fritillary butterfly. *Proceedings of the National Academy of Sciences*, *113*(10), 2678–2683.

Fu, Y. B. (2000). Effectiveness of bulking procedures in measuring population-pairwise similarity with dominant and co-dominant genetic markers. *Theoretical and Applied Genetics*, *100*(8), 1284–1289.

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., … Estoup, A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, *22*(11), 3165–3178.

Grabherr, G., Gottfried, M., & Pauli, H. (2010). Climate change impacts in alpine environments. *Geography Compass*, *4*(8), 1133–1153.

Gradish, A. E., Keyghobadi, N., & Otis, G. W. (2015). Population genetic structure and genetic diversity of the threatened White Mountain arctic butterfly (*Oeneis melissa semidea*). *Conservation Genetics*, *16*(5), 1253–1264.

Guppy, C. S., & Shepard, J. (2001). *Butterflies of British Columbia including western Alberta, southern Yukon, the Alaska Panhandle, Washington, northern Oregon, northern Idaho, northwestern Montana*. Vancouver [B.C.]: UBC Press.

Halliburton, R. (2004). *Introduction to population genetics*. Upper Saddle River, NJ: Pearson/Prentice Hall.

Hodel, R. G. J., Segovia-Salcedo, M. C., Landis, J. B., Crowl, A. A., Sun, M., Liu, X., … Soltis, P. S. (2016). The report of my death was an exaggeration: A review for researchers using microsatellites in the 21st century. *Applications in Plant Sciences*, *4*(6), 1600025.

Holderegger, R., Kamm, U., & Gugerli, F. (2006). Adaptive vs. neutral genetic diversity: Implications for landscape genetics. *Landscape Ecology*, *21*(6), 797–807.

Hubby, J. L., & Lewontin, R. C. (1966). A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics*, *54*(2), 577–594.

Jangjoo, M., Matter, S. F., Roland, J., & Keyghobadi, N. (2016). Connectivity rescues genetic diversity after a demographic bottleneck in a butterfly population network. *Proceedings of the National Academy of Sciences*, *113*(39), 10914–10919.

Jangjoo, M., Matter, S. F., Roland, J., & Keyghobadi, N. (2020). Demographic fluctuations lead to rapid and cyclic shifts in genetic structure among populations of an alpine butterfly, *Parnassius smintheus*. *Journal of Evolutionary Biology*, *33*(5), 668–681.

Keyghobadi, N., Roland, J., & Strobeck, C. (1999). Influence of landscape on the population genetic structure of the alpine butterfly *Parnassius smintheus* (Papilionidae). *Molecular Ecology*, *8*(9), 1481–1495.

Keyghobadi, N., Roland, J., & Strobeck, C. (2005). Genetic differentiation and gene flow among populations of the alpine butterfly, *Parnassius smintheus*, vary with landscape connectivity. *Molecular Ecology*, *14*(7), 1897–1909.

Kirk, H., & Freeland, J. R. (2011). Applications and implications of neutral versus non-neutral markers in molecular ecology. *International Journal of Molecular Sciences*, *12*(6), 3966–3988.

Ledoux, J.-B., Tarnowska, K., Gérard, K., Lhuillier, E., Jacquemin, B., Weydmann, A., … Chenuil, A. (2012). Fine-scale spatial genetic structure in the brooding sea urchin *Abatus cordatus* suggests vulnerability of the Southern Ocean marine invertebrates facing global change. *Polar Biology*, *35*(4), 611–623.

Levy, S. E., & Boone, B. E. (2019). Next-generation sequencing strategies. *Cold Spring Harbor Perspectives in Medicine*, *9*(7), a025791.

Lewontin, R. C. (1964). The interaction of selection and linkage. Ii.Optimum models. *Genetics*, *50*(4), 757–782.

Lewontin, R. C., & Hubby, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. Ii. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics*, *54*(2), 595–609.

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

Maeda, M., Murayama, N., Ishii, H., Uryu, N., Ota, M., Tsuji, K., & Inoko, H. (1989). A simple and rapid method for HLA-DQA1 genotyping by digestion of PCR-amplified DNA with allele specific restriction endonucleases. *Tissue Antigens*, *34*(5), 290–298.

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Molecular Ecology Resources*, *15*(1), 28–41.

Matter, S. F., Keyghobadi, N., & Roland, J. (2014). Ten years of abundance data within a spatial population network of the alpine butterfly, *Parnassius smintheus*. *Ecology*, *95*(10), 2985–2985.

Matter, S. F., & Roland, J. (2015). A priori prediction of an extreme crash in 2015 for a population network of the alpine butterfly, *Parnassius smintheus*. *Ecosphere*, *6*(10), 1–4.

Milligan, B. G., & McMurry, C. K. (1993). Dominant vs. Codominant genetic markers in the estimation of male mating success. *Molecular Ecology*, *2*(5), 275–283.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, *51*, 263–273.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5), e37135.

Ramos-Onsins, S. E., & Rozas, J. (2002). Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution*, *19*(12), 2092–2100.

Roland, J., Keyghobadi, N., & Fownes, S. (2000). Alpine Parnassius butterfly dispersal: Effects of landscape and population size. *Ecology*, *81*(6), 1642–1653.

Roland, J., & Matter, S. F. (2013). Variability in winter climate and winter extremes reduces population growth of an alpine butterfly. *Ecology*, *94*(1), 190–199.

Roland, J., & Matter, S. F. (2016). Pivotal effect of early-winter temperatures and snowfall on population growth of alpine *Parnassius smintheus* butterflies. *Ecological Monographs*, *86*(4), 412–428.

Scheben, A., Batley, J., & Edwards, D. (2017). Genotyping-by-sequencing approaches to characterize crop genomes: Choosing the right tool for the right application. *Plant Biotechnology Journal*, *15*(2), 149–161.

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically

impacts downstream population genetic inference. *Methods in Ecology and Evolution*, *8*(8), 907–917.

Sturtevant, A. H., & Dobzhansky, Th. (1936). Inversions in the third chromosome of wild races of *Drosophila pseudoobscura*, and their use in the study of the history of the species. *Proceedings of the National Academy of Sciences of the United States of America*, *22*(7), 448–450.

Suárez, N. M., Betancor, E., Fregel, R., Rodríguez, F., & Pestano, J. (2012). Genetic signature of a severe forest fire on the endangered Gran Canaria blue chaffinch (*Fringilla teydea polatzeki*). *Conservation Genetics*, *13*(2), 499–507.

Tay, W. T., Behere, G. T., Heckel, D. G., Lee, S. F., & Batterham, P. (2008). Exon-primed intron-crossing (EPIC) PCR markers of *Helicoverpa armigera* (Lepidoptera: Noctuidae). *Bulletin of Entomological Research*, *98*(5), 509–518.

Templeton, A. R., Routman, E., & Phillips, C. A. (1995). Separating population structure from population history: A cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics*, *140*(2), 767–782.

Vandergast, A. G., Bohonak, A. J., Weissman, D. B., & Fisher, R. N. (2007). Understanding the genetic effects of recent habitat fragmentation in the context of evolutionary history: Phylogeography and landscape genetics of a southern California endemic Jerusalem cricket (Orthoptera: Stenopelmatidae: Stenopelmatus). *Molecular Ecology*, *16*(5), 977–992.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., … Kuiper, M. (1995). AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Research*, *23*(21), 4407–4414.

Welsh, J., & McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research*, *18*(24), 7213–7218.

White, V. L., Endersby, N. M., Chan, J., Hoffmann, A. A., & Weeks, A. R. (2015). Developing Exon-Primed Intron-Crossing (EPIC) markers for population genetic studies in three Aedes disease vectors. *Insect Science*, *22*(3), 409–423.

Williams, J. G. K., Kubelik, A. R., Livak, K. J., Rafalski, J. A., & Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research*, *18*(22), 6531–6535.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, *16*(2), 97–159.

Wright, S. (1943). Isolation by distance. *Genetics*, *28*(2), 114–138.

Chapter 2

2 The effect of missing data in RADseq-generated SNP datasets on common population genetic analyses

2.1 Introduction

Population genetics studies have benefited from next-generation sequencing (NGS) techniques, which have allowed simultaneous *de novo* genotyping across many loci without the need for a reference genome (Davey & Blaxter, 2010). In comparison to traditional genetic markers such as microsatellites, single nucleotide polymorphism (SNP) datasets that are next generation sequenced have lower genotyping success rates (where loci are not genotyped in some individuals) and higher error rates (Hodel et al., 2017). However, the large number of loci (typically in the 1000's, compared to dozens for microsatellites) mean that analyses can recoup some of the power lost to inaccuracies (Hodel et al., 2017).

This trade-off between power, gained through a large number of genetic markers, and accuracy is also seen among different SNP datasets, especially those generated through reduced representation sequencing. Reduced representation sequencing approaches, where a subset of the genome is sequenced on a NGS platform, are increasingly a common way to generate large SNP datasets (Casillas & Barbadilla, 2017; Sunde, Yıldırım, Tibblin, & Forsman, 2020). These approaches can lead to both genotyping failure or miscalled SNPs, with a trade-off between the total number of SNPs included in a dataset and the inclusion of SNPs with high rates of genotyping failure. Genotyping failure or miscalling in such datasets can occur due to mutations in sites proximate to the SNP locus or stochastic errors in library preparation and sequencing. In many reduced representation sequencing protocols, restriction enzymes are used to fragment DNA, and DNA adjacent to cut sites in some of these fragments is then sequenced. A mutation in a particular cut site would prevent a restriction fragment from being produced, DNA sequencing of that fragment would not occur, and any SNPs present in that fragment would not be called for individuals carrying that mutation (Luca, Hudson, Witonsky, & Rienzo, 2011). This "allelic dropout" is analogous to null alleles in microsatellites, where amplification is

prevented due to mutations in the PCR primer binding site (Callen et al., 1993). In both cases, heterozygotes that carry a single copy of the mutated site are erroneously called as homozygotes, as only one of the two alleles is sequenced or observed. In addition to allelic dropout resulting from cut site mutations, errors in library preparation can occur and also result in heterozygotes being called as homozygotes (Rokas & Abbot, 2009). This occurs when there are insufficient reads (i.e., the number of times a DNA fragment was sequenced) of the second allele during sequencing, and only one allele is called. The second allele can be lost at several points in library preparation, including during PCR, where differences in amplification efficiency between alleles results in allele-biased PCR, during size selection of restriction fragments, where a small proportion of fragments at the desired size are removed along with the undesired fragments, and as a result of low read numbers during the sequencing process (Davey et al., 2013; Huang & Knowles, 2016).

The consequences of genotyping error in reduced representation datasets on population genetic analyses have been studied mostly in the context of allele dropout. Simulated datasets including allele dropout show inflated differentiation among populations, and either over- (Gautier et al., 2013) or under-estimated diversity within populations (Arnold, Corbett-Detig, Hartl, & Bomblies, 2013). In addition to allelic dropout (where a single allele is not called), reduced representation SNP datasets also face the problem of missing data. Missing data occurs when, for some individuals, neither allele is called. This occurs for the same reasons a single allele is not called – either both alleles have dropped out due to being homozygous for cut site mutations, or there are too few reads for either allele to be called. The problem of entirely missing genotypes at some loci resulting from library preparation error (and not allelic dropout) has been anticipated (Pool, Hellmann, Jensen, & Nielsen, 2010; Rokas & Abbot, 2009), but has not been broadly studied in the context of implications for population genetic analyses (but see Chattopadhyay, Garg, & Ramakrishnan, 2014 and Shafer et al., 2017).

Where genotyping has failed for some individuals at a particular locus, a key decision for the researcher is whether to retain that locus for analysis, given that some proportion of individuals will have missing data at that locus. This decision is made by applying a cut-

off for accepting a certain percent of individuals where genotyping has failed; the more stringent the criteria for maximum genotyping failure (i.e., the lower the proportion of individuals with missing data at a locus that are permitted), the more loci that will be excluded from the final dataset (Huang & Knowles, 2016). Using stringent cut-offs (<10% of individuals permitted to have missing genotypes at a given locus) results in biases when reconstructing phylogenetic histories (Huang and Knowles 2016), and reduces the power of individual genetic assignment (Chattopadhyay et al., 2014) and the estimated strength of isolation by distance (IBD; i.e., the correlation between genetic and geographic distances between populations; Shafer et al., 2017). These biases may result from having small SNP datasets and excluding especially informative SNP loci (i.e., those with low minor allele frequencies). Conversely, using lenient (>60%) cut-offs produces larger SNP datasets but arguably at the cost of adding lower quality loci. While estimation of some parameters, such as global genetic differentiation, appear robust to the use of lenient cut-offs for missing data, other parameters and analyses are more sensitive; for example, the degree of IBD (i.e., the correlation coefficient between geographic and genetic distance) is lower in datasets with lenient cut-offs compared to datasets with intermediate cut-offs (Shafer et al., 2017).

Existing studies on the effects of varying the cut-off for missing genotypes on population genetic analyses are limited by either using a small number of individuals (n=10; Chattopadhyay et al 2015) or few levels of missingness (3; Shafer et al., 2017). Quantifying the sensitivity of population genetic analyses to finer variations in the permitted amount of missing data is important for informing best practices in analyzing NGS SNP datasets. Here, I examine how different levels of maximum permitted missing data affect the accuracy of inferences about genetic diversity and differentiation in populations of the butterfly *Parnassius smintheus.* I use SNP datasets generated by double digest restriction site associated DNA sequencing (ddRADSeq; Peterson, Weber, Kay, Fisher, & Hoekstra, 2012) to explore how varying the maximum permitted percentage of individuals that have missing genotypes at a given locus, henceforth referred to as the permitted proportion of "missing data" (and as a result, the number of SNPs) affects several common population genetic analyses: genetic differentiation, IBD, and population assignment. This approach results in datasets that vary both in their

number of SNPs and their proportion of missing data; while these variables are confounded, this reflects the real trade-off between dataset size and quality present in SNP datasets and allows decisions around that trade-off to be examined.

Parnassius smintheus populations in western Alberta have been previously studied at two spatial scales; this previous research provides a baseline understanding of the ecology and genetics of these systems. The populations of *P. smintheus* that I genotyped have been previously characterized by microsatellite loci, and show a significant pattern of IBD (Keyghobadi, Roland, & Strobeck, 2005). In these *P. smintheus* populations, most sample sites are expected to contain a single population, with few or no individuals expected to be immigrants from another sampled population (or a nearby population that acted as a stepping stone for dispersal). Both microsatellite and mark-recapture studies have also been conducted on a different set of populations on Jumpingpound Ridge, Alberta, and used to characterize genetic patterns (including significant IBD), dispersal rates and changes in population size (Caplins et al., 2014; Jangjoo, Matter, Roland, & Keyghobadi, 2016; Keyghobadi, Roland, & Strobeck, 1999). These studies allow me to evaluate both the precision of SNP datasets that vary in the level of missing data, and their accuracy relative to known population genetic and ecological metrics. Analyses on the same set of populations can have different results when using SNPs versus microsatellites (e.g., Jeffries et al., 2016; Lemopoulos et al., 2019); however, the difference is typically power to detect a pattern (e.g., IBD) rather than in the type of pattern detected.

I hypothesized that including more SNPs would increase the power of population genetic analyses, but that including lower quality (i.e., with more missing data) SNPs would decrease accuracy by adding random errors to the data. I therefore predicted that there would be an intermediate level of missing data, and resulting SNP dataset size, that would most accurately estimate population genetic patterns. However, previous studies demonstrate that the effect of missing data depends on the population genetic analysis used. Analyses that require the estimation of fewer parameters (e.g., global genetic differentiation) are more robust to missing data, while analyses that require the estimation of more parameters (e.g., analyses of spatial genetic structure) are most powerful when using an intermediate level of missing data (Shafer et al, 2017). I predicted that

parameters of diversity (allelic richness and expected heterozygosity) and global genetic differentiation (FST) would be minimally impacted by the amount of missing data per SNP dataset, because these analyses require the estimation of fewer parameters. I predicted that analyses of IBD and population assignment, which require the estimation of many parameters, would be more sensitive to the amount of missing data. Specifically, I predicted that compared to datasets with intermediate levels of missing data, the strength of IBD and the number of populations identified would be lower in the most stringent and most permissive dataset. In the most stringent dataset this would be a result of the low number of SNPs, and in the most permissive dataset a result of introducing random errors to the data.

2.2 Methods

2.2.1 Study species and sampling

Parnassius smintheus is an alpine butterfly, typically occurring at high elevations in meadows above the treeline, whose range extends from Yukon to New Mexico (Guppy & Shepard, 2001). My study populations are located in alpine meadows in the Rocky Mountains of western Alberta. Here, *P. smintheus* populations fly from July to September and lay eggs which overwinter as first instar larvae. Larvae emerge after snowmelt and feed on their hostplant, *Sedum lanceolatum*. I examine a subset of the populations sampled in western Alberta that were previously characterized using microsatellite loci by Keyghobadi et al. (2005). Whole *Parnassius smintheus* adults were collected in 1995 and 1996 from 21 sites located in three regions (East Kananaskis, West Kananaskis, and Banff) of western Alberta (Keyghobadi et al. 2005; Figure 2.1; Table 2). Individuals were caught using hand nets and stored at -80 °C in glassine envelopes.

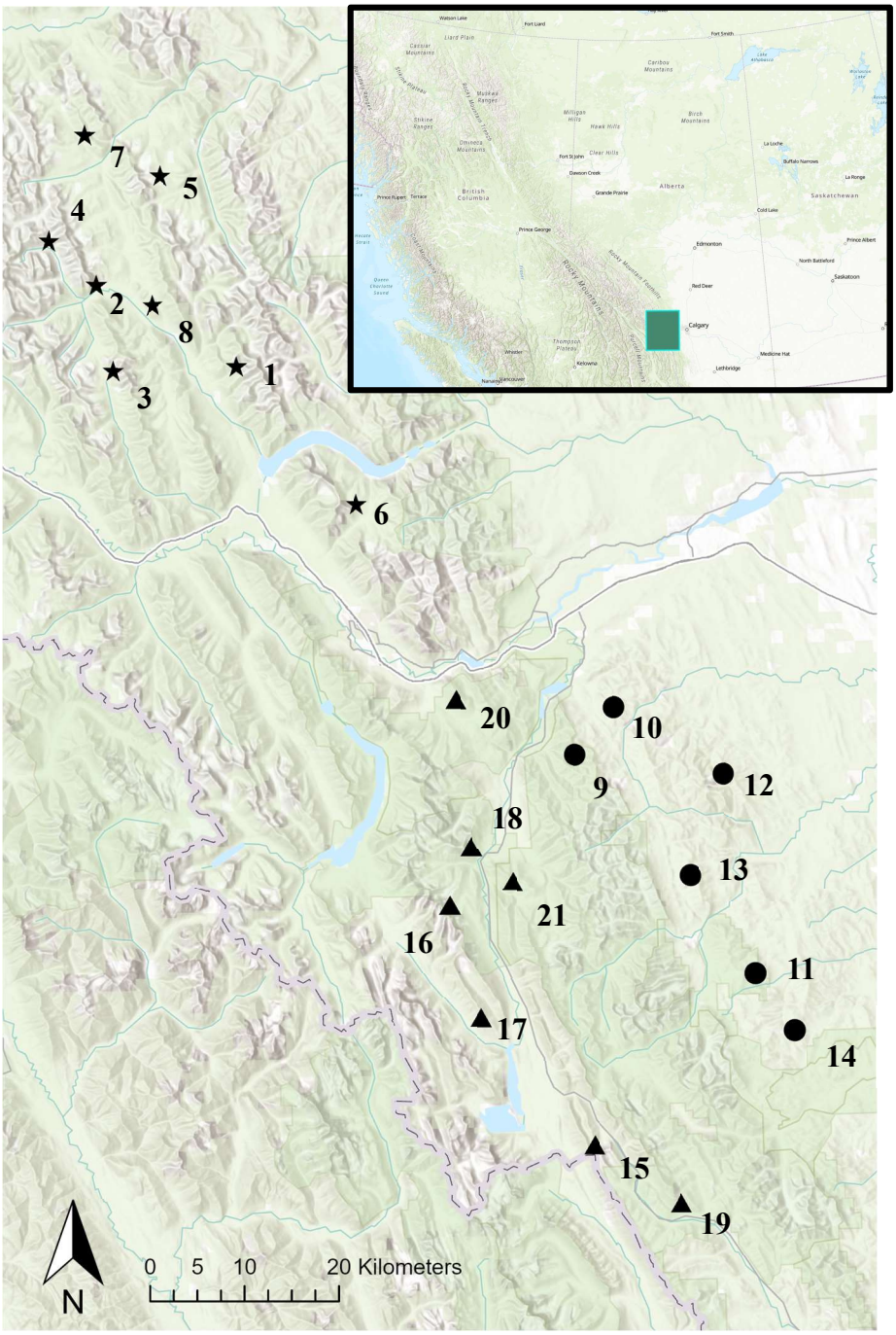

Figure 2.1 *Parnassius smintheus* individuals were sampled from 21 populations in Alberta, from 3 regions: Banff (★), East Kananaskis (●), and West Kananaskis (▲). Inset shows the extent of sampling sites in Alberta, Canada. Map data: Esri, NASA, NGA, USGS, Esri Canada, Esri, HERE, Garmin, Safegraph, FAO, METI/NASA, USGS, EPA, NRCan, Parks Canada.

Table 2.1 *Parnassius smintheus* individuals were sampled from 21 populations in Alberta, from 3 regions: Banff (B), East Kananaskis (EK), and West Kananaskis (WK). Numbers refer to the sampling locations on Figure 2.1.

Population	No. on map	Region	Sample size	Latitude	Longitude
Cascade 1	1	B	11	51.3424	-115.5287
Flint Peak	2	B	38	51.4227	-115.7399
FortyMile Creek	3	B	20	51.3405	-115.7165
North Cascade 1	4	B	16	51.4654	-115.8105
Panther Mtn	5	B	36	51.5257	-115.6395
Mount Peechee	6	B	9	51.2088	-115.3510
Snow Creek	7	B	32	51.5657	-115.7537
Stony Creek	8	B	38	51.4024	-115.6548
Mount Baldy	9	EK	13	50.9651	-115.0297
E (Lusk Ridge)	10	EK	13	51.0091	-114.9697
Forget-Me-Not Ridge	11	EK	20	50.7514	-114.7675
Moose Mtn	12	EK	19	50.9426	-114.8065
Powderface Ridge	13	EK	20	50.8470	-114.8602
Volcano Ridge	14	EK	9	50.6964	-114.7118
Elk	15	WK	8	50.5942	-115.0150
Fortress Mtn	16	WK	39	50.8260	-115.2239
Mount Kent	17	WK	36	50.7180	-115.1806
Mount Kidd	18	WK	12	50.8812	-115.1894
Mist Mtn	19	WK	12	50.5369	-114.8879
Pigeon Mtn	20	WK	19	51.0219	-115.2067
Wedge	21	WK	36	50.8470	-115.1278

2.2.2 SNP library preparation

I used DNeasy Blood and Tissue kits (Qiagen, Germantown, MD) to extract DNA from the head and thorax of between 8 and 38 individuals collected at each of 21 sites, for a total of 501 individuals. I treated DNA extractions with RNAse before digestion with the restriction enzymes *Nla*III and *Eco*RI-HF (New England Biolabs, Ipswich, MA). I performed restriction enzyme digestion for 3 hours at 37 ℃ in a total volume of 50 ul, with each digestion containing: 200 ng DNA, 1X CutSmart buffer, 10 Units *Nla*III, 20 Units *Eco*RI-HF, and 1 µg bovine serum albumin. I purified DNA after restriction enzyme digestion using Sera-Mag solid-phase reversible immobilization beads (GE Healthcare Life Sciences, Chicago, IL). For each individual, I used 75 µl of SPRI beads to bind DNA, followed by two washes with 75% ethanol and elution in 40 µl of nuclease-free water. I estimated the concentration of digested DNA using a Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA), and standardized concentrations to the lowest observed across all samples to be combined in a single lane for sequencing.

Adapter sequences, including Illumina sequences and barcodes for individual identification (Table A1), were ligated to the digested DNA. I performed ligation reactions at 16 ℃ for 8 hours in a 45 µl volume including: standardized digested DNA, 1100 Units T4 ligase (New England Biolabs, Ipswich, MA), 1X T4 ligase reaction buffer, 3.80 pmol adapter P1 (for *Nla*III cut sites), and 2.72 pmol adapter P2 (for *Eco*RI cut sites). I heat killed ligation reactions at 65 ℃ for ten minutes followed by a ramp down at 1.3 ℃ per minute to a final temperature of 21 ℃. I pooled ligated DNA across groups of 40-50 individuals, and size selected for fragments between 200 and 500 bp using Sera-Mag solid-phase reversible immobilization beads (SPRI; GE Healthcare Life Sciences, Chicago, IL). I amplified the pooled, size selected DNA by PCR using i5 and i7 Illumina primers with an additional barcode as per Rašić, Filipović, Weeks, & Hoffmann (2014) in a total volume of 10 µl including: 1X PCR Phusion master mix (New England Biolabs, Ipswich, MA), 2 µM forward primer, 2 µM reverse primer, 1 µl pooled ligated DNA. PCR conditions were: denaturation at 95 ℃ for 30 s, followed by 9 cycles of 98 ℃ for 10 s, 62 ℃ for 30 s, 72 ℃ for 120 s, and finally 72 ℃ for 5 min. I used SPRI beads (at 0.8 times the reaction volume) to clean the PCR reactions, and used an Agilent 2100

Bioanalyzer to verify size selection. Libraries containing 80-100 individuals (representing 2 pooled sets of 40-50 individuals, where each pool was amplified with one of two differently barcoded reverse PCR primers) were sequenced on a single lane of an Illumina HiSeq 2500 sequencer.

I used the program STACKS (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013) to process the raw sequence data and call SNPs. I used the *de novo* pipeline to align sequences and call SNPs, as no published reference genome was available for *Parnassius smintheus* or closely related species. I set parameters for alignment and SNP calling as recommended by Mastretta-Yanes et al. (2015): a minimum stack depth of three reads (m), a maximum of three nucleotides differing between reads in an assembled stack (n), a maximum difference of two nucleotides between combined stacks (M), and a maximum difference of four nucleotides between assembled stacks and additional aligned reads (N). I further filtered SNPs using a minor allele frequency cutoff of 5%. To minimize incidences of linked SNPs in the dataset, I included only SNPs from the forward reads for analysis. Also, in instances of multiple SNPs being present per read, I included only a single SNP per read. I excluded all individuals that had low genotyping success from further analysis, by removing individuals where alleles were identified at fewer than 50% of loci (in a preliminary SNP dataset generated by permitting a maximum of 20% missing data per locus).

2.2.3 Filtering SNP datasets for missing data

Starting with all SNP loci that were remaining after applying the above analysis parameters, I created six different SNP datasets that varied in the maximum permitted percentage of individuals (combined across all populations) that have missing genotypes at a given locus; that is, the datasets varied in what I will refer to as the permitted proportion of "missing data". To generate the six datasets, I used maximum cut-offs of 10, 15, 20, 25, 30, and 40% missing data per locus. For example, a cut-off of 10% meant that all loci where more than 10% of individuals were not successfully genotyped would be excluded. Therefore, the cut-off of 10% was the most stringent and resulted in the smallest SNP dataset. Conversely, the cut-off of 40% was the most lenient and resulted in the largest SNP dataset.

2.2.4 Genetic diversity, differentiation and isolation by distance

I calculated three metrics of genetic diversity for each population (averaged across all loci in each SNP dataset) using the statistical software R (R Core Team, 2015): allelic richness (using the hierfstat package, Goudet 2005), and expected and observed heterozygosity (using the adegenet package, Jombart, 2008). I assessed whether allelic richness differed among SNP datasets using linear mixed models with population as a random effect. I applied a logit transformation to expected and observed heterozygosity, and used linear mixed models (with population as a random effect) to assess whether SNP dataset predicted the transformed values. I used the multcomp package (Bretz, Hothorn, & Westfall, 2010) to assess whether either allelic richness or expected heterozygosity differed significantly (after a Bonferroni correction) between any pair of SNP datasets. I also calculated average minor allele frequency and genotype frequencies across all populations within each dataset (using the packages adegenet and mixIndependR, Song, Woerner, & Planz, 2021).

I calculated global F_{ST} using the hierfstat package (Goudet, 2005) to assess genetic differentiation among all sampled populations. I also calculated pairwise F_{ST} between all pairs of populations using the hierfstat package in R, which calculates Weir and Cockerham's F_{ST} estimate (Weir & Cockerham, 1984), and calculated ninety-five percent confidence intervals using 999 bootstraps. I used maximum likelihood population effects (MLPE) mixed models (Clarke, Rothery, & Raybould, 2002) from the R package nlme (Pinheiro et al., 2015) to assess the degree and significance of IBD. The response variable was pairwise F_{ST}. I estimated the fixed effect, geographic distance between populations, as the Euclidian distance between the centroids of each sampling site. To account for non-independence between the population pairs I used the corMLPE package (Pope, 2020) to apply random effects to each pairing. Models were estimated with a maximum likelihood (ML) approach, for comparison of models across datasets and to a null model of the form F_{ST}~1. I used a restricted maximum likelihood (REML) approach to obtain unbiased estimates of model coefficients to determine the slope of IBD from each dataset.

2.2.5 Population clustering and assignment

I used two population assignment software packages, fastSTRUCTURE (Raj, Stephens, & Pritchard, 2014) and Geneland (Guillot, Mortier, & Estoup, 2005), to assign individuals to putative populations of origin and estimate the number of distinct populations identifiable in each dataset. Both packages use a Bayesian approach to assign individuals to populations in Hardy-Weinberg equilibrium. FastSTRUCTURE is a Python package based on the commonly used population structure software STRUCTURE (Pritchard, Stephens, & Donnelly, 2000), and assigns individuals to a number of populations (K) set by the user. The user can then perform post-hoc analyses to determine the value of K best supported by their data. In fastSTRUCTURE, this is performed with the built-in chooseK function that estimates the number of model components (populations) that are contributing to the model. The function estimates two K values. K^*_{ε} represents the number of populations that optimizes marginal likelihood. $K^*_{\emptyset C}$ represents the number of populations that explain almost all of the ancestry in the dataset. K^*_{ε} is a more stringent indicator of strong contributions to population structure, while $K^*_{\emptyset C}$ is more permissive and includes populations that contribute more weakly to overall structure. K^*_{ε} therefore is typically smaller than $K^*_{\emptyset C}$. I used fastSTRUCTURE rather than STRUCTURE because it is designed to accommodate very large SNP datasets that would be prohibitively complex for STRUCTURE to run. Unlike STRUCTURE, the number of iterations stops automatically once the model converges. I ran fastSTRUCTURE from k=2 to k=30 and used the chooseK function to estimate K^*_{ε} and $K^*_{\emptyset C}$. I repeated this 10 times to get 10 estimates of K^*_{ε} and $K^*_{\emptyset C}$ for each SNP dataset. Due to the non-independence of the data (as each of the 10 replicates used the same SNP dataset) I could not use traditional statistical methods to test for significant differences in K^*_{ε} and $K^*_{\emptyset C}$, and instead looked for trends across the SNP datasets.

I also used Geneland (Guillot, Mortier, & Estoup, 2005) to estimate the number of populations in each dataset. Geneland was designed to directly estimate the number of populations best supported by the data, and subsequently assign individuals to those populations. I ran Geneland 10 times per dataset, setting the possible range of populations

from 1 to 30 over 100 000 iterations with a spatial model and a correlated allele frequency model. In the post-processing stage I removed the burn-in period as the first 200 out of 1000 saved iterations. The run with the lowest posterior probability after removing the burn-in was identified as the best representation of the data. I defined the estimated number of populations in a given dataset as the modal number of populations appearing throughout the MCMC iterations of the best run for that dataset. I calculated the standard deviation in population number within the best run, as well as the standard deviation in average population number among all runs for each dataset.

2.3 Results

I genotyped a total of 501 individuals from 21 sites. After filtering out individuals with low genotyping success, a total of 456 individuals were left for further analysis. The six SNP datasets with different cutoffs for missing data ranged in size from 12 291 SNPs at 40% missing data to 37 SNPs at 10% missing data (Table 2.1).

2.3.1 Genetic diversity, differentiation, and IBD

Mean allelic richness (averaged across all populations in each dataset) tended to decrease as the maximum missing data cut-off was decreased from 40 to 15%; however, the highest mean allelic richness (1.25±0.02) was measured in the 10% missing data dataset (Figure 2.2). Mean expected heterozygosity increased as the cut-off for missing data was increased from 10 to 40% (Figure 2.2). Mean observed heterozygosity did not differ significantly among datasets, except for at 10% missing data (Figure 2.2). Mean minor allele frequency increased consistently with the amount of missing data (Table 2.2). The proportion of major allele homozygote loci (across all loci and all individuals) declined from 84 to 80% as missing data increased, while minor allele homozygote loci increased from 1.8 to 7% and heterozygote loci remained steady at 13-14% (Table 2.2).

Mean pairwise F_{ST} across datasets ranged from 0.072 in the 10% missing data dataset to 0.086 in the 40% missing data dataset. Estimates of pairwise F_{ST} tended to increase with the allowed percentage of missing data, with pairwise F_{ST} being significantly higher when calculated using the 40% missing data dataset than when using the 10 and 15%

missing data datasets (Figure 2.3). Global F_{ST} was highest in the 40% missing data dataset (0.085) and lowest in the 15% missing data dataset (0.074; Table 2.2).

Significant IBD was observed using all SNP datasets. The coefficient of the relationship between geographic distance and genetic distance was similar across datasets, and did not show a trend with the level of missing data. (Table 2.2).

2.3.2 Population clustering and assignment

The largest dataset at 40% missing data and with >12,000 SNPs was used only for analyses of genetic diversity and differentiation; due to the large memory and time requirements it was excluded from population assignment analyses.

Using Geneland, when looking at both the single best run as well the modal number of populations identified across all 10 runs, results were consistent across most datasets: individuals were clustered into either 15 or 16 populations for all datasets except for the one with 10% missing data (Figure 2.4). At 10% missing data, individuals were clustered into 8 populations in the single best run and 9 populations when considering all runs.

For fastSTRUCTURE, I examined the number of populations strongly contributing to structure (K^*_{ε}) separately from the number of populations that had weaker contributions to structure ($K^*_{\emptyset C}$). Mean K^*_{ε} (number of populations strongly contributing to population structure, averaged across 10 replicates) was very similar at 15, 20, and 25% missing data (Figure 2.5a). Mean K^*_{ε} was somewhat lower at 10% missing data (where only a single population was identified) and higher at 30% missing data (where five populations were identified). $K^*_{\emptyset C}$ varied more among datasets than K^*_{ε}. While $K^*_{\emptyset C}$ remained similar at 15, 20, and 25% missing data, it was much higher for both the largest dataset at 30% missing data and the smallest at 10% missing data (Figure 2.5b). However, in the dataset with 10% missing data most individuals were being assigned to populations that did not correspond to the population from which they had been sampled. This was in contrast to the 30% missing data dataset, where a high number of populations were identified with individuals mostly assigned to their population of origin.

Table 2.2 Six ddRADSeq SNP datasets for the alpine butterfly *Parnassius smintheus* were generated by varying the maximum percent missing data per locus, and each was used to calculate basic population genetic parameters including global F_{ST} and pairwise F_{ST}. MAF is the mean minor allele frequency across all loci in the dataset. The strength of isolation by distance was estimated using MLPE models, and the coefficient of geographic distance was estimated using a REML approach.

Percent missing data	# SNPs	Global F_{ST}	Mean pairwise F_{ST} ± s.d.	MAF	Coefficient of geographic distance
10%	37	0.0760	0.0721 ± 0.0483	0.087	1.17E-03
15%	339	0.0745	0.0724 ± 0.0411	0.098	1.06E-03
20%	1098	0.0814	0.0823 ± 0.0473	0.110	1.23E-03
25%	2485	0.0790	0.0805 ± 0.0453	0.120	1.17E-03
30%	4760	0.0824	0.0845 ± 0.0471	0.130	1.20E-03
40%	12291	0.0846	0.0863 ± 0.0476	0.140	1.19E-03

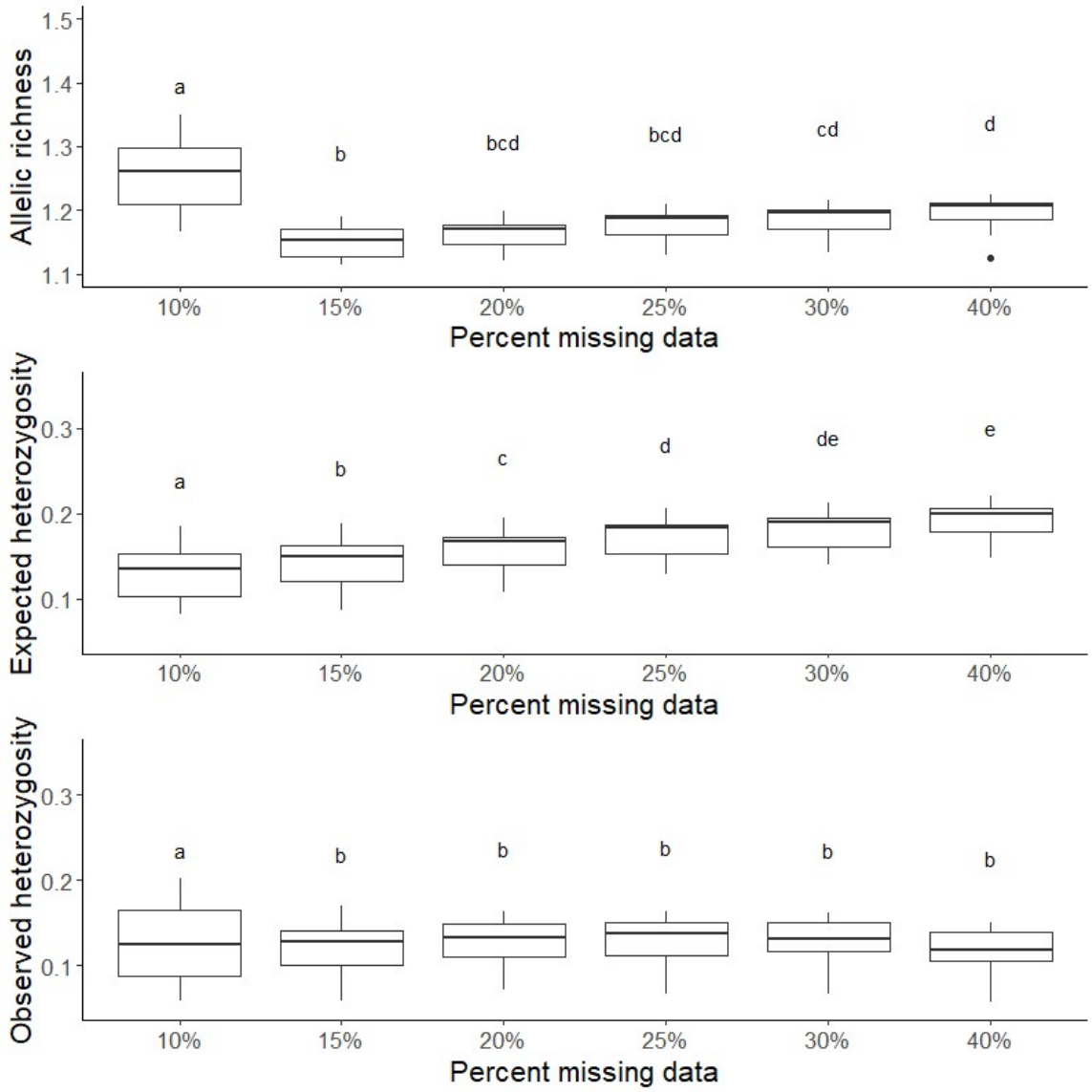

Figure 2.2 Boxplots for three measures of genetic diversity (allelic richness, expected heterozygosity, and observed heterozygosity) estimated for each of 21 *Parnassius smintheus* populations. Each metric was calculated using six SNP datasets (shown on x-axis) that differed in maximum permitted missing data and therefore the total number of loci. Boxes show central 50% of values and the median, across populations. Tails represent values within 1.5 times the interquartile range, and points represent values outside 1.5 times the interquartile range. Letters indicate significant differences calculated from linear mixed models, with population as a random effect.

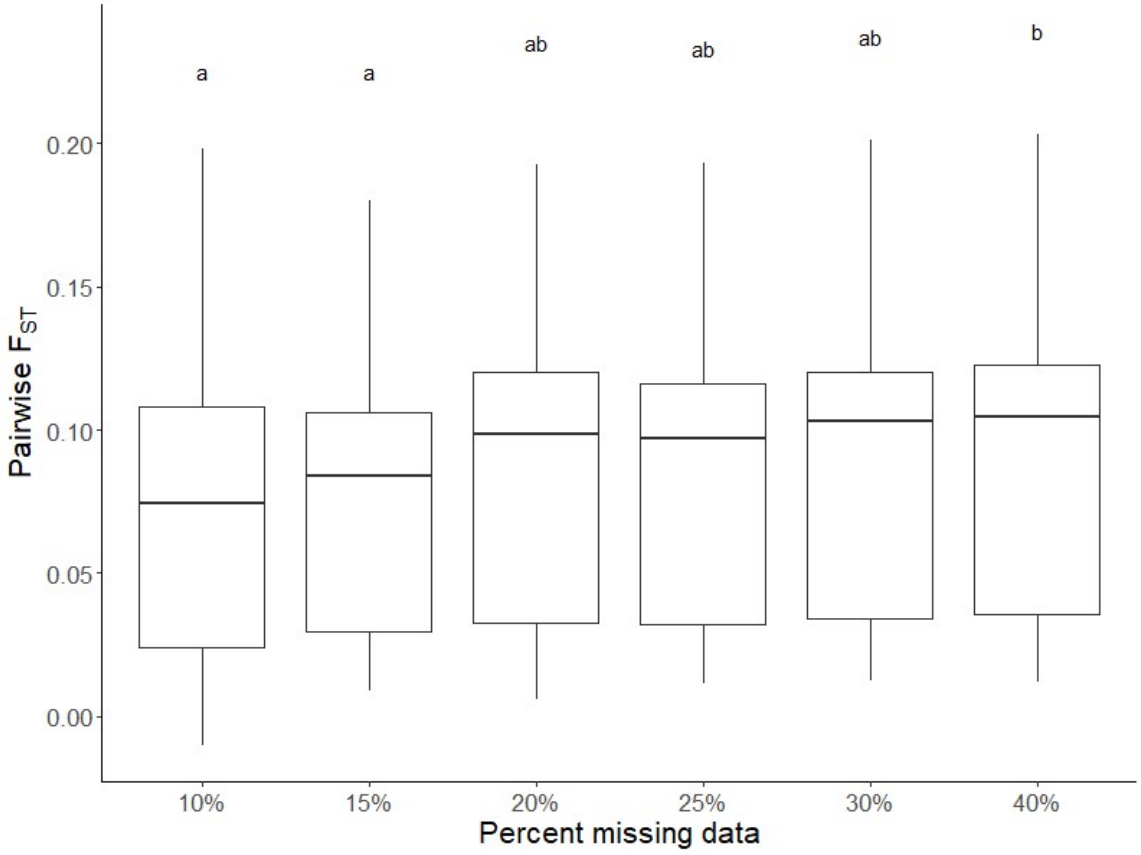

Figure 2.3 Boxplots for pairwise F_{ST}, estimated between each pairwise combination of 21 *Parnassius smintheus* populations, for six SNP datasets (shown on x-axis) that differed in maximum permitted missing data and therefore the total number of loci. Boxes show central 50% of values and the median, across populations. Tails represent values within 1.5 times the interquartile range. Letters indicate significant differences calculated from linear mixed models, with population as a random effect.

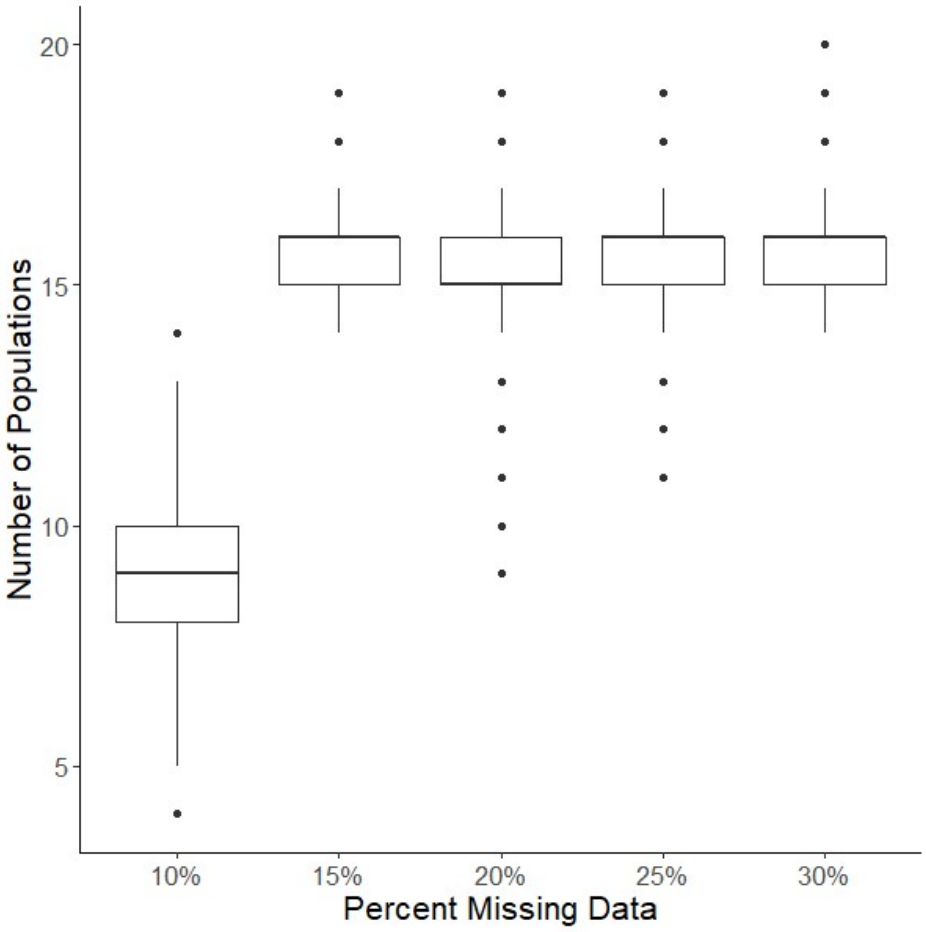

Figure 2.4 Boxplots of the number of populations estimated along the MCMC chain using Geneland in the single best run (lowest maximum likelihood), for 21 *Parnassius smintheus* populations across six SNP datasets (shown on x-axis). Boxes show central 50% of values and the median, for the single best of 10 runs. Tails represent values within 1.5 times the interquartile range, and points represent values outside 1.5 times the interquartile range.

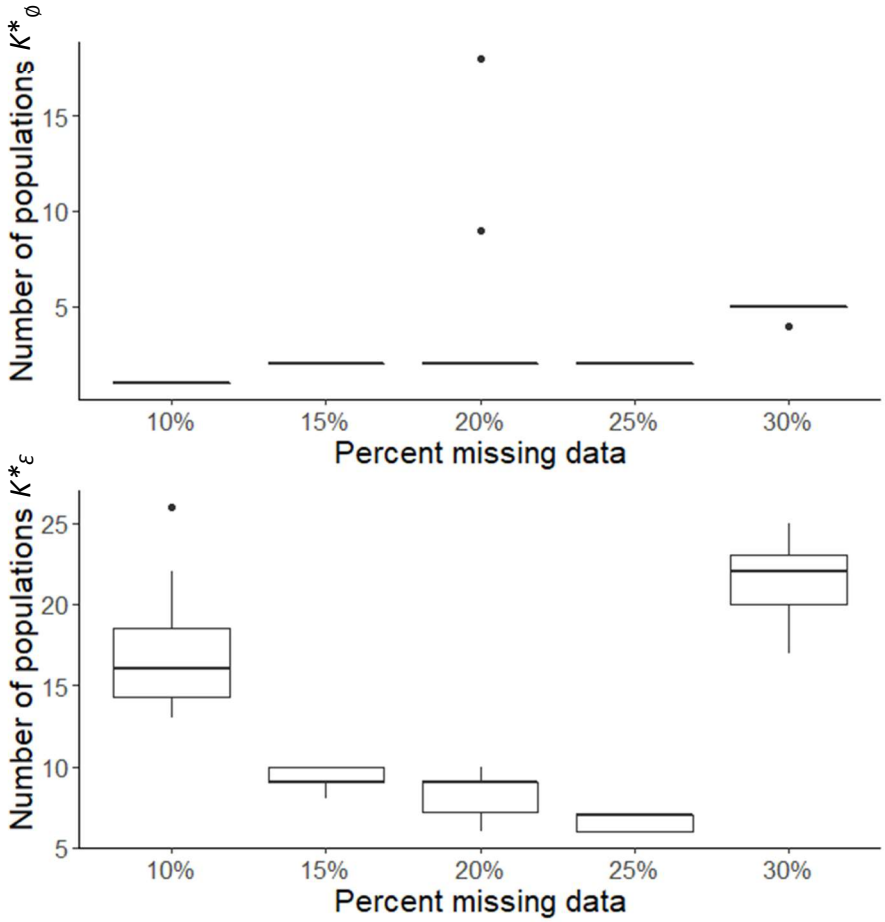

Figure 2.5 Boxplots of a) the K^*_{ε} estimator in fastSTRUCTURE and b) the $K^*_{\emptyset C}$ estimator in fastSTRUCTURE, for 21 *Parnassius smintheus* populations across six SNP datasets (shown on x-axis). Both estimators were generated using the chooseK.py function in fastSTRUCTURE. Boxes show central 50% of values and the median, across 10 runs. Tails represent values within 1.5 times the interquartile range, and points represent values outside 1.5 times the interquartile range.

2.4 Discussion

2.4.1 Genetic diversity

Two measures of genetic diversity, allelic richness and expected heterozygosity, were the most sensitive to the level of missing data, with both increasing as more missing data was permitted. However, observed heterozygosity did not show the same pattern, and did not differ among SNP datasets. These trends were driven by increases in the minor allele frequency in datasets with more missing data, which caused both allelic richness and expected heterozygosity to increase. Importantly, the increased minor allele frequency was driven primarily by having more minor allele homozygotes in the datasets with more missing data; however, observed heterozygosity did not increase because the overall proportion of heterozygotes did not significantly change. As the frequency of minor allele homozygotes increased, without additional observed heterozygotes, expected and observed heterozygosity diverged increasingly with the amount of missing data.

Loci with higher proportions of missing data (i.e., more genotyping failures) therefore tended to have higher minor allele frequencies, and fewer heterozygotes than expected given their allele frequencies. It appears that as the amount of permitted missing data increases, heterozygote genotypes either fail to be called altogether, or may be incorrectly genotyped as homozygotes (i.e., only one of the two alleles is called). Heterozygote genotypes are more difficult to call correctly from RADseq data than are homozygotes. When processing RADseq data, multiple reads of a given sequenced fragment are identified and combined to identify alleles (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013); setting a minimum stack depth (number of identical reads necessary to designate an allele) prevents spurious alleles and loci from being called (Catchen et al., 2013; Mastretta-Yanes et al., 2015). However, for a heterozygote to be genotyped correctly, both alleles must meet the minimum stack depth to be identified and assigned to the same locus. This makes it possible for heterozygotes to be erroneously genotyped as homozygotes if one allele is not sequenced in enough copies to be called, or not to be genotyped at all if sequences of both alleles fall short of the cut-off to be called.

2.4.2 Genetic differentiation and isolation by distance

Estimates of genetic differentiation, both globally and between population pairs, were mostly robust across datasets, despite F_{ST} being calculated based on expected heterozygosity, which had more significant differences across datasets. Shafer et al (2017) found a similar pattern, with pairwise F_{ST} and observed heterozygosity being robust across three datasets, while nucleotide diversity (as with allelic richness and minor allele frequency here) was lower in the dataset with the least missing data. Wright's F_{ST} essentially describes the reduction in mean expected heterozygosity in a group of isolated or semi-isolated populations compared to if all individuals in those populations were able to mate randomly (Wright, 1969); F_{ST} is higher with greater differences between the expected heterozygosity of the total population and the mean expected heterozygosity of the subpopulations. I found that F_{ST} did not change significantly across most SNP datasets; this indicates that the ratio of mean expected heterozygosity of individual populations to the expected heterozygosity for the entire set of individuals is mostly consistent across the datasets. As a result, patterns of IBD among populations also remained consistent, as neither geographic nor genetic (i.e., pairwise F_{ST}) distances differed significantly among most SNP datasets.

2.4.3 Population assignment

The accuracy of population assignment can be assessed by two criteria: whether the number of populations identified is close to the number of known populations sampled, and whether individuals are being assigned to a reasonable population (i.e., their sample site, or a site within dispersal distance). For both criteria, I found that population assignment was broadly robust to missing data, with the clear exception of the smallest, most stringent dataset (10% missing data; 39 SNPs). At 10% missing data, individuals were either assigned to many fewer populations than were in the sample (and were identified using the larger SNP datasets) or were assigned to a larger number of populations that often did not correspond to the spatial locations of sample sites. Datasets with more permissive thresholds for missing data did not introduce noise and reduce the accuracy of population assignment; in some cases (here, using fastSTRUCTURE to identify populations that contribute weakly to structure) the most permissive 30%

missing data dataset was closest at identifying the expected number of populations in the sample. Larger SNP datasets (independent of the level of missing data) typically provide greater accuracy for both the number of populations identified and in assigning individuals to their source population (Guillot & Santos, 2010). As discussed above, the SNP loci with higher amounts of missing data are also more likely to have genotyping errors (specifically, misgenotyping heterozygotes as homozygotes). However, these added loci also have higher minor allele frequencies. Rare alleles (i.e., at loci with low minor allele frequency) are often considered highly informative and important to include in population genetic studies (Marandel et al., 2020).

However, loci with low minor allele frequencies can also add noise to analyses; for example, using loci with higher minor allele frequencies better reflected population structure of the golden-crowned kinglet (*Regulus satrapa*) than when including loci with lower minor allele frequencies (Linck & Battey, 2019). This may be because for loci with very low minor allele frequencies, very few individuals will be heterozygous or homozygous at the minor allele. As a result, there is a greater impact of sampling error and the true allele frequencies are harder to estimate accurately. With respect to estimating population differentiation, even if the minor allele frequency is similar across all sampled populations, the likelihood of sampling an individual carrying a copy of the minor allele will be low in all populations; on the rare occasion that such an individual is sampled, its population will erroneously appear more genetically distinct than is truly the case. Overall, the positive impact of including loci with higher minor allele frequencies may balance out the negative impact of higher genotyping errors in more permissive datasets, allowing the additional SNPs in datasets with more missing data to increase the accuracy of population assignment.

2.4.4 The usefulness of small SNP datasets

For many population genetic analyses, datasets as small as approximately 350 to 1000 SNPs (with 15-20% permitted missing data) were sufficient to infer population genetic patterns. While these datasets may appear small compared to the numbers of SNPs that are often used in population genetic studies using RADseq (i.e., in the 1000s to tens of thousands, Puckett, 2017), it is congruent with the number of SNPs sufficient to assess

genetic diversity, differentiation, and population structure in other studies. Genetic differentiation (i.e., F_{ST}) can be assessed using as few as 20 to 75 SNPs (depending on the degree of differentiation), when sample sizes are high (i.e., more than 40 individuals per population, Morin, Martien, & Taylor, 2009). In a set of five simulated populations, 1000 SNPs with less than 10% missing data were always sufficient to detect population structure (using STRUCTURE) even when populations had only recently diverged; with more differentiated populations, population structure was detectable with fewer than 200 SNPs (Haasl & Payseur, 2011). For AFLPs, a similar biallelic (although dominant) molecular marker, the number of loci required for population assignment (using the methods of Paetkau, Calvert, Stirling, & Strobeck, 1995) ranged from 50 to 400, depending on the degree of differentiation and the number of populations considered (Campbell, Duchesne, & Bernatchez, 2003).

With the many variations of RADseq available, several of which result in even more SNPs being called than ddRADseq, it is common to see SNP dataset sizes in the thousands to tens of thousands used in population genetic studies (Puckett, 2017). However, limitations to the amount and quality of DNA that can be extracted from some sampled tissues (particularly when samples are highly degraded, or are contaminated) can limit the success of these techniques (Blair, Campbell, & Yoder, 2015; Graham et al., 2015; Hart, Meyer, Johnson, & Ericsson, 2015). RADseq and related techniques amplify and sequence all DNA fragments in a sample; low starting concentrations or degradation of the target DNA increases the chances of amplifying background DNA and identifying spurious SNPs (Leese et al., 2012). Calling spurious SNPs is especially a problem in species without a reference genome, which would otherwise allow loci that do not align to the reference genome to be filtered out (Leese et al., 2012). One option for researchers in this situation is to use RADseq, or related methods, to identify SNPs in a small number of high quality samples, then use more direct methods of SNP genotyping that are robust to low DNA quantity and quality to genotype lower quality samples at a subset of identified loci (e.g., Norman, Street, & Spong, 2013; Siccha-Ramirez et al., 2018; Tysklind et al., 2019). In the Western rattlesnake *Crotalus oreganus*, a small number of SNPs (362) genotyped using the Genotyping-in-Thousands by sequencing method provided comparable results to a larger RADseq dataset (8281 SNPs) in estimating

genetic differentiation and population assignment, and allowed non-invasive clocal swabs rather than blood samples as a source of DNA (Schmidt, Campbell, Govindarajulu, Larsen, & Russello, 2020). Overall, when RADseq is difficult to apply due to sample quality problems, or when large SNP datasets are too computationally demanding, trade-offs in the quality, informativeness and number of loci appear such that SNP datasets as small as approximately 350 SNPs are sufficient.

2.4.5 Recommendations for filtering SNP datasets

I found that the optimal level of missing data depended on the intended use of the dataset. For some analyses, including assessing F_{ST} and IBD, the missing data threshold does not seem to be an important consideration. For others, especially the number of populations identified by fastSTRUCTURE and Geneland population assignment, a more permissive threshold and a resulting larger SNP dataset allows more populations to be identified. However, very permissive missing data thresholds result in very large SNP datasets that are difficult to work with. Here, the largest SNP dataset (~12 000 SNPs) could be used for some analyses (e.g., estimating genetic diversity and differentiation) but not for population assignment, as a result of computational and time constraints. Given that for many analyses continuing to add SNPs past a moderately sized dataset does not affect the interpretation of results (and can be computationally intensive), a moderately sized dataset (~350-1000 SNPs), generated with an intermediate level of permitted missing data (15-20%), is appropriate for most analyses.

2.5 Literature cited

Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, *22*(11), 3179–3190.

Blair, C., Campbell, C. R., & Yoder, A. D. (2015). Assessing the utility of whole genome amplified DNA for next-generation molecular ecology. *Molecular Ecology Resources*, *15*(5), 1079–1090.

Bretz, F., Hothorn, T., & Westfall, P. (2010). Multiple comparisons using R. New York: Chapman and Hall/CRC.

Callen, D. F., Thompson, A. D., Shen, Y., Phillips, H. A., Richards, R. I., Mulley, J. C., & Sutherland, G. R. (1993). Incidence and origin of "null" alleles in the (AC)n microsatellite markers. *American Journal of Human Genetics*, *52*(5), 922–927.

Campbell, D., Duchesne, P., & Bernatchez, L. (2003). AFLP utility for population assignment studies: Analytical investigation and empirical comparison with microsatellites. *Molecular Ecology*, *12*(7), 1979–1991.

Caplins, S. A., Gilbert, K. J., Ciotir, C., Roland, J., Matter, S. F., & Keyghobadi, N. (2014). Landscape structure and the genetic effects of a population collapse. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1796), 20141798.

Casillas, S., & Barbadilla, A. (2017). Molecular population genetics. *Genetics*, *205*(3), 1003–1035.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140.

Chattopadhyay, B., Garg, K. M., & Ramakrishnan, U. (2014). Effect of diversity and missing data on genetic assignment with RAD-Seq markers. *BMC Research Notes*, *7*, 841.

Clarke, R. T., Rothery, P., & Raybould, A. F. (2002). Confidence limits for regression relationships between distance matrices: Estimating gene flow with distance. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*(3), 361.

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: Next-generation population genetics. *Briefings in Functional Genomics*, *9*(5–6), 416–423.

Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., & Blaxter, M. L. (2013). Special features of RAD Sequencing data: Implications for genotyping. *Molecular Ecology*, *22*(11), 3151–3164.

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., … Estoup, A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, *22*(11), 3165–3178.

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, *5*(1), 184–186.

Graham, C. F., Glenn, T. C., McArthur, A. G., Boreham, D. R., Kieran, T., Lance, S., … Somers, C. M. (2015). Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Molecular Ecology Resources*, *15*(6), 1304–1315.

Guillot, G., Mortier, F., & Estoup, A. (2005). Geneland: A computer package for landscape genetics. *Molecular Ecology Notes*, *5*(3), 712–715.

Guppy, C. S., & Shepard, J. (2001). *Butterflies of British Columbia including western Alberta, southern Yukon, the Alaska Panhandle, Washington, northern Oregon, northern Idaho, northwestern Montana*. Vancouver [B.C.]: UBC Press.

Haasl, R. J., & Payseur, B. A. (2011). Multi-locus inference of population structure: A comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, *106*(1), 158–171.

Hart, M. L., Meyer, A., Johnson, P. J., & Ericsson, A. C. (2015). Comparative evaluation of DNA extraction methods from feces of multiple host species for downstream next-generation sequencing. *PLOS ONE*, *10*(11), e0143334.

Hodel, R. G. J., Chen, S., Payton, A. C., McDaniel, S. F., Soltis, P., & Soltis, D. E. (2017). Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: Comparing microsatellites and RAD-Seq and investigating loci filtering. *Scientific Reports*, *7*(1), 17598.

Huang, H., & Knowles, L. L. (2016). Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of RAD sequences. *Systematic Biology*, *65*(3), 357–365.

Jangjoo, M., Matter, S. F., Roland, J., & Keyghobadi, N. (2016). Connectivity rescues genetic diversity after a demographic bottleneck in a butterfly population network. *Proceedings of the National Academy of Sciences*, *113*(39), 10914–10919.

Jeffries, D. L., Copp, G. H., Handley, L. L., Olsén, K. H., Sayer, C. D., & Hänfling, B. (2016). Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. *Molecular Ecology*, *25*(13), 2997–3018.

Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405.

Keyghobadi, N., Roland, J., & Strobeck, C. (1999). Influence of landscape on the population genetic structure of the alpine butterfly *Parnassius smintheus* (Papilionidae). *Molecular Ecology*, *8*(9), 1481–1495.

Keyghobadi, N., Roland, J., & Strobeck, C. (2005). Genetic differentiation and gene flow among populations of the alpine butterfly, *Parnassius smintheus*, vary with landscape connectivity. *Molecular Ecology*, *14*(7), 1897–1909.

Leese, F., Brand, P., Rozenberg, A., Mayer, C., Agrawal, S., Dambach, J., … Sands, C. J. (2012). Exploring Pandora's box: Potential and pitfalls of low coverage genome surveys for evolutionary biology. *PLOS ONE*, *7*(11), e49202.

Lemopoulos, A., Prokkola, J. M., Uusi-Heikkilä, S., Vasemägi, A., Huusko, A., Hyvärinen, P., … Vainikka, A. (2019). Comparing RADseq and microsatellites for estimating genetic diversity and relatedness—Implications for brown trout conservation. *Ecology and Evolution*, *9*(4), 2106–2120.

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33.

Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, *19*(3), 639–647.

Luca, F., Hudson, R. R., Witonsky, D. B., & Rienzo, A. D. (2011). A reduced representation approach to population genetic analyses and applications to human evolution. *Genome Research*, *21*(7), 1087–1098.

Marandel, F., Charrier, G., Lamy, J.-B., Cam, S. L., Lorance, P., & Trenkel, V. M. (2020). Estimating effective population size using RADseq: Effects of SNP selection and sample size. *Ecology and Evolution*, *10*(4), 1929–1937.

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Molecular Ecology Resources*, *15*(1), 28–41.

Morin, P. A., Martien, K. K., & Taylor, B. L. (2009). Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, *9*(1), 66–73.

Norman, A. J., Street, N. R., & Spong, G. (2013). *De novo* SNP discovery in the Scandinavian brown bear (*Ursus arctos*). *PLOS ONE*, *8*(11), e81012.

Paetkau, D., Calvert, W., Stirling, I., & Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, *4*(3), 347–354.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5), e37135.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & Deepayan, S. (2015). *nlme: Linear and nonlinear mixed effects models*.

Pool, J. E., Hellmann, I., Jensen, J. D., & Nielsen, R. (2010). Population genetic inference from genomic sequence variation. *Genome Research*, *20*(3), 291–300.

Pope, N. (2020). *CorMLPE: A correlation structure for symmetric relational data*.

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

Puckett, E. E. (2017). Variability in total project and per sample genotyping costs under varying study designs including with microsatellites or SNPs to answer conservation genetic questions. *Conservation Genetics Resources*, *9*(2), 289–304.

R Core Team. (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573–589.

Rašić, G., Filipović, I., Weeks, A. R., & Hoffmann, A. A. (2014). Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, *Aedes aegypti*. *BMC Genomics*, *15*.

Rokas, A., & Abbot, P. (2009). Harnessing genomics for evolutionary insights. *Trends in Ecology & Evolution*, *24*(4), 192–200.

Roland, J., Keyghobadi, N., & Fownes, S. (2000). Alpine parnassius butterfly dispersal: Effects of landscape and population size. *Ecology*, *81*(6), 1642–1653.

Schmidt, D. A., Campbell, N. R., Govindarajulu, P., Larsen, K. W., & Russello, M. A. (2020). Genotyping-in-Thousands by sequencing (GT-seq) panel development and application to minimally invasive DNA samples to support studies in molecular ecology. *Molecular Ecology Resources*, *20*(1), 114–124.

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, *8*(8), 907–917.

Siccha-Ramirez, Z. R., Maroso, F., Pardo, B. G., Fernández, C., Martínez, P., & Oliveira, C. (2018). SNP identification and validation on genomic DNA for studying genetic diversity in *Thunnus albacares* and *Scomberomorus brasiliensis* by combining

RADseq and long read high throughput sequencing. *Fisheries Research*, *198*, 189–194.

Song, B., Woerner, A. E., & Planz, J. (2021). mixIndependR: A R package for statistical independence testing of loci in database of multi-locus genotypes. *BMC Bioinformatics*, *22*(1), 12.

Sunde, J., Yıldırım, Y., Tibblin, P., & Forsman, A. (2020). Comparing the performance of microsatellites and RADseq in population genetic studies: Analysis of data for pike (*Esox lucius*) and a synthesis of previous studies. *Frontiers in Genetics*, *11*, 218.

Tysklind, N., Blanc-Jolivet, C., Mader, M., Meyer-Sand, B. R. V., Paredes-Villanueva, K., Honorio Coronado, E. N., … Degen, B. (2019). Development of nuclear and plastid SNP and INDEL markers for population genetic studies and timber traceability of Carapa species. *Conservation Genetics Resources*, *11*(3), 337–339.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358–1370. JSTOR.

Wright, S. (1969). *Evolution and genetics of populations. Volume 2. The theory of gene frequencies*. Chicago: University of Chicago Press.

Chapter 3

3 Minimum sample sizes for population genetic analyses when using RADseq-generated SNP datasets

3.1 Introduction

Estimating genetic diversity and differentiation is core to empirical population genetic studies and provides insight into many processes with population-level consequences, such as dispersal, changes in population size, and selection (Allendorf, 1986; Kreitman, 2000; Slatkin, 1987). When designing population genetics studies, there is a trade-off between the number of individuals sampled per population and the number of populations surveyed. For a fixed cost, many individuals from few populations, few individuals from many populations, or an intermediate number of both individuals and populations may be studied. Both population number and sample size per population are important aspects of sampling design (Aguirre-Liguori, Luna-Sánchez, Gasca-Pineda, & Eguiarte, 2020; Hale, Burg, & Steeves, 2012; Morin, Martien, & Taylor, 2009). Achieving a minimum sample size per population is important to accurately reflect measures of genetic diversity and differentiation (e.g., allelic richness and F_{ST}: Morin et al., 2009; Nazareno, Bemmels, Dick, & Lohmann, 2017). The number of populations sampled is also important however, as sampling multiple populations distributed across the landscape minimizes the possibility of spurious inferences (Anderson et al., 2010; Beerli, 2004; Oyler-McCance, Fedy, & Landguth, 2013). In addition to increasing the total number of different populations that can be surveyed and minimizing genotyping costs, there are other compelling reasons to sample the minimum necessary number of individuals per population to accurately reflect the genetic structure. This includes when using lethal sampling or sampling that otherwise decreases fitness, as well as when sampling is difficult or time consuming, such as when sampling cryptic species or in locations that are difficult to access.

The minimum number of individuals per population required to accurately reflect commonly used population genetic metrics varies depending on the type and number of molecular marker used, the metric in question, and idiosyncrasies of the study system

(Flesch, Rotella, Thomson, Graves, & Garrott, 2018; Sunde, Yıldırım, Tibblin, & Forsman, 2020). Increasing the total information provided per individual by changing the type or number of molecular markers used is one potential approach to offset lower sample sizes. For microsatellites, both the number and the variability of loci (i.e., the number of alleles per locus) are important. For a given sample size, increasing the number of independent microsatellite loci increases the power to detect historical bottlenecks (Hoban, Gaggiotti, & Bertorelle, 2013), and both microsatellite number and variability contribute to the power to detect isolation by distance (IBD) and isolation by resistance (Landguth et al., 2012). Increasing the number of microsatellite loci used can also compensate for sampling fewer individuals per population; adding microsatellite markers was more efficient than adding additional individuals when testing for historical bottlenecks (Hoban et al., 2013). When using STRUCTURE to estimate population number, increasing either the number of microsatellites used or the number of individuals sampled allowed the known total number of populations to be detected (Evanno, Regnaut, & Goudet, 2005).

For microsatellites, a sample size of 20-30 individuals per population is commonly recommended for calculating parameters such as allelic richness, expected heterozygosity, and F_{ST} (Hale, Burg, & Steeves, 2012; Pruett & Winker, 2008). However, as single nucleotide polymorphisms (SNPs) have become more commonly used in population genetic studies, this broadly recommended sample size has been called into question (Willing, Dreyer, & Oosterhout, 2012). At the level of individual loci, SNPs are less informative than microsatellites because they are biallelic (while microsatellites are typically multiallelic) and therefore more SNP loci are necessary to achieve a similar power to detect genetic patterns. For small SNP panels, the power to detect genetic differentiation is determined both by SNP number and number of individuals sampled. When simulating populations with different levels of genetic differentiation, sampling more than 40 individuals per population greatly increased the power for all sizes of SNP panel, while for the smallest SNP panel (20 SNPs) the power to detect differentiation never approached that of the larger panels (50 and 75 SNPs) even at the largest sample sizes (Morin, Martien, & Taylor, 2009). With the development of next-generation sequencing techniques it is now routine to generate very large SNP datasets (i.e., with

thousands of loci) for costs comparable to genotyping a microsatellite panel, although there is debate as to whether the per sample cost is in fact greater for SNPs (Puckett, 2017) or microsatellites (Kraus et al., 2015). Extending the trend observed by Morin et al. (2009), these very large SNP panels were predicted to allow for even fewer individuals to be sampled per population while retaining the power to detect genetic patterns (Willing et al., 2012). This prediction had held true for some study systems and some analyses; for example, sample sizes of four to six (Willing, Dreyer, & Oosterhout, 2012) or two (Nazareno, Bemmels, Dick, & Lohmann, 2017) have been shown to be sufficient to accurately estimate F_{ST}, and sample sizes of 6-8 were sufficient to accurately estimate heterozygosity and the number of effective alleles (Nazareno et al., 2017). However, these studies were limited to examining variation within and between only two populations. In a slightly larger study of four bighorn sheep populations, using approximately 14 000 SNPs, a minimum sample size of 25 was recommended for both kinship analyses and estimates of F_{ST} (Flesch, Rotella, Thomson, Graves, & Garrott, 2018). The effect of sample size on population clustering when using SNPs is even less clear than for analyses of diversity and differentiation, and has not been addressed using some of the most commonly applied genetic clustering and assignment approaches such as STRUCTURE (Pritchard, Stephens, & Donnelly, 2000) and Geneland (Guillot, Mortier, & Estoup, 2005). Where population clustering and assignment has been examined using an alternative Bayesian approach (Paetkau, Slade, Burden, & Estoup, 2004), the success of population assignment continued to increase with sample size up to the maximum sample of 34 individuals (Benestan et al., 2015).

Here, I use populations of the alpine butterfly *Parnassius smintheus* to explore how sample size affects estimation of genetic diversity, differentiation, and population clustering when using SNPs as a molecular marker. These populations, located in western Alberta, are among a larger set of populations that have been previously characterized using microsatellites (Keyghobadi, Roland, & Strobeck, 2005). In the larger study, populations were located in three distinct geographic regions, each displaying different patterns of genetic differentiation; here, the populations I examine are drawn from two of the geographic regions (Banff and West Kananaskis). Populations in these regions had lower genetic differentiation (global F_{ST}) and significant IBD (i.e., a correlation between

pairwise genetic and geographic distances) compared to populations from the third region (East Kananaskis). *Parnassius smintheus* has also been studied extensively using mark-release-recapture methods (Keyghobadi, Roland, & Strobeck, 1999; Matter, Keyghobadi, & Roland, 2014; Roland, Keyghobadi, & Fownes, 2000). The maximum observed dispersal distance of a *P. smintheus* individual was approximately 1.7 km (Roland et al., 2000); given that *P. smintheus* inhabits distinct alpine meadows and the closest populations in my dataset are separated by approximately 6 km, it is unlikely that there is extensive dispersal and gene flow among these populations.

These characteristics – low but detectable genetic differentiation, the presence of IBD, and the availability of relatively large samples from multiple distinct populations – provide an opportunity to test the effect of sample size in an empirical system in which considerable demographic and genetic data are already available. I use a reduced representation sequencing approach – double digest restriction site associated DNA sequencing (ddRADseq; Peterson, Weber, Kay, Fisher, & Hoekstra, 2012) – to genotype *P. smintheus* individuals at hundreds to thousands of SNP loci. I tested the hypothesis that sampling a larger portion of the genome compensates to an extent for sampling a smaller proportion of the population, such that SNP datasets with several hundreds or thousands of loci (e.g., generated through ddRADseq) allow accurate inferences to be made with relatively small numbers of sampled individuals (Willing et al, 2012; Nazareno et al., 2017). I defined the minimum sample size as the lowest sample size that produces results similar to those from the complete dataset for a given analysis. Minimum sample size depends not only on the number of genetic markers in a dataset (Morin, Martien, & Taylor, 2009), but also on the type of analysis; estimation of expected heterozygosity and F_{ST} (Willing et al, 2012; Nazareno et al., 2017) require fewer individuals than population clustering analyses (Benestan et al., 2015). I therefore also hypothesized that analyses where fewer parameters are estimated (e.g., expected heterozygosity, global F_{ST}) require fewer individuals to be sampled to make accurate inferences, as compared to analyses where many parameters are estimated (e.g., population assignment).

I predicted that as more individuals are sampled, the results of common population genetic analyses (including estimation of allelic richness, expected heterozygosity, F_{ST}, IBD, and population assignment) would approach that of the complete dataset. I predicted that for SNP datasets with hundreds or thousands of loci, the minimum sample size would be lower than that commonly recommended for microsatellite datasets (i.e., fewer than 20 individuals). I also predicted that there would be an interaction between SNP dataset size and minimum sample size, where the minimum sample size required for a given analysis would decrease when using larger SNP datasets. Finally, I predicted that the minimum sample size would be lower for estimating basic parameters of diversity and differentiation (i.e., global F_{ST}, allelic richness, expected heterozygosity) than for more complex analyses of population structure (IBD, population clustering).

3.2 Methods

3.2.1 Data collection

Twenty seven alpine meadows with populations of *Parnassius smintheus* were initially sampled in 1995 - 1999; individuals were captured using hand nets and stored in glassine envelopes at -80°C (Keyghobadi, Roland, & Strobeck, 2005). Of these, seven sampling sites had greater than 35 remaining voucher samples available for DNA extraction and genotyping. For these seven sites, I extracted DNA from the head and thorax of either 40 individuals or the maximum number of available samples (Table 3.1) using a DNeasy Blood and Tissue kit (Qiagen, Germantown, MD). I used this extracted DNA to prepare a ddRADseq library. I digested DNA with the restriction enzymes *Nla*III and *Eco*RI, labelled with adapters as per Râsíc et al (2014), and selected for fragments between 200 and 500 bp in length using Sera-Mag solid-phase reversible immobilization beads (SPRI; GE Healthcare Life Sciences, Chicago, IL). After PCR amplification, I sent libraries for sequencing on an Illumina HiSeq 2500 platform.

I assembled reads *de novo* using the Stacks pipeline (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013). I set the assembly parameters as: a minimum stack depth of 3 (m), a maximum difference of 3 nucleotides among reads per stack (n), a maximum difference of 2 nucleotides when combining stacks (M), and a maximum difference of 4

nucleotides when adding additional reads to an assembled stack (N). I excluded SNPs with a minor allele frequency of less than 5% from analysis, and included only one SNP per restriction fragment for analysis. I excluded any individuals with fewer than 50% of loci successfully genotyped (for loci with a maximum of 20% missing data across all individuals). Genotyping succeeds when there are sufficient reads at a locus for the Stacks pipeline to call a SNP for a given individual.

3.2.2 SNP datasets and subsampling

First, I generated three SNP datasets of differing size by including loci with a maximum of 15, 20, and 30% missing data. Missing data refers to the percent genotyping success per locus; setting the maximum percent missing data to 15% means that any locus where more than 15% of all individuals failed to genotype would be excluded. Datasets with lower permitted missing data thresholds are more exclusive, and therefore include fewer SNPs. This resulted in three SNP panels with 339, 1098, and 4760 loci, respectively. For each of the three SNP panels, I first calculated all population genetic metrics for the complete dataset (i.e., using all individuals sampled for each population) as the baseline for comparison to subsampled datasets. To investigate the effects of sample size on population genetic analyses, I randomly subsampled individuals from each population.

I simulated different sample sizes by randomly subsampling a set number of individuals from each population, without replacement, and performed population genetic analysis using each of the subsampled datasets at each of the three SNP panels. I did not re-run the entire Stacks pipeline for each set of subsampled individuals, but instead included all SNPs called from each of the three original panels (i.e., derived from the full dataset of all samples individuals). I did this to control for the effect of calling SNPs from different numbers of individuals, which would have resulted in fewer SNPs called when fewer individuals were originally included. While this would be a real consequence of designing a study with a smaller individual sample size, I was interested here in isolating only the effects of having fewer individuals for the same SNP panel(s).

Table 3.1 Basic information for each of seven *Parnassius smintheus* populations (AR: allelic richness; H_E: expected heterozygosity). Populations were sampled from three regions: Banff (B), East Kananaskis (EK), and West Kananaskis.

Population	Sample size	Region	AR	H_E
Fortress Mtn	39	WK	1.80	0.17
Flint Peak	38	B	1.75	0.18
Mount Kent	36	WK	1.77	0.16
Panther Mtn	36	B	1.79	0.18
Snow Creek	32	B	1.83	0.19
Stony Creek	38	B	1.83	0.20
Wedge	36	WK	1.81	0.17

I simulated sample sizes of 2, 5, 10, 15, 20, and 25 individuals per population by randomly selecting individuals, without replacement, using the gpSampler command from the diveRsity package in the statistical software R (R Core Team, 2015). I subsampled 100 replicate groups of individuals for each sample size, resulting in 500 subsampled individual datasets. Each of these 500 individual datasets was analyzed at each of the three SNP panels, for a total of 1500 analyzed datasets.

3.2.3 Genetic diversity and differentiation

For all 1500 resulting datasets, I calculated allelic richness (using the hierfstat package, Goudet 2005), and expected and observed heterozygosity (using the adegenet package, Jombart, 2008). So that allelic richness would be comparable across datasets, I rarified to four alleles (the lowest number of sampled alleles across all datasets, corresponding to sample sizes of n=2 individuals). To assess population differentiation, I calculated global F_{ST} as described by Nei (1972) using the hierfstat package (Goudet, 2005). To assess IBD I used two common approaches: Mantel tests (Mantel, 1967; Smouse, Long, & Sokal, 1986), and maximum likelihood population effects (MLPE) mixed models (Clarke, Rothery, & Raybould, 2002). I calculated Mantel's r (the correlation coefficient between pairwise distances, Mantel, 1967) as the correlation between Nei's pairwise F_{ST} (Nei, 1973), estimated using the adegenet package, and geographic distance, measured as the linear distance between the centroids of the meadows where each population was sampled. I also calculated the coefficient of geographic distance using mixed models in the R package nlme (Pinheiro et al., 2015), with a random effect implemented in the corMLPE package (Pope, 2018) to account for the pairwise nature of the data. I estimated coefficients using a restricted maximum likelihood approach. For both Mantel's r and MLPE model coefficients, I recorded the proportion of subsampled datasets where significant IBD ($p < 0.05$) was observed.

3.2.4 Population clustering

I used two approaches, as implemented in the softwares fastSTRUCTURE (Raj, Stephens, & Pritchard, 2014) and Geneland (Guillot et al., 2005), to test the effects of sample size on the number of populations identified. The populations sampled were

sufficiently geographically distant from one another that little or no gene flow should occur between most pairs of populations (Roland, Keyghobadi, & Fownes, 2000). Therefore, I would expect the number of populations identified to be equal to the number of populations sampled (n=7). I did not use all 1500 SNP datasets for population clustering, due to the constraints of computing time. Instead, for both approaches, and for each sample size (n=2, 5, 10, 15, 20, 25), and missing data combination, I selected 10 replicate subsampled datasets (out of the 100 used to calculate population differentiation above), for a total of 210 separate analyses for each approach.

The first approach I used, implemented in fastSTRUCTURE, is a variant of the commonly used STRUCTURE program, and is designed to accommodate large SNP panels (Raj, Stephens, & Pritchard, 2014). FastSTRUCTURE assigns individuals to each of *K* populations, where a range of possible *K*'s is set by the user. I set the range of *K* to be 1 to 10 populations, as it is common practice to estimate k for several more populations than are expected to be present in the sample (in this case, seven). After running across the selected range of *K'*s, I used the chooseK.py function to calculate two metrics, K^*_{ε} and $K^*_{\emptyset C}$, which reflect the likely numbers of populations in the dataset. K^*_{ε} is the number of populations that maximizes the maximum likelihood of the model, and is typically a lower value than $K^*_{\emptyset C}$, which is the number of populations required to explain nearly all of the ancestry in the dataset. Each fastSTRUCTURE analysis was run through 10 replicates, and across these replicates I calculated the median K^*_{ε} and $K^*_{\emptyset C}$ value for each subsampled dataset.

The second approach I used was implemented in the program Geneland (Guillot, Mortier, & Estoup, 2005). Like fastSTRUCTURE, Geneland uses a Bayesian approach to population clustering. The approach differs from fastSTRUCTURE in two major ways: it directly estimates the number of populations in a sample, and it incorporates the geographic coordinates of each sample. FastSTRUCTURE, and the original STRUCTURE program, were designed originally for individual assignment to a number of populations as set by the user; in STRUCTURE if the user wanted to estimate the number of populations in the dataset, they had to use additional post hoc analyses (e.g., estimating delta *K*, Earl & vonHoldt, 2012; Evanno, Regnaut, & Goudet, 2005). The

model used by Geneland estimates the number of populations at each iteration of the model, and then assigns each individual the probability of originating from each identified population (Guillot et al., 2005). As with the fastSTRUCTURE analysis, I set the range of possible population numbers to be between 1 and 10, and ran each of the 210 subsampled datasets through 10 replicates. I allowed the model to run over 100 000 iterations, using the correlated allele frequency model. I used the post-processing tools in Geneland to calculate average posterior probability for each replicate, and kept the results from the highest probability replicate for each dataset. For each of these best replicates, I recorded the modal population number (i.e., the number of populations identified most often across all iterations of the model).

3.2.5 Comparisons across different sample sizes

Where the effects of sample size or number of loci on population genetic analyses have been examined, the main objective has been to determine the minimum number of individuals sampled (or number of SNPs used) at which the value of the metric of interest stabilizes (Willing, Dreyer, & Oosterhout, 2012). Typically, increasing the sample size changes the metric of interest (bringing it closer to the true population value), but each individual added provides diminishing returns in terms of additional information. Eventually, the metric stops changing directionally with additional sample size and reaches a plateau. Studies do not typically attempt to use statistical methods to identify the sample size at which this plateau is reached; instead, the median value of the metric (averaged across the subsampled datasets) with the interquartile range is plotted against sample size, and the sample size at which the metric appears to reach an asymptote is visually determined. I use this approach here to determine the minimum sample size required to estimate diversity, differentiation, and the number of distinct populations identified using fastSTRUCTURE and Geneland, and to compare this sample size across the three SNP panels.

3.3 Results

3.3.1 Complete dataset

Overall, results for the complete datasets were similar for all SNP panels, with the 1098 and 4760 SNP panels being the most similar (Table 3.2). Values of expected heterozygosity, allelic richness, global F_{ST}, and Mantel's r were all somewhat lower when calculated using 339 SNPs compared to 1098 or 4760 SNPs. The number of populations identified using Geneland was five when using both 339 and 1098 SNPs, and six when using 4760 SNPs. K^*_{ε} from fastSTRUCTURE was the same ($K^*_{\varepsilon} = 2$) across all SNP panels. $K^*_{\emptyset C}$ from fastSTRUCTURE was the only metric for which the value calculated using 4760 SNPs was lower ($K^*_{\emptyset C} = 2.5$) than when using either 339 or 1098 SNPs ($K^*_{\emptyset C} = 4$).

3.3.2 Genetic diversity and differentiation

Across the 100 subsampled datasets for each SNP panel and sample size combination, median allelic richness ranged from 1.31 (339 SNPs, five individuals) to 1.50 (4760 SNPs, two individuals). For all SNP panels, using sample sizes of two resulted in much higher and more variable (i.e., with a larger interquartile range) estimates of allelic richness than all other sample sizes (Figure 3.1). For two of the three SNP panels (339 and 1098 SNPs), estimates of median allelic richness for sample sizes larger than two were close to allelic richness estimated with the complete dataset. Median allelic richness estimated using the 4760 SNP panel was noticeably different than the other two panels. For this panel, all subsampled datasets resulted in higher estimates of allelic richness than the complete dataset.

Median expected heterozygosity (across all populations for each subsampled dataset) ranged from 0.18 (339 SNPs, all sample sizes) to 0.22 (4760 SNPs, all sample sizes). Median expected heterozygosity varied little across sample sizes, and was equal or very close to the expected heterozygosity of the complete dataset for all sample sizes (Figure 3.2). Larger sample sizes resulted in smaller interquartile ranges, except for the increase from 20 to 25 individuals where the interquartile range remained consistent.

Median global F_{ST} ranged from 0.061 (339 SNPs, 25 individuals) to 0.078 (4760 SNPs, two individuals). For all SNP datasets (339, 1098, and 4760 SNPs), median global F_{ST} for sample sizes above two was close to global F_{ST} calculated from the complete dataset (Figure 3.3). Sample sizes of two and to a lesser extent five had more outliers and larger interquartile ranges than larger sample sizes.

Median Mantel's r ranged from 0.19 (4760 SNPs, two individuals) to 0.97 (4760 and 1098 SNPs, complete dataset). Sample sizes of two, five, and ten resulted in visibly lower estimates of Mantel's r than larger sample sizes, and tended to have both larger interquartile ranges and more extreme outliers (Figure 3.4). Sample sizes of 15 and above resulted in estimates of Mantel's r that were consistently close to Mantel's r calculated from the complete dataset. For all SNP panels, Mantel's r calculated using the complete dataset was high (r = 0.92, 0.96, 0.97, for 339, 1098, and 4760 SNPs, respectively). Significant IBD was observed in all subsampled datasets when using all sample sizes higher than two individuals, for all SNP datasets (except when using 4760 SNPs and a sample size of five, where 99% of subsampled datasets had significant IBD). When using a sample size of two, 55% of subsampled datasets had significant IBD when using the 339 and 1098 SNP panels, and 27% had significant IBD when using the 4760 SNP panel.

Median coefficients estimated using MLPE mixed models ranged from 5.1E-04 (4760 SNPs, five individuals) to 7.0E-04 (1098 SNPs, two individuals). Sample sizes of two resulted in more variable coefficient estimates than larger sample sizes, with larger interquartile ranges and more extreme outliers (Figure 3.5). For all sample sizes larger than two, median coefficient estimates were very close to the MLPE coefficient calculated using the complete dataset, with little variation and very few outliers. The coefficient for the effect of geographic distance was significant for all subsampled datasets when using sample sizes higher than two. When using a sample size of two, the percentage of subsampled datasets with significant coefficients were as follows: 92% using 339 SNPs, 99% using 1098 SNPs, and 93% using 4760 SNPs.

Table 3.2 Population genetic metrics estimated using all individuals sampled from each of seven *Parnassius smintheus* populations (n=36-40 per population). Estimates were derived using each of three SNP datasets of different sizes (339, 1098, and 4760 SNPs), where SNP number was increased by changing the level of maximum missing data at each locus (15, 20 and 30%, respectively). Metrics included: expected heterozygosity (H_E) averaged over loci and populations, allelic richness (AR) averaged over loci and populations, global F_{ST}, the Mantel test coefficient (Mantel r) for the correlation between pairwise F_{ST} and geographic distance, the number of population clusters detected using Geneland, and the number of population clusters detected using the $K^*_{\mathcal{E}}$ and $K^*_{\emptyset C}$ estimators in fastSTRUCTURE.

	339 SNPs	1098 SNPs	4760 SNPs
H_E	0.18	0.19	0.22
AR	1.32	1.34	1.34
Global F_{ST}	0.061	0.064	0.064
Mantel r	0.92	0.97	0.97
Geneland	5	5	6
$K^*_{\mathcal{E}}$	2	2	2
$K^*_{\emptyset C}$	4	4	2.5

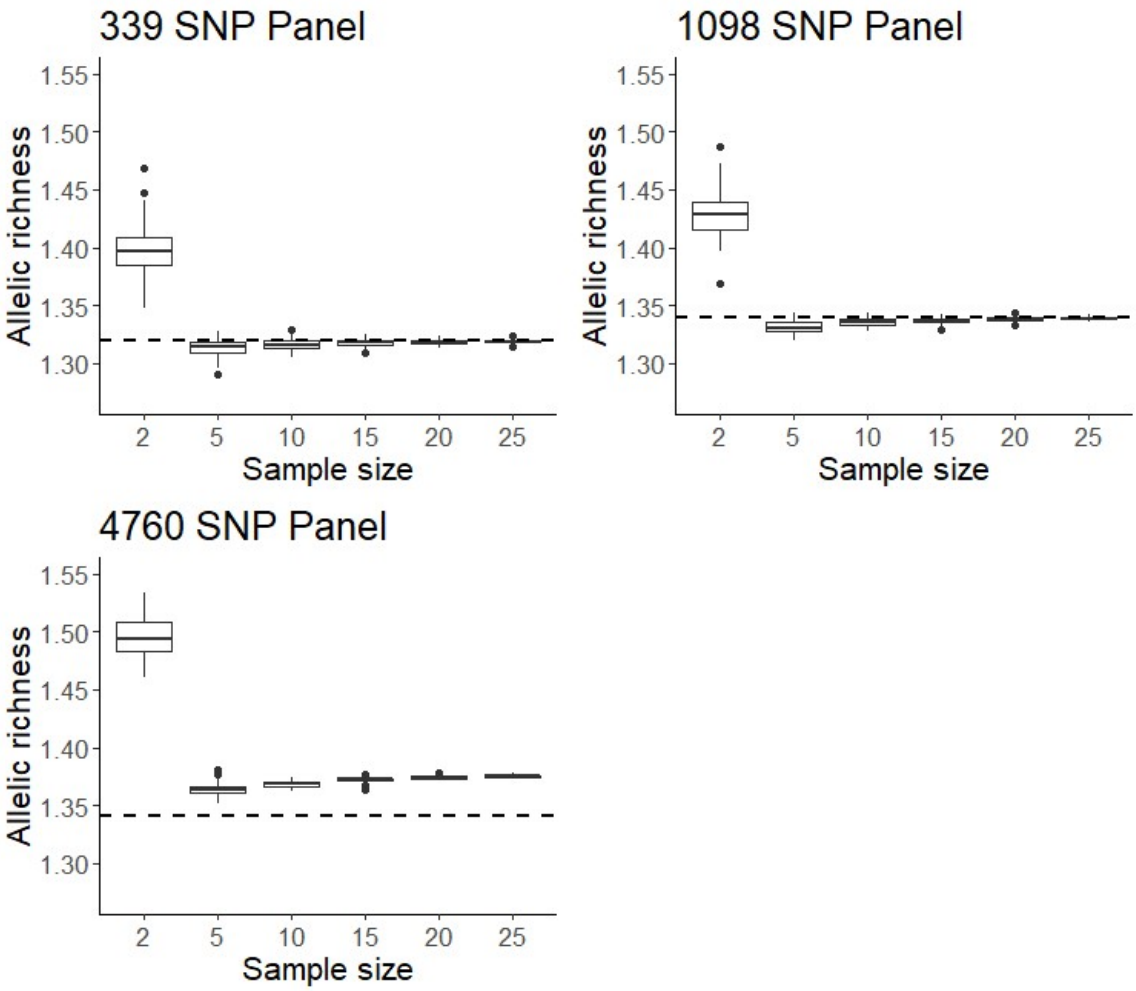

Figure 3.1 Effects of differing samples sizes (number of individuals sampled per population) on estimates of allelic richness across a set of seven *Parnassius smintheus* populations. Populations were subsampled 100 times for each sample size of between two and 25 individuals per population. Boxes show central 50% of values and the median, across subsampled datasets. Tails represent values within 1.5 times the interquartile range, and points represent values outside 1.5 times the interquartile range. Allelic richness estimated with all available individuals per population (n=36-40) is represented with a dashed line. Allelic richness was estimated using three SNP datasets of different sizes, generated with different levels of maximum missing data per locus: 339 SNPs (15% missing data), 1098 SNPs (20% missing data), and 4760 SNPs (30% missing data).

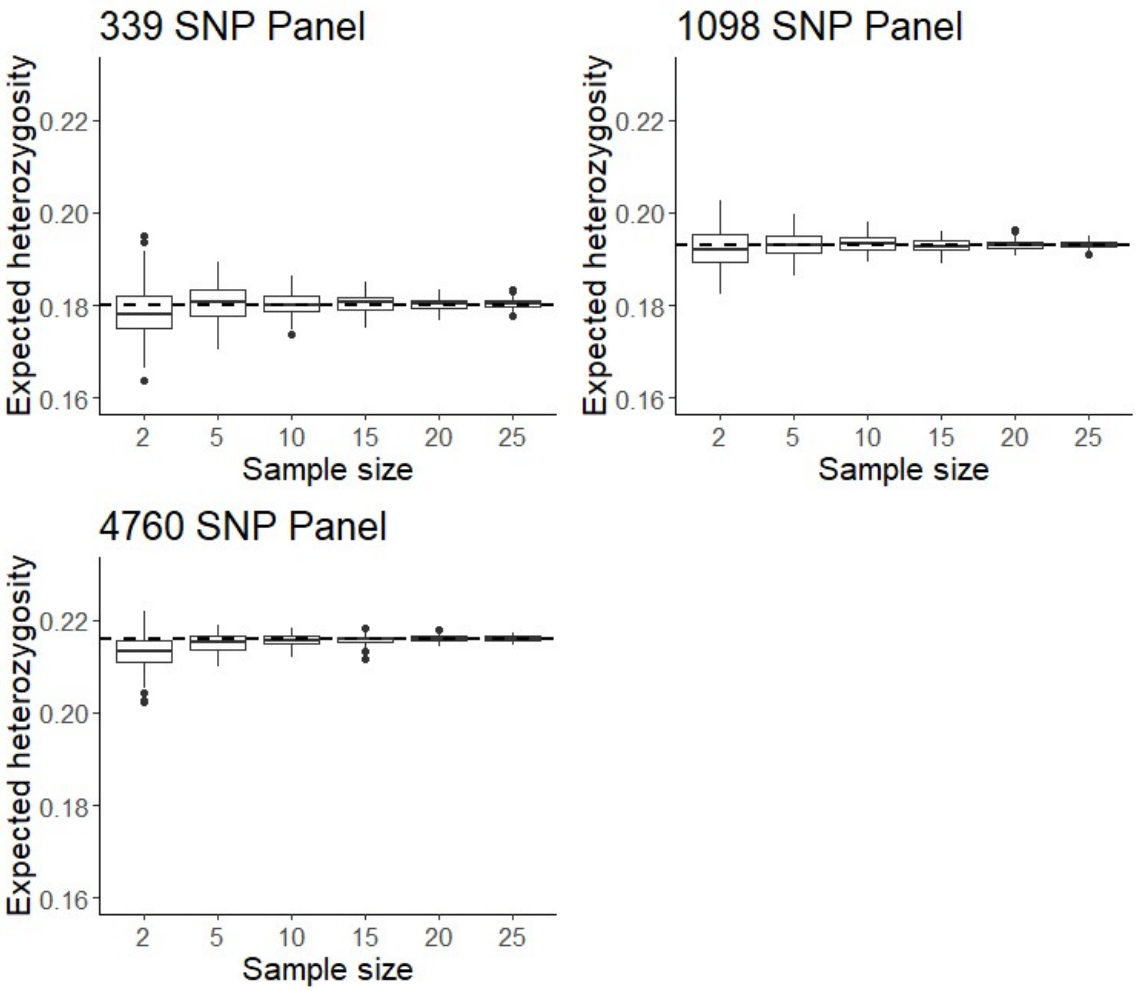

Figure 3.2 Effects of differing samples sizes (number of individuals sampled per population) on estimates of expected heterozygosity across a set of seven *Parnassius smintheus* populations. Populations were subsampled 100 times for each sample size of between two and 25 individuals per population. Boxes show central 50% of values and the median, across subsampled datasets. Tails represent values within 1.5 times the interquartile range, and points represent values outside 1.5 times the interquartile range. Expected heterozygosity estimated using all available individuals per population (n=36-40) is represented with a dashed line. Expected heterozygosity was estimated using three SNP datasets of different sizes, generated with different levels of maximum missing data per locus: 339 SNPs (15% missing data), 1098 SNPs (20% missing data), and 4760 SNPs (30% missing data).

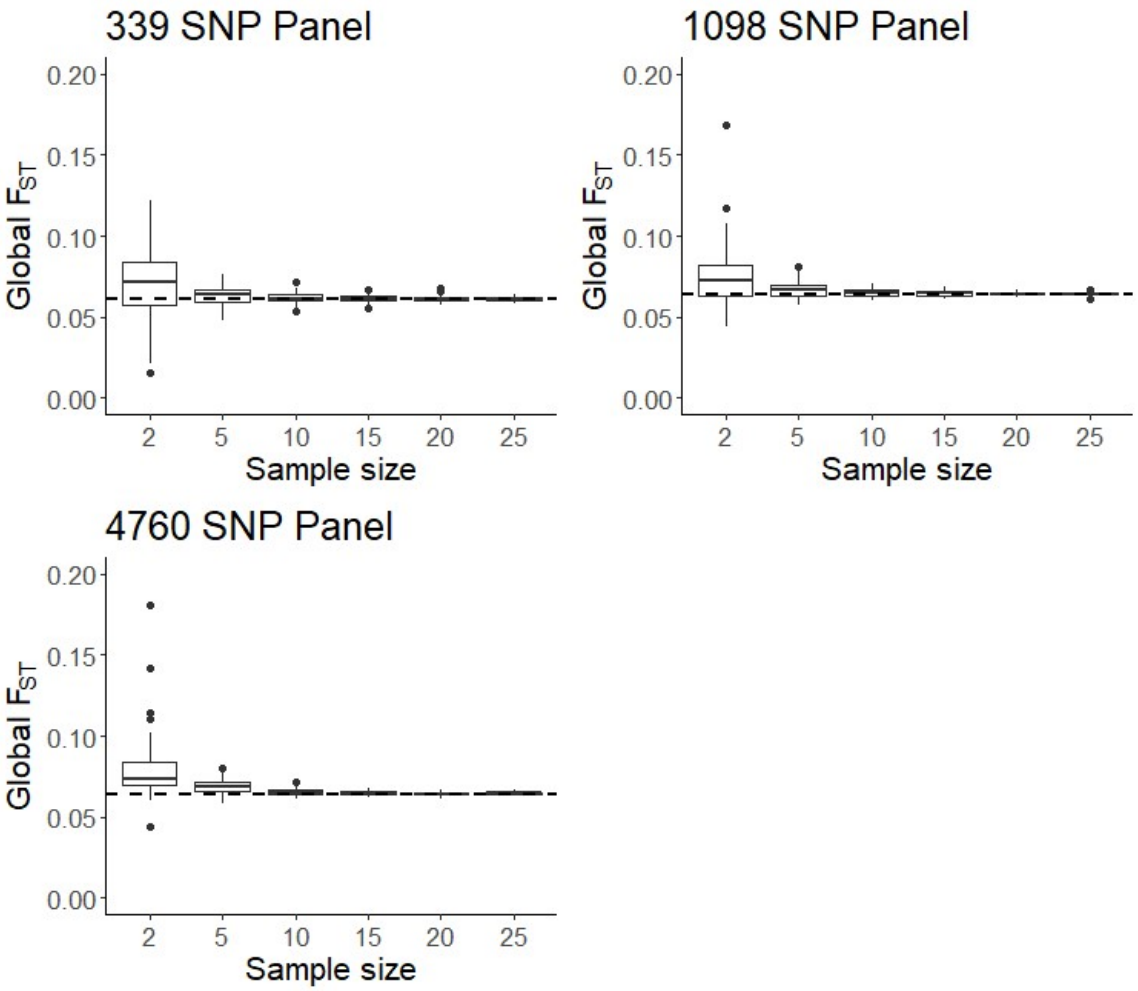

Figure 3.3 Effects of differing samples sizes (number of individuals sampled per population) on estimates of global F_{ST} among seven *Parnassius smintheus* populations. Populations were subsampled 100 times for each sample size of between two and 25 individuals per population. Boxes show central 50% of values and the median, across subsampled datasets. Tails represent values within 1.5 times the interquartile range, and points represent values outside 1.5 times the interquartile range. Global F_{ST} estimated using all available individuals per population (n=36-40) is represented with a dashed line. Global F_{ST} was estimated using three SNP datasets of different sizes, generated with different levels of maximum missing data per locus: 339 SNPs (15% missing data), 1098 SNPs (20% missing data), and 4760 SNPs (30% missing data).

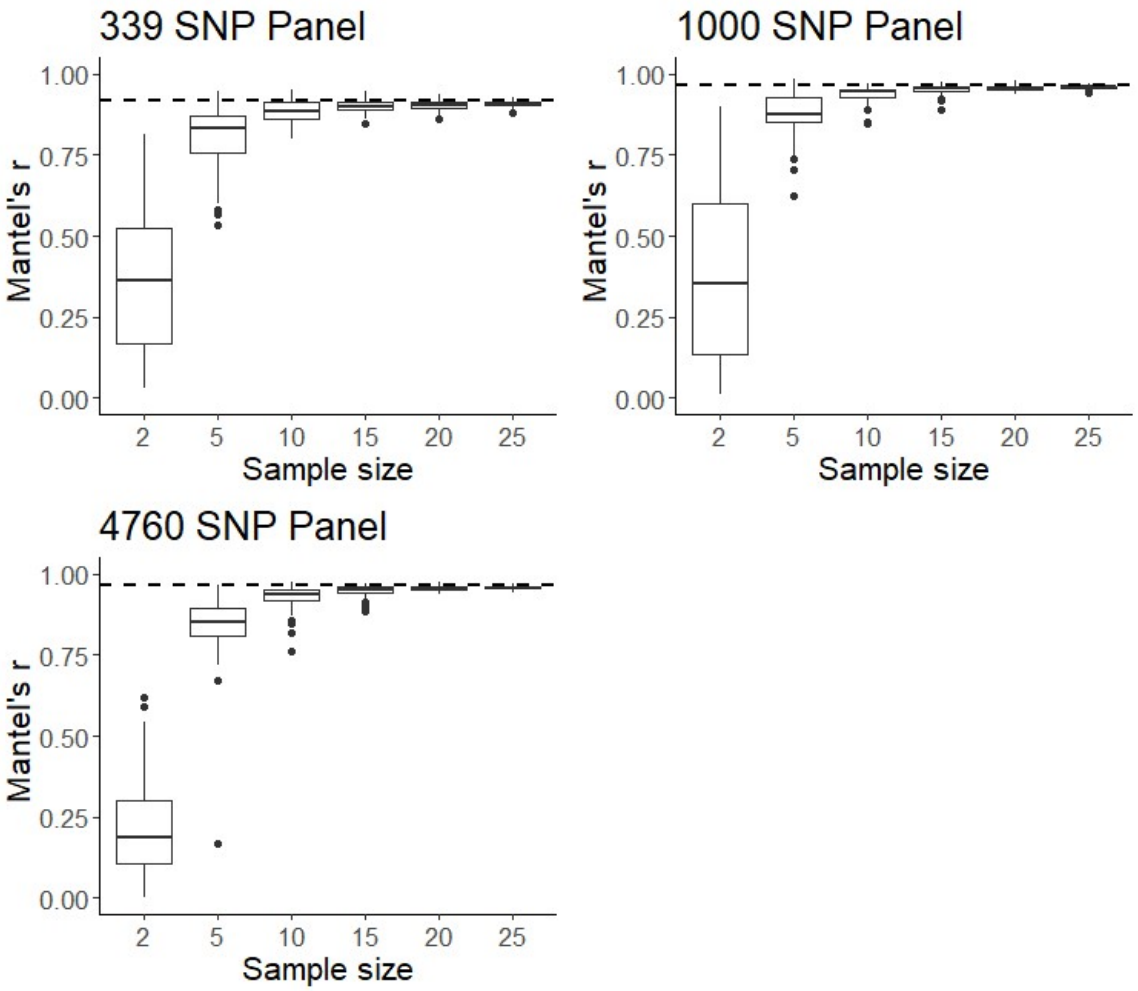

Figure 3.4 Effects of differing samples sizes (number of individuals sampled per population) on estimates of Mantel's r, for the correlation between pairwise F_{ST} and geographic distance, among seven *Parnassius smintheus* populations. Populations were subsampled 100 times for each sample size of between two and 25 individuals per population. Boxes show central 50% of values and the median, across subsampled datsets. Tails represent values within 1.5 times the interquartile range, and points represent values outside 1.5 times the interquartile range.. Mantel's r estimated using all available individuals per population (n=36-40) is represented with a dashed line. Mantel's r for the correlation between pairwise F_{ST} and geographic distance was estimated using three SNP datasets of different sizes, generated with different levels of maximum missing data per locus: 339 SNPs (15% missing data), 1098 SNPs (20% missing data), and 4760 SNPs (30% missing data).

3.3.3 Population clustering

The median number of populations identified using Geneland (for the highest likelihood run for each of ten subsampled datasets per sample size) ranged from three (4760 and 339 SNPs, five individuals) to six (4760 SNPs, complete dataset). Lower sample sizes tended to result in fewer populations being identified, although the importance of sample size varied depending on which SNP panel was used (Figure 3.6). For the 339 SNP panel, sample sizes of 15, 20 and 25 were equivalent both in their median estimate of population number and their interquartile ranges, with a median overall estimate equal to the number of populations identified in the complete dataset. For the 1098 SNP panel, sample sizes of 20 and 25 were equivalent to the complete dataset, with almost no variation (with the exception of a single outlier each) among subsampled datasets. The 4760 SNP panel produced more variable results. No subsample dataset resulted in a median number of populations identified equal to that of the complete dataset, and sample size was not predictive of interquartile range.

Median $K^*_{\mathcal{E}}$ (the number of populations identified using fastSTRUCTURE that maximized maximum likelihood) ranged from two (for almost all cases) to three (all SNP panels, two individuals). For sample sizes of ten and above for all SNP panels, there was no variation in the value of $K^*_{\mathcal{E}}$ identified across replicates (Figure 3.7). The estimated value of $K^*_{\emptyset C}$ was more variable both across different sample sizes and SNP panels. Each increase in sample size resulted in an increase in the number of populations identified up to the number of populations identified using the complete dataset (Figure 3.8). The exception to this was when using the 4760 SNP dataset, where the number of populations identified using the complete dataset was lower than when using sample sizes of 15, 20, or 25 individuals.

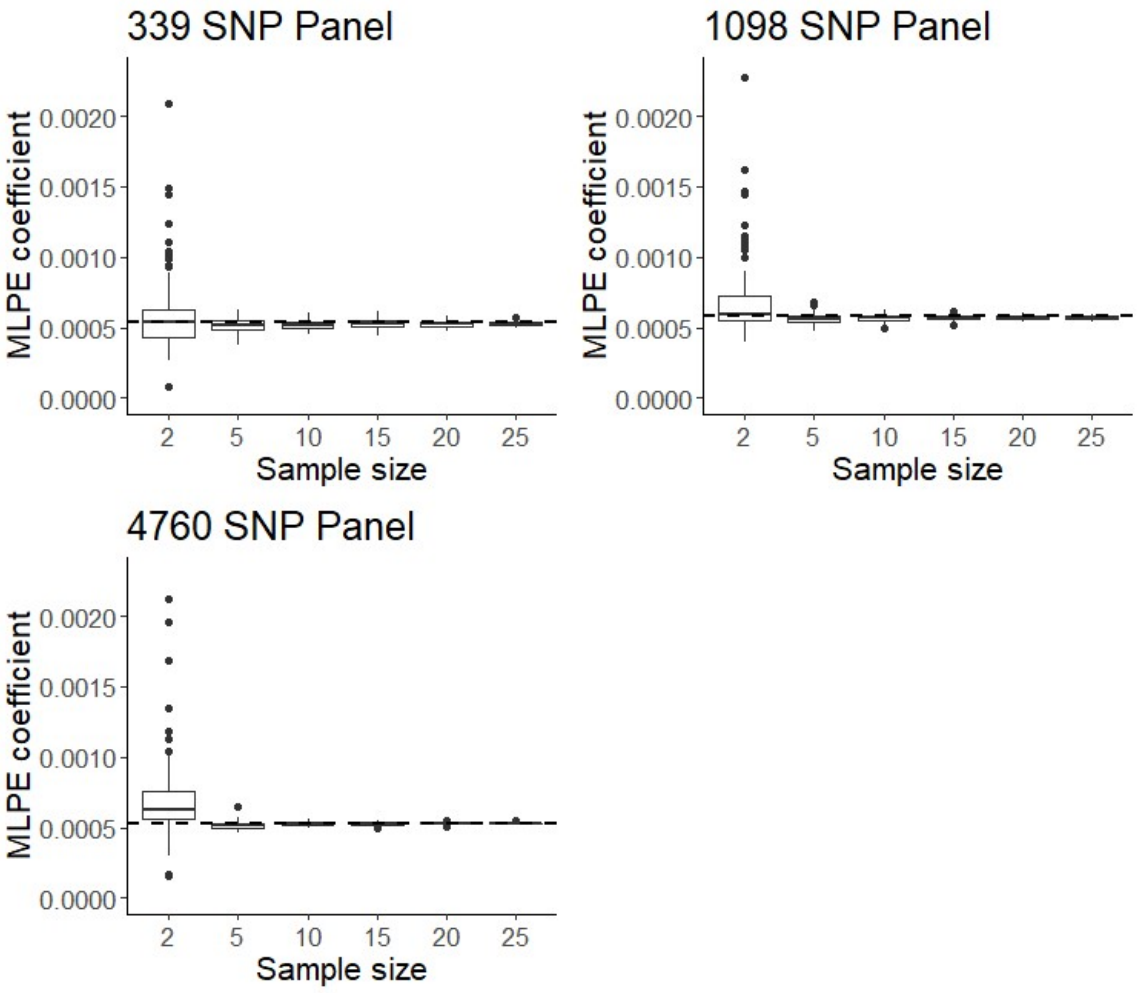

Figure 3.5 Effects of differing samples sizes (number of individuals sampled per population) on estimates of the coefficient of the relationship between pairwise F_{ST} and geographic distance estimated with MLPE mixed models, among seven *Parnassius smintheus* populations. Populations were subsampled 100 times for each sample size of between two and 25 individuals per population. Boxes show central 50% of values and the median, across subsample datasets. Tails represent values within 1.5 times the interquartile range, and points represent values outside 1.5 times the interquartile range. The coefficient as estimated using all available individuals per population (n=36-40) is represented with a dashed line. The MLPE mixed model coefficient was estimated using three SNP datasets of different sizes, generated with different levels of maximum missing data per locus: 339 SNPs (15% missing data), 1098 SNPs (20% missing data), and 4760 SNPs (30% missing data).

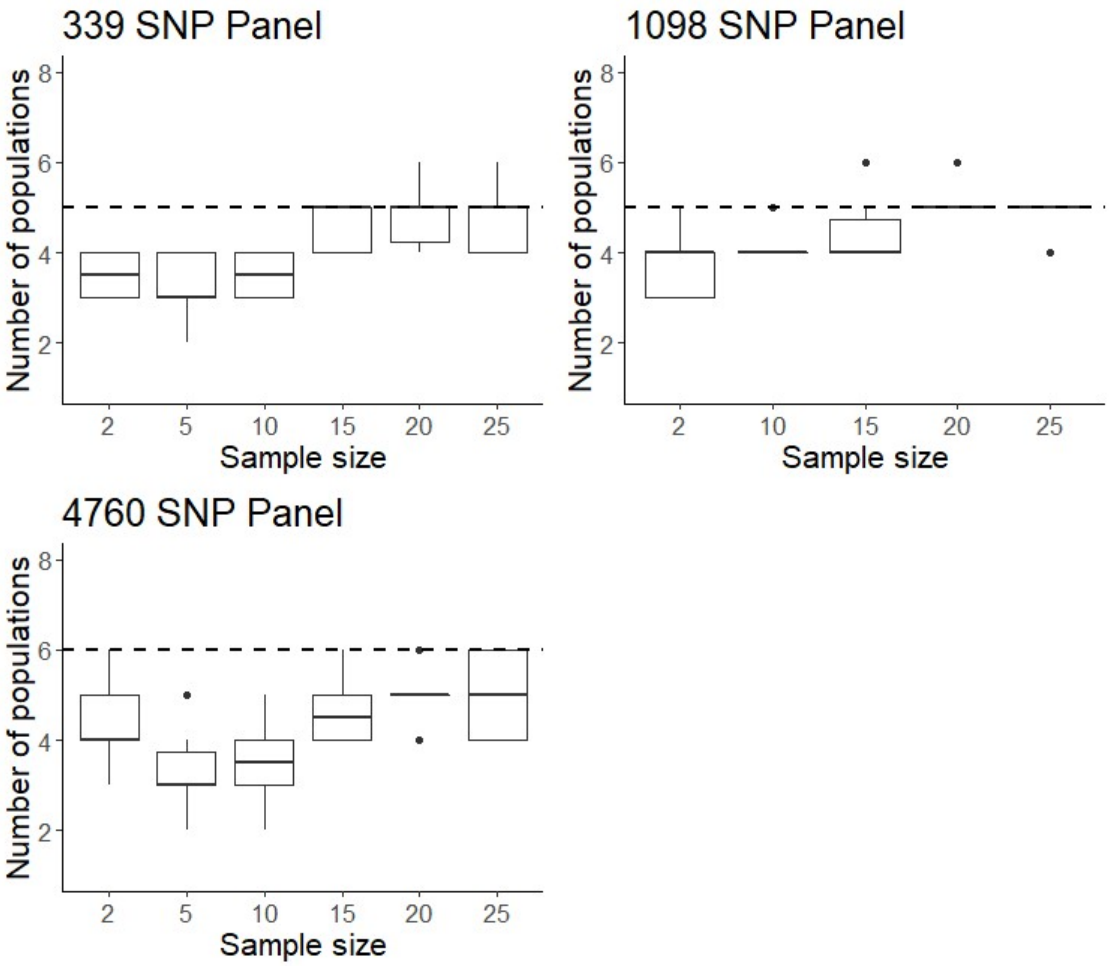

Figure 3.6 Effects of differing samples sizes (number of individuals sampled per population) on the number of population clusters identified using Geneland for seven *Parnassius smintheus* populations. Populations were subsampled 10 times for each sample size of between two and 25 individuals per population, and run for 10 iterations. The modal number of populations identified along the Monte Carlo chain was identified for the best (i.e., maximum likelihood) iteration. The number of population clusters estimated using all available individuals per population (n=36-40) is represented with a dashed line. The number of population clusters was estimated using three SNP datasets of different sizes, generated with different levels of maximum missing data per locus: 339 SNPs (15% missing data), 1098 SNPs (20% missing data), and 4760 SNPs (30% missing data). Boxes show central 50% of values and the median. Tails represent values within 1.5 times the interquartile range, and points represent values outside 1.5 times the interquartile range.

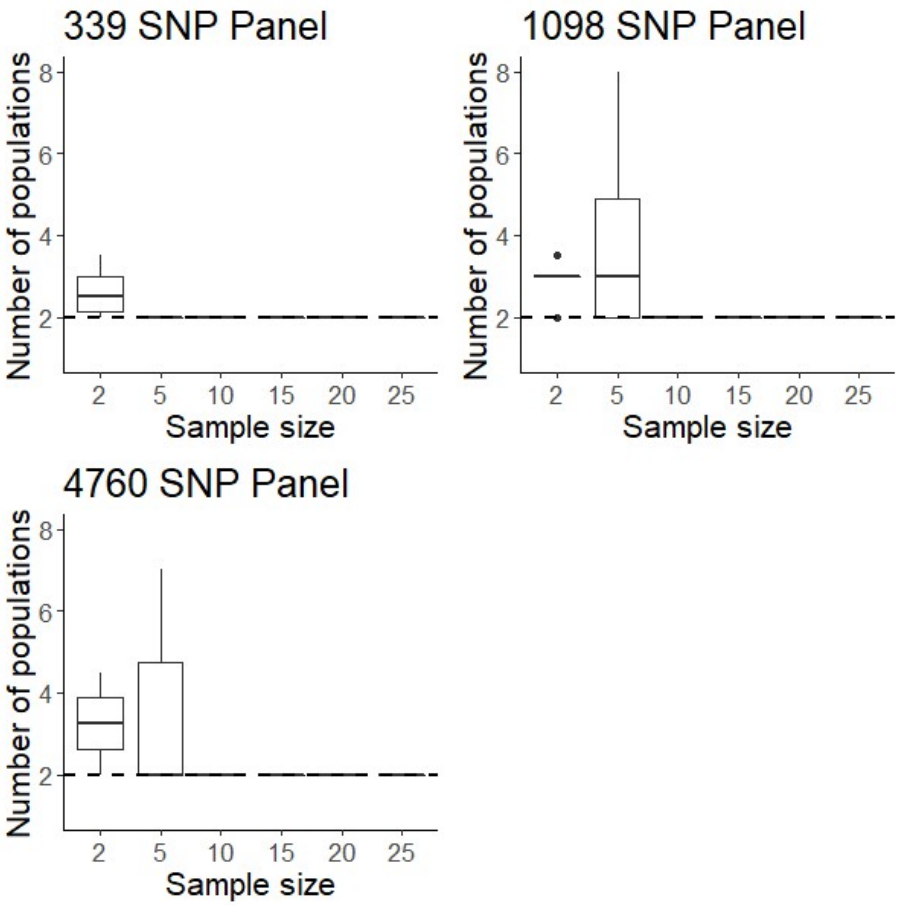

Figure 3.7 Effects of differing samples sizes (number of individuals sampled per population) on the number of population clusters identified using the K^*_ε estimator in fastSTRUCTURE for a set of seven *Parnassius smintheus* populations. Populations were subsampled 10 times for each sample size of between two and 25 individuals per population, and run for 10 iterations. K^*_ε (the number of populations that maximizes the maximum likelihood of the model) was estimated using the chooseK.py function in fastSTRUCTURE. The number of populations clusters estimated using all available individuals per population (n=36-40) is represented with a dashed line. The number of population clusters was estimated using three SNP datasets of different sizes, generated with different levels of maximum missing data per locus: 339 SNPs (15% missing data), 1098 SNPs (20% missing data), and 4760 SNPs (30% missing data). Boxes show central 50% of values and the median, across subsample datasets. Tails represent values within 1.5 times the interquartile range, and points represent values outside 1.5 times the interquartile range.

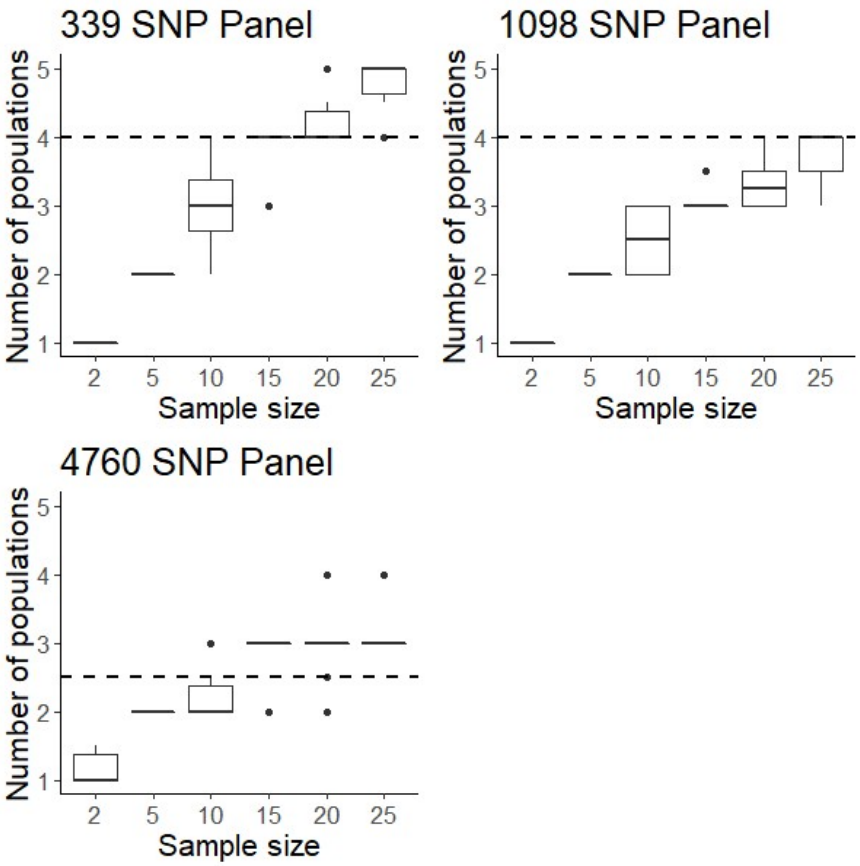

Figure 3.8 Effects of differing samples sizes (number of individuals sampled per population) on the number of population clusters identified using the $K^*_{\text{øC}}$ estimator in fastSTRUCTURE for a set of seven *Parnassius smintheus* populations. Populations were subsampled 10 times for each sample size of between two and 25 individuals per population, and run for 10 iterations. $K^*_{\text{øC}}$ (the number of populations required to explain nearly all of the ancestry in the dataset) was estimated using the chooseK.py function in fastSTRUCTURE. The number of populations clusters estimated using all available individuals per population (n=36-40) is represented with a dashed line. The number of population clusters was estimated using three SNP datasets of different sizes, generated with different levels of maximum missing data per locus: 339 SNPs (15% missing data), 1098 SNPs (20% missing data), and 4760 SNPs (30% missing data). Boxes show central 50% of values and the median, across subsampled datasets. Tails represent values within 1.5 times the interquartile range, and points represent values outside 1.5 times the interquartile range.

3.4 Discussion

The power to accurately estimate population genetic parameters is constrained both by the number and information content of molecular markers, and the number of individuals sampled (Hale et al., 2012; Morin et al., 2009; Schopen, Bovenhuis, Visker, & Arendonk, 2008). Using more molecular markers, in theory, allows for accurate analyses at lower sample sizes (Landguth et al., 2012); based on this premise, large SNP panels developed using next-generation sequencing (including RADseq) were proposed to allow for the use of much lower sample sizes than possible when using markers such as microsatellites (Willing et al., 2012). Here, SNP panels developed using double digest RADseq did not compensate for extremely low sample sizes (two individuals per population) for any type of analysis. For most analyses, a sample size of 10 was the minimum for results to be comparable with larger sample sizes (~40 individuals), and population clustering analyses required at least 15 individuals per population. SNP dataset size did not affect minimum sample size for any analysis, indicating for this range of dataset size (339-4760 SNPs) the number of individuals sampled is much more important than the number of genetic markers for determining the outcome of analyses.

3.4.1 Genetic diversity and differentiation

For estimating measures of genetic diversity, different minimum sample sizes have been reported for different measures and in different systems. In a study of two Acacia species, sample sizes of two were sufficient to estimate allelic richness and observed heterozygosity, but sample sizes of six or eight (depending on population) were required to estimate expected heterozygosity (Nazareno et al., 2017). Using two populations of the beetle *Harmonia axyridis* and 3000 SNPs, sample sizes of four were the minimum required to estimate allelic richness and observed and expected heterozygosity (expect in one population, where a sample size of six was required for observed heterozygosity) (Li et al., 2020). My results were more consistent with Li et al. (2020), with sample sizes of two being too low and sample sizes of five and higher producing estimates reasonably consistent with the estimates from the complete sample. For allelic richness particularly, my results suggest that using sample sizes of ten and higher with 1098 or more SNPs is

produces the most accurate results, but sample sizes as low as five result in slightly lower but overall very similar estimates.

Sample sizes as low as two individuals per population have been reported to be sufficient to accurately estimate global F_{ST} when using at least 1098 SNPs (Nazareno, Bemmels, Dick, & Lohmann, 2017; Willing et al., 2012). Here, a sample size of two resulted in higher median F_{ST} and markedly greater variance in F_{ST} than sample sizes of five and higher, even when using 1098 or more SNPs. A notable difference between the design of my study and those concluding that two individuals are sufficient for F_{ST} estimation is that those studies included only two empirical (Nazareno et al., 2017) or modelled (Willing et al., 2012) populations, whereas I examined differentiation among seven populations. While this means that I necessarily included more individuals in total in my analyses (e.g., a total of 14 rather than four individuals, for a sample size of two individuals per population), estimating differentiation among more than two populations allows for additional sources of variation. With more populations considered, there is a greater overall probability that a sample, particularly a small one, from any one population will have outlying allele frequencies; this would result in greater variation among subsampled datasets than might be observed in a simpler, two-population system. At higher sample sizes the chance for outlying subsamples decreases, and my results become more consistent with previous studies that included fewer populations: for sample sizes of five and above (and particularly above ten), global F_{ST} is consistent across both sample size and SNP number.

For accurate estimation of F_{ST} using the Weir and Cockerham method (as here) both high (>0.4) and low (<0.01) levels of true population differentiation require somewhat larger minimum sample sizes (six individuals) than moderately differentiated (0.05-0.2) populations (Nazareno et al., 2017). For populations with low genetic differentiation, low sample sizes can result in genetic differentiation not being detected, and for highly differentiated populations low sample sizes are more likely to overestimate differentiation (Nazareno et al., 2017). Among the populations of *P. smintheus* examined in my study, global F_{ST} estimated using the complete dataset ranged from 0.061 to 0.064, which falls

within the moderately differentiated range where low sample sizes are less likely to result in over- or under-estimates of F_{ST}.

The minimum sample size required to estimate IBD depended both on the test used, and whether the strength, or simply the significant presence, of IBD was estimated. The capacity to detect significant IBD (i.e., $p < 0.05$) was more robust to low sample sizes than the capacity to estimate the strength of IBD, (i.e., Mantel's *r* or MLPE model coefficients). The effect of sample size on the quantification of IBD has not been previously studied in detail. In simulated populations, sample size did not affect the estimation of the strength of isolation by resistance using partial Mantel tests (Landguth et al., 2012); however, the smallest sample size examined in that study was ten individuals per population. While I found that using a sample size of ten did result in somewhat more variable and lower Mantel's *r* than larger sample sizes, the effect of a small sample size was much more noticeable when using sample sizes of two and five individuals. Nonetheless, in many cases (i.e., when using MLPE mixed models, or estimating the significance of IBD using Mantel tests), I found that a sample size of five individuals per population was sufficient regardless of the number of SNPs. However, I note that the correlation between geographic and genetic distance among my sampled populations was unusually high, with values of Mantel's *r* approaching one (i.e., perfect correlation) when using all samples per population. Among populations with weaker IBD, higher sample sizes may be required to more accurately estimate pairwise genetic distances and allow IBD to be detected.

3.4.2 Population clustering

For any molecular marker, the accuracy of population clustering is improved by increasing sample size, and accurate population clustering often requires higher sample sizes than analyses of genetic diversity and differentiation (Fumagalli, 2013). For example, when using microsatellites sample sizes of 20-30 are routinely recommended for estimating genetic diversity metrics (Hale, Burg, & Steeves, 2012; Pruett & Winker, 2008). The number of individuals per population recommended for population clustering varies, but is generally more than 20 individuals and in some cases more than 50 individuals per population (Bjørnstad & Røed, 2002; Evanno, Regnaut, & Goudet, 2005).

For SNPs, the effect of sample size on population clustering has not been extensively studied using approaches such as Structure and Geneland. Using assignment approaches developed by Paetkau et al. (2004), the assignment of individuals of the lobster *Homarus americanus* to their source populations was assessed using large SNP panels (3000 SNPs) and sample sizes of ten to 34 individuals (Benestan et al., 2015). Assignment success was poor at ten individuals, and at the maximum of 34 assignment was improved but approximately 20% of individuals were still not successfully assigned to their source population. Here, I found that the minimum sample size needed to accurately estimate the number of distinct populations depended on the approach. When using the $K^*_{\text{øC}}$ estimator of fastSTRUCTURE, the effect of sample size was consistent with Benestan et al. (2015), where sample sizes of 20 and above resulted in population numbers close to that calculated from the complete dataset, but still with imperfect clustering; here, the number of populations identified was always fewer than the total number of populations sampled (seven). A similar pattern was seen when using Geneland, but with a lower minimum sample size of 15 individuals.

Overall, the number of distinct populations I identified was greater when using Geneland compared to either fastSTRUCTURE estimator. Estimates derived from Geneland and fastSTRUCTURE are rarely directly compared; when both were used to examine the population structure of *Osteoglossum* species (Souza et al., 2019) and *Arapaima gigas* populations (Oliveira et al., 2020), the same number of populations were identified using both approaches. Geneland is more frequently used in conjunction with the original Structure approach, especially when the use of Structure is not limited by a large SNP dataset. The number of populations identified are often similar between the two approaches (Coulon et al., 2008; Pometti, Bessega, Saidman, & Vilardi, 2014). While fastSTRUCTURE was developed to allow for similar but faster population assignment (and therefore allowing the use of large SNP datasets), it does not incorporate all of the model components included in Structure (Raj et al., 2014). Notably, it does not have an option to include the population of origin as a prior, whereas Geneland uses sampling coordinates as a prior (Guillot et al., 2005; Raj et al., 2014), which may be contributing to the greater number of populations identified here when using Geneland.

3.4.3 Recommendations for minimum sample size

The minimum sample size required to accurately assess genetic differentiation, diversity, and population clustering depends on the study system: in addition to the number and variability of the molecular markers used (Bjørnstad & Røed, 2002; Nazareno et al., 2017), factors such as the number of populations sampled and the true degree of genetic structuring among populations affect the necessary minimum sample size (Flesch et al., 2018). For example, my results suggest that including more populations may increase the required minimum sample size when assessing genetic differentiation (e.g., as compared to Nazareno et al., 2017 and Willing et al., 2012). Differences among study systems (i.e., the number of populations sampled and their degree of genetic differentiation) make it difficult to provide universal recommendations for a minimum sample size across study systems and analyses. When trying to minimize sampling, one option is to conduct a pilot study to examine how sample size affects the outcomes of all anticipated analyses, and to choose the highest minimum sample size across analyses. Given that the number of populations sampled affects this outcome, this pilot study would have to include many if not all of the populations included in the final analyses. This approach would be useful for populations in long term studies where sampling is expected to reoccur. When sampling occurs only once, and a low sample size is desirable (due to factors such as cost, difficulty in locating individuals, or when studying species at risk), the choice of sampling size should consider the number of populations sampled and the expected degree of genetic differentiation among populations (with increased sample size required when genetic differentiation is expected to be lower). For systems like the *P. smintheus* populations examined here, where multiple populations are sampled and genetic differentiation is detectable but not high, the following sample sizes may be used as a baseline when using large SNP datasets. If only analyses of differentiation and diversity are of interest, five individuals per population will likely be sufficient in many systems. If analyses of IBD will be conducted, sample sizes should be increased to 10-15 individuals. Likewise, if population clustering analyses will be conducted, using fewer than 15 individuals seems likely to underestimate the number of clusters detected.

3.5 Literature cited

Aguirre-Liguori, J. A., Luna-Sánchez, J. A., Gasca-Pineda, J., & Eguiarte, L. E. (2020). Evaluation of the minimum sampling design for population genomic and microsatellite studies: An analysis based on wild maize. *Frontiers in Genetics*, *11*, 870.

Allendorf, F. W. (1986). Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology*, *5*(2), 181–190.

Anderson, C. D., Epperson, B. K., Fortin, M.-J., Holderegger, R., James, P. M. A., Rosenberg, M. S., … Spear, S. (2010). Considering spatial and temporal scale in landscape-genetic studies of gene flow. *Molecular Ecology*, *19*(17), 3565–3575.

Beerli, P. (2004). Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Molecular Ecology*, *13*(4), 827–836.

Benestan, L., Gosselin, T., Perrier, C., Sainte-Marie, B., Rochette, R., & Bernatchez, L. (2015). RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular Ecology*, *24*(13), 3299–3315.

Bjørnstad, G., & Røed, K. H. (2002). Evaluation of factors affecting individual assignment precision using microsatellite data from horse breeds and simulated breed crosses. *Animal Genetics*, *33*(4), 264–270.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140.

Clarke, R. T., Rothery, P., & Raybould, A. F. (2002). Confidence limits for regression relationships between distance matrices: Estimating gene flow with distance. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*(3), 361.

Coulon, A., Fitzpatrick, J. W., Bowman, R., Stith, B. M., Makarewich, C. A., Stenzler, L. M., & Lovette, I. J. (2008). Congruent population structure inferred from dispersal behaviour and intensive genetic surveys of the threatened Florida scrub-jay (*Aphelocoma cœrulescens*). *Molecular Ecology*, *17*(7), 1685–1701.

Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*(2), 359–361.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. *Molecular Ecology*, *14*(8), 2611–2620.

Flesch, E. P., Rotella, J. J., Thomson, J. M., Graves, T. A., & Garrott, R. A. (2018). Evaluating sample size to estimate genetic management metrics in the genomics era. *Molecular Ecology Resources*, *18*(5), 1077–1091.

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLOS ONE*, *8*(11), e79667.

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, *5*(1), 184–186.

Guillot, G., Mortier, F., & Estoup, A. (2005). Geneland: A computer package for landscape genetics. *Molecular Ecology Notes*, *5*(3), 712–715.

Hale, M. L., Burg, T. M., & Steeves, T. E. (2012). Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PLoS ONE*, *7*(9).

Hoban, S. M., Gaggiotti, O. E., & Bertorelle, G. (2013). The number of markers and samples needed for detecting bottlenecks under realistic scenarios, with and without recovery: A simulation-based study. *Molecular Ecology*, *22*(13), 3444–3450.

Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405.

Keyghobadi, N., Roland, J., & Strobeck, C. (1999). Influence of landscape on the population genetic structure of the alpine butterfly *Parnassius smintheus* (Papilionidae). *Molecular Ecology*, *8*(9), 1481–1495.

Keyghobadi, N., Roland, J., & Strobeck, C. (2005). Genetic differentiation and gene flow among populations of the alpine butterfly, *Parnassius smintheus*, vary with landscape connectivity. *Molecular Ecology*, *14*(7), 1897–1909.

Kraus, R. H. S., vonHoldt, B., Cocchiararo, B., Harms, V., Bayerl, H., Kühn, R., … Nowak, C. (2015). A single-nucleotide polymorphism-based approach for rapid and cost-effective genetic wolf monitoring in Europe based on noninvasively collected samples. *Molecular Ecology Resources*, *15*(2), 295–305.

Kreitman, M. (2000). Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics*, *1*(1), 539–559.

Landguth, E. L., Fedy, B. C., OYLER-McCANCE, S. J., Garey, A. L., Emel, S. L., Mumma, M., … Cushman, S. A. (2012). Effects of sample size, number of markers, and allelic richness on the detection of spatial genetic pattern. *Molecular Ecology Resources*, *12*(2), 276–284.

Li, H., Qu, W., Obrycki, J. J., Meng, L., Zhou, X., Chu, D., & Li, B. (2020). Optimizing sample size for population genomic study in a global invasive lady beetle, *Harmonia axyridis*. *Insects*, *11*(5), 290.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, *27*(2 Part 1), 209–220.

Matter, S. F., Keyghobadi, N., & Roland, J. (2014). Ten years of abundance data within a spatial population network of the alpine butterfly, *Parnassius smintheus*. *Ecology*, *95*(10), 2985–2985.

Morin, P. A., Martien, K. K., & Taylor, B. L. (2009). Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, *9*(1), 66–73.

Nazareno, A. G., Bemmels, J. B., Dick, C. W., & Lohmann, L. G. (2017). Minimum sample sizes for population genomics: An empirical study from an Amazonian plant species. *Molecular Ecology Resources*, *17*(6), 1136–1147.

Nei, M. (1972). Genetic distance between populations. *The American Naturalist*, *106*(949), 283–292.

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

Oliveira, E. A., Perez, M. F., Bertollo, L. a. C., Gestich, C. C., Ráb, P., Ezaz, T., … Cioffi, M. B. (2020). Historical demography and climate driven distributional changes in a widespread Neotropical freshwater species with high economic importance. *Ecography*, *43*(9), 1291–1304.

Oyler-McCance, S. J., Fedy, B. C., & Landguth, E. L. (2013). Sample design effects in landscape genetics. *Conservation Genetics*, *14*(2), 275–285.

Paetkau, D., Slade, R., Burden, M., & Estoup, A. (2004). Genetic assignment methods for the direct, real-time estimation of migration rate: A simulation-based exploration of accuracy and power. *Molecular Ecology*, *13*(1), 55–65.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5), e37135.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & Deepayan, S. (2015). *nlme: Linear and Nonlinear Mixed Effects Models*.

Pometti, C. L., Bessega, C. F., Saidman, B. O., & Vilardi, J. C. (2014). Analysis of genetic population structure in *Acacia caven* (Leguminosae, Mimosoideae), comparing one exploratory and two Bayesian-model-based methods. *Genetics and Molecular Biology*, *37*(1), 64–72.

Pope, N. (2020). *CorMLPE: A correlation structure for symmetric relational data*.

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

Pruett, C. L., & Winker, K. (2008). The effects of sample size on population genetic diversity estimates in song sparrows *Melospiza melodia*. *Journal of Avian Biology*, *39*(2), 252–256.

Puckett, E. E. (2017). Variability in total project and per sample genotyping costs under varying study designs including with microsatellites or SNPs to answer conservation genetic questions. *Conservation Genetics Resources*, *9*(2), 289–304.

R Core Team. (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573–589.

Roland, J., Keyghobadi, N., & Fownes, S. (2000). Alpine parnassius butterfly dispersal: Effects of landscape and population size. *Ecology*, *81*(6), 1642–1653.

Schopen, G. C. B., Bovenhuis, H., Visker, M. H. P. W., & Arendonk, J. A. M. V. (2008). Comparison of information content for microsatellites and SNPs in poultry and cattle. *Animal Genetics*, *39*(4), 451–453.

Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science*, *236*(4803), 787–792.

Smouse, P. E., Long, J. C., & Sokal, R. R. (1986). Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Systematic Zoology*, *35*(4), 627–632.

Souza, F. H. S. de, Perez, M. F., Bertollo, L. A. C., Oliveira, E. A. de, Lavoué, S., Gestich, C. C., … Cioffi, M. de B. (2019). Interspecific genetic differences and historical demography in South American arowanas (Osteoglossiformes, Osteoglossidae, Osteoglossum). *Genes*, *10*(9), 693.

Sunde, J., Yıldırım, Y., Tibblin, P., & Forsman, A. (2020). Comparing the performance of microsatellites and RADseq in population genetic studies: Analysis of data for pike (*Esox lucius*) and a synthesis of previous studies. *Frontiers in Genetics*, *11*, 218.

Willing, E.-M., Dreyer, C., & Oosterhout, C. van. (2012). Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. *PLOS ONE*, *7*(8), e42649.

Chapter 4

4 Weather and landscape affect genomic diversity and differentiation in the alpine butterfly *Parnassius smintheus*

4.1 Introduction

A population's genetic diversity both predicts its evolutionary potential and reflects its history. Quantifying patterns of genetic diversity is therefore useful in itself when surveying specific populations, such as those of conservation or management concern, and understanding their likelihood of persisting under changing conditions (Lande, 1988). But, patterns of genetic diversity can also be used more broadly to test hypotheses about what environmental factors, such as weather or habitat amount, have shaped or continue to shape a population's dynamics (e.g., population size, rates of immigration/emigration, birth/death rates; Manel, Schwartz, Luikart, & Taberlet, 2003). This approach is especially useful when genetic samples are more easily collected than behavioral or demographic data; because a population's dynamics ultimately shape genetic diversity, the factors affecting population dynamics can be indirectly identified through those that are found to significantly predict patterns of genetic diversity (Beichman, Huerta-Sanchez, & Lohmueller, 2018; Manel et al., 2003). Understanding what factors and processes affect population dynamics is important from a conservation perspective, where it is critical to understand the conditions necessary to maintain populations at risk. By analyzing data from natural populations we can also support or refute hypotheses about which environmental factors are important in shaping population dynamics, and which of these relationships are generalizable across taxa and which are species-specific.

Rates of dispersal and fluctuations in population size are both important demographic factors that drive patterns of genetic diversity in natural populations. When loci are not under balancing selection, genetic diversity tends to be lost over time; alleles are lost from a population when, by chance, they are not passed to the next generation (i.e., individuals carrying an allele do not reproduce, their progeny do not survive, or that allele is not inherited by surviving progeny; Wright, 1931). This loss of genetic diversity, called

genetic drift, is mediated by population size. Smaller populations undergo drift more quickly, as individual chance events have a greater proportional impact when there are fewer individuals reproducing (Nei, Maruyama, & Chakraborty, 1975). Populations often experience higher rates of drift than would be expected for their census (i.e., total count) population size (Nei et al., 1975). Factors that increase the pace of drift include unequal sex ratios, non-random mating (Frankham, 1995), and historical fluctuations in population size (Vucetich, Waite, & Nunney, 1997); these factors result in an "effective" population size that is almost always smaller than the census (i.e., total count) population size (Nei et al., 1975).

In addition to lower genetic diversity within populations, a set of populations undergoing drift will become more different from each other over time as alleles are lost independently and by chance from each of the separate populations (Wright, 1943). The effects of drift can be countered by gene flow, where alleles move among populations (Slatkin, 1987). Dispersal, or the movement and settlement of individuals away from their natal populations, results in the movement of alleles among populations and is the basis for gene flow. Higher rates of dispersal are thus associated with both lower genetic differentiation among populations, as those populations will tend to share more alleles, and higher genetic diversity within populations, as alleles lost through drift are replaced by alleles introduced by gene flow (Bohonak, 1999).

One key environmental factor that can influence population size and dispersal is land cover. Land cover refers to the physical material (e.g., vegetation such as forest or grassland, bare rock or earth, or open water) that covers the land surface in a particular, spatially delimited area (Kerr & Ostrovsky, 2003). The amount and distribution of land cover surrounding a population, both locally and regionally, is a key determinant of the availability of suitable habitat (Store & Jokimäki, 2003), and the carrying capacity for that species (e.g., Livolsi, Williams, Coluccy, & Dibona, 2021). Land cover also contributes to patterns of dispersal. For many species, land cover facilitates or impedes movement among populations (Henein & Merriam, 1990). Individuals may move more readily or successfully through habitat that provides necessary resources (e.g., food, water, shelter from predators; Henein & Merriam, 1990; Pérez-Espona, McLeod, &

Franks, 2012), whereas other types of land cover may reduce movement or survival by presenting a physical barrier to movement or by simply lacking resources (e.g., Castillo et al., 2016; Funk et al., 2005).

Weather is both a spatially and temporally varying factor that can affect population size and dispersal. Local weather conditions can be advantageous or detrimental for survival or reproduction of a particular species, which can drive population size up or down, respectively (e.g., Chase, Nur, & Geupel, 2005; Lewellen & Vessey, 1998). Weather can directly reduce or enhance individual movement; for example, in the butterfly *Parnassius mnemosyne* both average rates of emigration and movement within habitat patches increases with temperature and solar radiation (Kuussaari, Rytteri, Heikkinen, Heliölä, & von Bagh, 2016). As with land cover, there are long-term patterns of weather (i.e., climate) that vary in predictable ways among populations, resulting in geographic variation in population size and dispersal. Weather is also variable across different time scales; in some cases, this variation is cyclical (e.g., in association with El Niño and other climatic oscillations), and in others is stochastic (e.g., isolated severe weather events). This variability in weather across years can drive cycles of population growth and decline; more variable weather conditions may then be associated with greater variability in population size and a lower resulting effective population size. This is especially true for insect species with short and/or non-overlapping generations where a single year with poor conditions can result in a population size collapse. Weather is an important driver of population growth in insects including the mountain pine beetle *Dendroctonus ponderosae* (Preisler, Hicke, Ager, & Hayes, 2012), the seed feeding bug *Lygaeus equestris* (Solbreck, 1991), the spittlebug *Philaenus spumarius* (Halkka, Halkka, Halkka, Roukka, & Pokki, 2006), and in butterfly species including the cabbage white *Pieris brassicae* (Roy, Rothery, Moss, Pollard, & Thomas, 2001) and the Rocky Mountain Apollo *Parnassius smintheus* (Roland & Matter, 2016). In *P. spumarius* and *P. smintheus*, the fluctuations in weather believed to be driving fluctuations in population size are associated with long-term climate cycles (the North Atlantic Oscillation and the Pacific Decadal Oscillation, respectively). The periodic nature of these climate cycles and the associated periodic unfavourable conditions mean that observed declines in population size are likely to be experienced repeatedly rather than as a one-time event.

A population network of the alpine butterfly *Parnassius smintheus* in western Alberta, Canada has been the subject of a long-term mark recapture study (since 1995), providing insight into how landscape and weather affect dispersal and population growth. Dispersal rates are associated with land cover (Keyghobadi, Roland, & Strobeck, 1999); based on both recapture frequency and genetic distances, *P. smintheus* disperses less frequently through forest than through meadow. Population growth rates are associated with early-winter weather (Roland & Matter, 2016). Periodic network-wide population bottlenecks have been observed in these populations and are likely driven by early-winter egg mortality during particularly cold or warm years (Matter, Doyle, Illerbrun, Wheeler, & Roland, 2011; Roland & Matter, 2013, 2016); eggs freeze when exposed to temperatures below -28 °C if they are not protected by an insulating layer of snow. This occurs both in cold years with an early onset of low temperatures prior to sufficient snowfall, as well as in warm years where frequent thaws reduce snow cover (Matter et al., 2011; Roland & Matter, 2013, 2016); Additionally, in warm years with little snow cover eggs exposed to warm temperatures for several consecutive days in early winter may emerge prematurely (Matter et al., 2011; Roland & Matter, 2013, 2016). These bottlenecks are associated with changes in genetic diversity and differentiation; patterns of isolation by distance broke down as pairwise genetic differentiation increased after two documented bottlenecks, while allelic richness decreased immediately after the more severe of the two documented bottlenecks.

These documented relationships among weather, demography, and genetics of *P. smintheus* are based on a geographically restricted set of populations; importantly, weather varies only among years with little spatial variation. Here, I explore whether these relationships are generalizable at a larger scale, among populations separated by tens of kilometers. I test hypotheses about which, if any, landscape and weather variables contribute to patterns of genetic differentiation and diversity in *P. smintheus*, based on the variables identified in previous research. I predict that weather variables predicted to be associated with higher *P. smintheus* egg survival (moderate average temperature, fewer extreme temperature events, and greater snow cover) will be associated with higher genetic diversity and lower genetic differentiation, as a result of less frequent and/or less severe bottlenecks and overall more stable population size. I predict that higher total

amounts of open landcover surrounding sampling sites will reflect better habitat quality and quantity and support larger populations, and thus be associated with greater diversity and lower differentiation. I also predict that sampling sites with greater connectivity to surrounding patches of open habitat will allow greater dispersal among populations, thus maintaining higher genetic diversity and lower differentiation as alleles are exchanged. I thereby test the relative importance of weather versus landscape variables on genetic differentiation and diversity, as well as the relative importance of mean weather conditions over time versus variability and extremes in those conditions.

4.2 Methods

4.2.1 Genetic data collection

Parnassius smintheus individuals were collected from 21 alpine meadows in western Alberta in 1995, 1996 and 1999. Sampling sites were separated by ~6 to 120 km. The sites were located in three distinct geographic regions: Banff (n=8), West Kananaskis (n=7), and East Kananaskis (n=6). Adults were captured using hand nets and whole body samples were stored in glassine envelopes at -80 °C until DNA extraction. I used DNeasy blood and tissue kits (Qiagen, Germantown, MD) to extract DNA from the head and thorax of 501 individuals. I conducted double digest restriction site associated DNA sequencing using a starting amount of 200 ng of DNA per individual. I digested DNA with the restriction enzymes *Nla*III and *Eco*RI-HF (New England Biolabs, Ipswich, MA). I tagged restricted DNA as per Rašić, Filipović, Weeks, & Hoffmann (2014) before pooling and size selecting for fragments between 200 and 500 bp using Sera-Mag solid-phase reversible immobilization beads (SPRI; GE Healthcare Life Sciences, Chicago, IL). I amplified the size selected library by PCR (see Chapter 2 for PCR conditions). Size selection was verified using an Agilent 2100 Bioanalyzer before libraries were sequenced on an Illumina HiSeq 2500.

I called single nucleotide polymorphisms (SNPs) using the Stacks pipeline (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013) with the following parameter values: a minimum stack depth of 3 reads (m), a maximum of 3 nucleotides differing between reads in an assembled stack (n), a maximum difference of 2 nucleotides between

combined stacks (M), and a maximum difference of 4 nucleotides between assembled stacks and additional aligned reads (N). I filtered the SNP dataset to exclude individuals with low coverage (less than 50% of loci genotyped), and subsequently by data coverage per locus (loci with less than 80% of individuals genotyped were excluded).

4.2.2 Population genetic variables

With the filtered SNP dataset, I estimated two variables that represent genetic diversity and differentiation at the level of individual populations, respectively: expected heterozygosity and distance-weighted mean Nei's D (Nei, 1972). I chose Nei's D specifically because it is not dependent on heterozygosity as are other measures of genetic differentiation (e.g., F_{ST}; Meirmans & Hedrick, 2011). I used the R (R Core Team 2015) package pegas (Paradis, 2010) to determine the percent of loci in each population that departed from Hardy-Weinberg equilibrium. I used the R package diveRsity (Keenan, McGinnity, Cross, Crozier, & Prodöhl, 2013) to calculate the mean expected heterozygosity (H_E) across loci. I used the R package adegenet (Jombart, 2008; Jombart & Ahmed, 2011) to calculate pairwise Nei's D between each population and all other sampled populations. For each population, I estimated the mean pairwise Nei's D to all other populations in the dataset, weighted by the geographic distance to each other population (i.e., mean distance-weighted Nei's D). This provided a site-specific, distance-weighted measure of genetic differentiation. The weights were set as: e^{-d}, where d is the pairwise geographic distance in kilometers between the centroids of sampling sites.

4.2.3 Landscape data

I obtained landcover maps as 25x25m rasters from the Canadian Forest Service's Earth Observation for Sustainable Development of Forests. I used ArcMap to combine all high elevation meadow and barren rock into a single category of "open" landcover. All other landcover types were binned to a second category. This includes forests, which restrict movement in mark-recapture studies, and areas below elevations of 1900m where *P. smintheus* is not typically observed regardless of landcover type (Guppy & Shepard, 2001). I extracted landcover rasters from a 1, 2, and 5 km radius around the centroids of each sampling site using the clip function in ArcMap. I used FragStats to calculate the

percent of open landcover (PLAND) and the mean distance between open patches for each site (ENN_MN).

4.2.4 Weather data

For each sampling site, I obtained modelled historical weather data from Natural Resources Canada's interpolated spatial models (Hutchinson et al., 2009). I chose a set of weather variables representing several hypotheses of how weather could affect population genetic variables, based on variables that predict changes in *P. smintheus* population growth rates at a smaller spatial scale (Table 4.1). I summarized these variables for each site across all available years of data (1960-2015), as well as for the 5, 10, and 20 years preceding sampling in 1995.

I primarily examined early-winter (November) weather variables because of their importance in determining *P. smintheus* population growth (Roland & Matter, 2016). I also included two variables that had been previously examined and did not predict changes in *P. smintheus* population size: July temperature and February snow depth (Roland & Matter, 2016). I included July temperature to determine whether any trends in November temperature could be differentiated from temperature during another biologically significant time of the year (i.e., flight season), and February snow depth to examine whether any trends in November snow depth were specific to November or would be observed throughout the winter.

I obtained 1 km gridded snow depth estimates from the National Operational Hydrologic Remote Sensing Center's SNODAS (snow data assimilation system) model (National Operational Hydrologic Remote Sensing Center, 2004). The SNODAS model is updated daily using a suite of modelled and remotely sensed variables including snow cover, precipitation, temperature, and solar radiation. I extracted snow depth from the grid cells containing the centroids of each sampled meadow for all years of available data (2010-2019). Although the SNODAS snow depth data were collected after the genetic data, it seems reasonable to extrapolate that meadows (as a result of their geography) are likely to experience similar relative patterns of snowfall even if average snowfall may be shifting over time.

4.2.5 Models

I used all landscape and weather variables individually as fixed effects in separate models for each of two response variables: expected heterozygosity and Nei's D. For all predictors other than mean November temperature, I predicted a linear relationship between the predictor and the genetic response variable. However, I hypothesized that moderate mean November temperatures would provide the best conditions for *P. smintheus* larval survival, with survival being reduced by both extreme warm and cool conditions (Roland & Matter, 2016). I thus used a quadratic model of the form mean November temperature + (mean November temperature) ^{2}~Nei's D or H_E.

I built linear models in R, and used the MuMIn package to calculate corrected Akaike's Information Criterion (AIC_C). I considered models within two AIC_C of the best (i.e., lowest AIC_C) model to also have strong support, and models within 4 of the best model to have some support (Burnham & Anderson, 2002). I fit models using maximum likelihood estimation for assessing model performance with AIC_C, and again with restricted maximum likelihood to estimate model parameters. After running models with all predictors individually, I also examined models with two predictor variables: the best (i.e., lowest AIC_C) landscape and the best weather variables of interest (i.e., November weather variables). I compared the AIC_C of these models to the models with the individual best landscape and weather predictors to determine whether the combined model including both weather and landscape variables better explained the data than models with single predictors. I also compared the proportion of variance explained (R^2) by the best landscape, best weather, and combined models.

I examined the residuals of all models that performed better than the null for linearity, normality, and equal variance. I also calculated Moran's I using the package ape (Paradis & Schliep, 2019) to test for spatial autocorrelation in residuals. I used the R package nlme to examine the same set of fixed effects with the addition of region (Banff, East Kananskis, and West Kananaskis) as a random effect.

Table 4.1 A summary of all weather and landcover variables used as predictors in linear models explaining genetic diversity and differentiation in populations of *Parnassius smintheus*. All factors hypothesized to result in higher frequency of demographic bottlenecks would be expected to reduce effective population size, and therefore lead to reduced genetic diversity and greater genetic differentiation.

Variable	Category	Definition	Associated Hypothesis
Mean November temperature (°C)	Weather	The mean temperature recorded daily for November, averaged across all November days 1960-2015	Moderate November temperatures maintain *P. smintheus* population size by preventing egg/larval mortality in early winter.
Mean November daily high (°C)	Weather	The highest temperature recorded daily in November, averaged across all November days 1960-2015.	Warmer November temperatures may maintain *P. smintheus* population size by preventing egg freezing. However, warm temperatures associated with snow melt, increased metabolic rate or precocious hatch may also increase the risk of mortality and therefore bottlenecks.
Mean November daily low (°C)	Weather	The lowest temperature recorded daily in November, averaged across all November days 1960-2015.	Sites with colder nighttime temperatures on average may carry greater risk of egg freezing, and therefore of bottlenecks.
Mean November snow depth (m)	Weather	The daily November snow depth as modelled by SNODAS, averaged across all November days 2010-2018.	Snow cover insulates eggs from supercooling temperatures. Sites with greater snow cover should experience fewer bottlenecks in population size.
November extreme temperature events	Weather	The total number of times temperatures in November were observed above 6 °C or below -28 °C from 1960 to 2015.	November temperatures above 6 °C or below -28 °C may result in higher egg mortality (see below).

Table 4.1 Continued

November extreme high temperature (°C)	Weather	The total number of times temperatures in November were observed above 6 °C from 1960 to 2015.	Same hypothesis as for 'Mean November daily high', but the number of extreme events may be more important that the mean value. November temperatures above 6 °C specifically are associated with negative population growth.
November extreme low temperature (°C)	Weather	The total number of times temperatures in November were observed below -28 °C from 1960 to 2015.	Same hypothesis as for 'Mean November daily low, but the number of extreme events may be more important that the mean value. Parnassius smintheus eggs freeze below -28 °C.
Lowest mean November snow depth (m)	Weather	The mean snow depth as modelled by SNODAS across all days in November, for the year between 2010 and 2018 with the lowest mean November snow depth.	Same hypothesis as for 'Mean November snow depth', but extreme lows in snow cover may be more important than overall means.
Percent landscape as open (%)	Landscape	For a 5 km buffer around the centroid of each sampled meadow, the percent of the landscape composed of high elevation open land cover.	Open land cover facilitates dispersal. Greater amounts of open land cover should maintain greater genetic diversity and reduce genetic differentiation between sites.
Mean of Euclidean distance between open patches (m)	Landscape	For a 5 km buffer around the centroid of each sampled meadow, the mean distance between neighbouring patches of high elevation open land cover.	Open land cover facilitates dispersal. Patches of open land cover that are more contiguous and less fragmented should allow for greater dispersal, maintaining genetic diversity and reducing differentiation between sites.

4.3 Results

4.3.1 Genetic data

I sequenced 501 *P. smintheus* individuals and generated an initial permissive SNP dataset filtering for 60% coverage at each locus. Using this initial dataset, individuals with data at fewer than 50% of the 12 291 loci were dropped, leaving 456 individuals for further analysis. The final sample size per population ranged from 8 to 39 individuals (Table 4.2). Loci with genotypes for at least 80% of these individuals were retained for analysis, for a final dataset of 1098 SNPs.

On average, 15% of all loci across all populations were out of Hardy-Weinberg equilibrium. The lowest observed proportion of loci out of equilibrium in any population was 7%, and the highest was 28%. On average, loci in Banff and West Kananaskis populations were more likely (both at 16%) than in East Kananaskis populations (11%) to be out of equilibrium. This was typically (for ~75% of loci) due to an excess of homozygotes. Expected heterozygosity ranged between 0.11 – 0.2, and mean distance-weighted Nei's D ranged between 0.0075 – 0.0401 (Table 4.2).

Table 4.2 Basic information for each of 21 *Parnassius smintheus* populations (Nei's D: mean distance-weighted Nei's D; H_E: expected heterozygosity; HWE: proportion of loci out of Hardy-Weinberg equilibrium). Populations were sampled from three regions: Banff (B), East Kananaskis (EK), and West Kananaskis. Nei's D, H_E, and HWE were estimated from 1098 SNP loci.

Population	Region	Sample size	Nei's D	H_E	HWE
Cascade 1	B	11	0.015	0.17	0.08
Flint Peak	B	38	0.012	0.18	0.22
FortyMile Creek	B	20	0.015	0.17	0.13
North Cascade 1	B	16	0.016	0.17	0.15
Panther Mtn	B	36	0.009	0.19	0.22
Mount Peechee	B	9	0.040	0.12	0.07
Snow Creek	B	32	0.009	0.19	0.19
Stony Creek	B	38	0.010	0.20	0.24
Mount Baldy	EK	13	0.030	0.14	0.08
E (Lusk Ridge)	EK	13	0.030	0.15	0.09
Forget-Me-Not Ridge	EK	20	0.018	0.17	0.11
Moose Mtn	EK	19	0.015	0.14	0.16
Powderface Ridge	EK	20	0.014	0.17	0.13
Volcano Ridge	EK	9	0.018	0.12	0.07
Elk	WK	8	0.027	0.11	0.08
Fortress Mtn	WK	39	0.011	0.17	0.28
Mount Kent	WK	36	0.008	0.17	0.24
Mount Kidd	WK	12	0.013	0.13	0.09
Mist Mtn	WK	12	0.027	0.15	0.08
Pigeon Mtn	WK	19	0.027	0.16	0.12
Wedge	WK	36	0.012	0.17	0.23

4.3.2 Landscape and weather

The average temperature in November among all sites and all years of available data was -6.4 °C; the warmest site had an average November temperature of -4.8 °C and the coolest site had an average temperature of -7.8 °C. (Supplemental Table 4.1). Mean November snow depth ranged from 17.5 cm to 63.5 cm, with an average of 34.8 cm.

Several pairs of weather variables were highly correlated ($r>0.7$). These included: July and November daily highs ($r=0.96$); July and November daily lows ($r=0.89$); mean February snowfall and mean November snowfall ($r=0.93$); minimum February snowfall and minimum November snowfall ($r=0.78$); mean and minimum November snowfall ($r=0.78$); and mean and minimum February snowfall ($r=0.91$). The two landscape variables (percent total open landscape and mean distance between open patches) were moderately correlated ($r=-0.56$, $p=0.0077$). In addition, the daily low November temperature (i.e., lowest daily temperature averaged across all November days and years) was moderately negatively correlated ($r=-0.53$, $p=0.013$) with the lowest observed mean November snow depth, indicating that sites with higher minimum daily temperatures tended to experience more extreme low snow depths.

4.3.3 Model output

I considered weather data over varying time spans and landscape data with varying buffer sizes; only results for the full time span (all data; 1960-2016) and largest buffer size (5 km) are presented here, as there were the scales for which the fixed effects best predicted genetic distance and heterozygosity (Table 4.3). I also examined models with a random effect incorporating region and found no difference compared to models without random effects, and so only present here the linear models without random effects. All supported linear models were examined for spatial autocorrelation in their residuals, and no spatial autocorrelation was found ($p=0.2$-0.9).

Expected heterozygosity was best explained by the mean distance between open-cover patches, for the landscape in a 5 km radius around each site (Table 4.3). Lower distances between patches were associated with higher H_E (Figure 4.1). Four additional single variable models had some support (i.e., were within 4 AIC_c of the best model), including

mean November daily lows, mean July daily lows, lowest mean November snow depth, and lowest February mean snow depth. For mean November and July daily lows, colder temperatures were associated with higher H_E (Figure 4.1). For the lowest mean November and February snow depth, greater snow depth was associated with higher H_E (Figure 4.1). I examined models combining mean distance between open-cover patches with either lowest mean November snow depth or mean November daily lows. Both of these combined models were within 2 AIC_C of the best model and were therefore supported.

Mean distance-weighted Nei's D was best explained by the model combining mean distance between open-cover patches, and mean November snow depth (Table 4.4). Among the single variable models the best supported model was mean November snow depth, where greater snow depth was associated with lower genetic distances (Figure 4.1). One other single variable model – mean distance between open-cover patches – was strongly supported (i.e., within 2 AIC_C of the best model); lower distances between patches were associated with lower genetic differentiation. Four additional single variable models had some support (i.e., were within 4 AIC_C of the best model); these included lowest mean February snow depth, lowest mean November snow depth, mean February snow depth, and mean November daily lows.

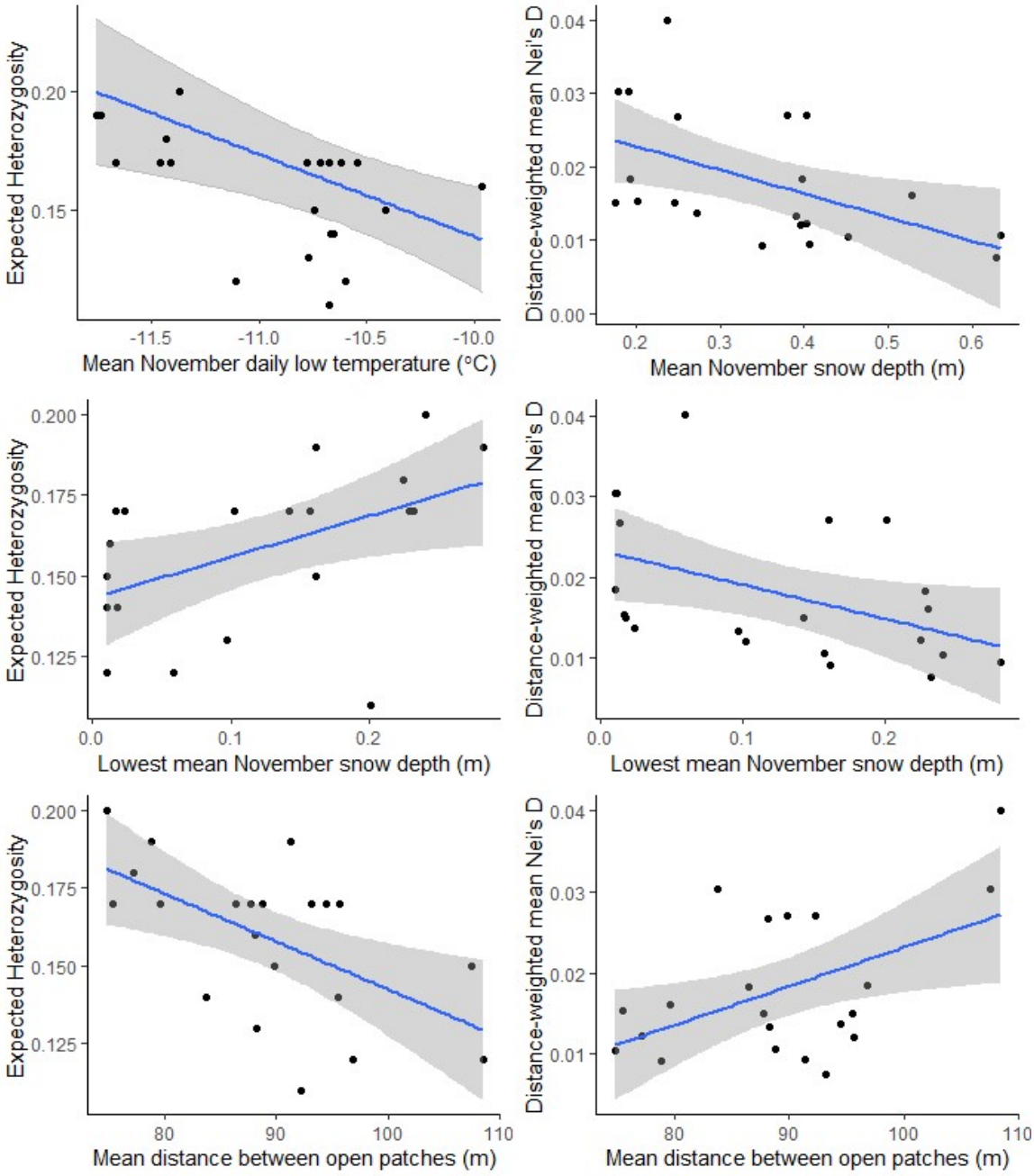

Figure 4.1 Relationships between expected heterozygosity or mean distance-weighted Nei's D and their best single predictors (including weather variables only in November) for *Parnassius smintheus* populations. The blue line indicates the predicted variable and the grey bar indicates the 95% confidence interval.

Table 4.3 Output for linear models predicting the expected heterozygosity of *Parnassius smintheus* populations. Δ AIC_C is calculated as the difference in AIC_C from the best model (for which Δ AIC_C =0). Models within 4 AIC_C of the best model are bolded.

	Model components	Coefficients	p-values	AIC_C	Δ AIC_C
Long-term weather variables	mean November temperature	-9.71E-02; -7.39E-03	0.33; 0.35	-87.72	10.10
	mean November daily highs	6.32E-04	0.87	-89.50	8.30
	mean November daily lows	**-2.53E-02**	**0.02**	**-95.66**	**2.14**
	mean November snow depth	6.89E-02	0.08	-92.96	4.85
	mean February snow depth	2.36E-02	0.07	-93.13	4.67
	July mean temp	6.26E-04	0.91	-89.50	8.30
	mean July daily highs	1.80E-03	0.55	-89.90	7.90
	mean July daily lows	**-2.98E-02**	**0.02**	**-95.40**	**2.40**
Extreme weather variables	number of extreme temperature events	-3.73E-04	0.70	-89.66	8.14
	number of extreme temperature highs	-2.62E-04	0.68	-89.70	8.10
	number of extreme temperature lows	5.36E-04	0.70	-89.60	8.20
	lowest mean November snow depth	**1.28E-01**	**0.02**	**-95.37**	**2.43**
	lowest mean February snow depth	**3.51E-02**	**0.02**	**-95.29**	**2.51**
Landscape variables	% landscape as open (1 km buffer)	2.69E-05	0.92	-89.51	8.29
	% landscape as open (2 km buffer)	1.15E-04	0.65	-89.72	8.08
	% landscape as open (5 km buffer)	3.64E-04	0.19	-91.44	6.36
	mean distance between open patches (1 km buffer)	3.30E-04	0.33	-90.59	7.21
	mean distance between open patches (2 km buffer)	-7.76E-04	0.16	-91.73	6.07
	mean distance between open patches (5 km buffer)	**-1.54E-03**	**0.01**	**-97.80**	**0**
Combined models	**mean distance between open patches + lowest mean November snow depth**	**-1.19E-03; 8.21E-02**	**0.04; 0.12**	**-97.56**	**0.24**
	mean distance between open patches + mean November daily lows	**-1.14E-03; -1.45E-02**	**0.06;0.19**	**-96.78**	**1.02**

Table 4.4 Output for linear models predicting distance-weighted mean Nei' s D among *Parnassius smintheus* populations. Distance-weighted mean Nei' s D was calculated for each population as the weighted mean of pairwise Nei' s D to all other populations, with weights calculated as the negative exponent of geographic distance. Δ AIC_C was calculated as the difference in AIC_C from the best model (Δ AIC_C =0). Models within 4 AIC_C of the best model are bolded.

	Model components	Coefficients	p-values	AIC_C	Δ AIC_C
Long-term weather variables	mean November temperature	5.58E-02; 4.23-03	0.11; 0.12	-133.59	6.92
	mean November daily highs	6.40E-04	0.65	-133.12	7.39
	mean November daily lows	**7.17E-03**	**0.07**	**-136.56**	**3.95**
	mean November snow depth	**-3.22E-02**	**0.02**	**-139.28**	**1.22**
	mean February snow depth	**-9.52E-03**	**0.04**	**-137.65**	**2.85**
	July mean temp	-8.26-02; 3.92E-03	0.08; 0.08	-133.05	7.46
	mean July daily highs	1.24E-04	0.91	-132.89	7.62
	mean July daily lows	7.04E-03	0.14	-135.29	5.21
Extreme weather variables	number of extreme temperature events	3.09E-04	0.37	-133.77	6.73
	number of extreme temperature highs	1.56E-04	0.49	-132.97	7.54
	number of extreme temperature lows	-1.43E-04	0.77	-133.43	7.08
	lowest mean November snow depth	**-4.20E-02**	**0.04**	**-137.80**	**2.70**
	lowest mean February snow depth	**-1.18E-02**	**0.04**	**-137.91**	**2.60**
	% landscape as open (1 km buffer)	-7.55E-05	0.41	-133.66	6.84
	% landscape as open (2 km buffer)	-8.07E-05	0.37	-133.77	6.73
	% landscape as open (5 km buffer)	-1.09E-04	0.27	-134.23	6.28
	mean distance between open patches (1 km buffer)	-8.60E-05	0.48	-133.45	7.05
	mean distance between open patches (2 km buffer)	2.41E-04	0.22	-134.57	5.94
	mean distance between open patches (5 km buffer)	**4.79E-04**	**0.02**	**-138.93**	**1.58**
Combined models	**mean distance between open patches + mean November snow depth**	**3.69E-04; -2.54E-02**	**0.06; 0.05**	**-140.50**	**0.00**
	mean distance between open patches + lowest mean November snow depth	**3.62E-04; -2.83E-02**	**0.09; 0.16**	**-138.21**	**2.30**
	mean distance between open patches + mean November daily lows	**3.82E-04; 3.57E-03**	**0.10; 0.40**	**-136.69**	**3.81**

4.4 Discussion

Both landscape connectivity and early winter weather explain genetic diversity and differentiation in *P. smintheus* populations. The genetic consequences of population collapses putatively driven by early-winter weather have been observed in a smaller scale study with time series data (Caplins et al., 2014; Jangjoo, Matter, Roland, & Keyghobadi, 2016); I demonstrate here that these effects can also be seen in a single, larger scale snapshot of populations across Alberta.

4.4.1 Relative importance of dispersal and population size fluctuation

Landscape connectivity and weather models were both supported as predictors of *P. smintheus* genetic differentiation and diversity. Landscape connectivity primarily reflects potential for dispersal, while November temperature and snow depth likely drive changes in population size by affecting larval mortality. Based on model performance neither dispersal nor changes in population size seem to play a dominant role in shaping genetic diversity or differentiation; rather, both have similarly important roles. In other systems, the roles of landscape and weather in shaping genetic structure are highly dependent on the species and area under study. In some cases, including in the cotton bollworm *Helicoverpa armigera* (Zhang et al., 2018) and the dragonfly *Orthetrum coerulescens* (Herzog & Hadrys, 2017), weather affects population growth (and as a result genetic diversity) more strongly than land cover.

It is often difficult to disentangle the effects of weather from landscape – weather can affect the landscape, and the landscape can also affect weather patterns. In *Papilio machaon* and *Papilio zelicaon*, climate variables such as solar radiation accounted for more genetic variation than habitat variables; however, the open, high elevation hilltops preferred by the adults are also where total solar radiation tends to be the highest (Dupuis, Cullingham, Nielsen, & Sperling, 2019). Similarly, in the Tasmanian devil (*Sarcophilus harrisii*) temperature is a significant predictor of gene flow, but likely because temperature affects land cover (e.g., vegetation). For the cyprinid *Squalius valentinus*

strong seasonal fluctuations in weather are important drivers of genetic structure, but it is the effects of weather (especially precipitation) on the seascape that affects population size and dispersal (Perea & Doadrio, 2015). The populations of *P. smintheus* I examine here are valuable in that while there is likely some interaction between climate and landscape (as is always the case), the effects of weather and land cover on population size and dispersal are discrete and can be evaluated separately. The relationships between snow depth and genetic diversity and differentiation support the hypothesis that weather affects larval survival and therefore population size directly. Likewise, landscape configuration (i.e., the connectivity of open land cover) appears to affect dispersal rates directly; the amount and configuration of open land cover may be partially determined by climate, but the actual effects of land cover on *P. smintheus* dispersal are not dependent on weather conditions (Matter et al., 2011).

Understanding the separate effects of weather and landscape on populations is important when trying to predict how those populations will respond to changing environmental conditions. Many populations are experiencing weather changes as a result of climate change; as a result, trying to analyze how populations are responding to specific changes in landscape (e.g., changes in human land use) is made more difficult by the background of climate change. In the dragonfly *O. coerulescens* genetic diversity declined following the clearing of their canal habitat; however, the decline was likely due to a drought that occurred several years later than the changes to their habitat (Herzog & Hadrys, 2017). Without accounting for the effects of weather conditions, erroneous conclusions about the cause of the decline may have been reached; in conservation efforts, this can lead to recommendations (e.g., habitat restoration) that will not address the true source of population declines (e.g., more severe weather brought on by climate change).

4.4.2 Landscape connectivity and configuration

Landscape connectivity within a 5 km radius, but not landscape composition, predicted both expected heterozygosity and Nei's genetic distance for populations of *P. smintheus*. In a study of the same set of *P. smintheus* populations examined here, both landscape composition and configuration across entire regions were associated with genetic differentiation and diversity (as estimated with microsatellites); specifically, the East

Kananaskis region, which had both the most forest cover and most fragmented high altitude open land cover, also had the greatest global genetic differentiation and lowest within-population diversity compared to the less forested and thus higher connectivity West Kananaskis and Banff regions (Keyghobadi, Roland, & Strobeck, 2005). In that study, landscape composition was measured across entire regions as opposed to locally (within the 5 km radius around each sampling site) and was thus strongly confounded with connectivity across the region – if there is more open land cover, populations are more likely to have more open land cover connecting them. At a smaller scale (<12 km), *P. smintheus* dispersal and patterns of genetic differentiation are correlated with land cover composition between pairs of meadows (Keyghobadi et al., 1999; Roland, Keyghobadi, & Fownes, 2000), which is another measure of the landscape that potentially confounds composition and connectivity. By using separate measures of composition (percent open land cover) and connectivity (distance between open patches), I show that landscape connectivity has distinct effects on genetic diversity and differentiation.

The importance of landscape connectivity but not composition informs what mechanisms are driving patterns of genetic differentiation and diversity. Landscape connectivity is associated with the ability of *P. smintheus* to disperse between populations, whereas landscape composition at this scale (a 5 km radius) represents the amount of available habitat. An association between landscape composition and population genetic parameters could indicate that available habitat was limiting population size (i.e., with less habitat, smaller populations would be more vulnerable to the effects of drift). Habitat area does limit effective population size in some species, including the tiger salamander *Ambystoma californiense*, where effective population size is correlated with breeding pond size (Wang, Johnson, Johnson, & Shaffer, 2011). This does not appear to be the case here; although there are no available estimates of population size, based on the expected heterozygosity there is no evidence that effective population size is limited by the amount of open landcover. Other factors and associated mechanisms may play a greater role than habitat size in limiting effective population size and thus genetic diversity. This could include population bottlenecks, as observed in a *P. smintheus* population network on Jumpingpound Ridge in the East Kananaskis region; average

effective population size would be limited by the low population numbers immediately following a bottleneck, rather than the maximum population size supported by the meadow. Specifically, the effective population size would be estimated as the harmonic mean of annual population sizes, where years with extremely low population size have an outsized effect on N_E (Vucetich, Waite, & Nunney, 1997). Alternatively, landscape composition as measured by open landcover may not reflect appropriate *P. smintheus* habitat - the host plant, *Sedum lanceolatum,* is likely not present in all areas categorized as "open". This could explain why landscape connectivity and not composition predicted genetic patterns; landscape that contributes to connectivity need merely be unforested to allow for *P. smintheus* movement, whereas landscape that contributes to habitat must contain the host plant. This could be addressed in future research by modelling the distribution of *S. lanceolatum* based on its habitat requirements and observed occurrences. Using a map of *S. lanceolatum* distribution instead of open landcover would better represent potential *P. smintheus* habitat and allow for more accurate quantification of habitat composition. Using both open landcover and predicted host plant landcover, and comparing separately how the composition and connectivity of each is associated with genetic metrics could also illuminate the relative contributions of *P. smintheus* dispersal (mediated by open landcover) and potential maximum effective population size (mediated by host plant availability) on genetic patterns.

That the configuration of habitat is important to a population's persistence and size, beyond the total amount of habitat available, has been accepted in conservation, population ecology, and landscape genetics (Cushman, Shirk, & Landguth, 2012; Hanski & Ovaskainen, 2000; Robert, 2009). The habitat amount hypothesis was proposed as an alternative to this paradigm; Fahrig (2013) hypothesized that patch size and isolation did not contribute significantly to the total species richness or size of populations, and that total habitat area (and not how it is configured) was the primary factor. Fahrig suggested that in studies where patch size and isolation had been identified as predictors of species richness or population size (and arguably by extension, genetic structure), these factors were just a proxy for habitat area. Since the proposal of the habitat amount hypothesis, there have been many studies that both support (e.g., Camargo, Boucher-Lalonde, & Currie, 2018; Gardiner, Bain, Hamer, Jones, & Johnson, 2018; Melo, Sponchiado,

Cáceres, & Fahrig, 2017; Watling et al., 2020) and refute (e.g., Evju & Sverdrup-Thygeson, 2016; Haddad et al., 2017; Lindgren & Cousins, 2017) it. Although not designed to test the habitat amount hypothesis, this study joins those that do not fully support it; specifically, the presence of a significant relationship of genetic distance and diversity in *P. smintheus* with the distance between open patches, but not with total open area, indicates that the configuration of land cover is important. Here, distance among open patches and total open area are not independent (and are moderately correlated), which in theory makes it more difficult to distinguish between the effects of patch isolation and habitat area. However, because habitat amount was not a significant predictor at all, it cannot be underlying the relationship between patch distance and genetic distance and diversity. As discussed above, one caveat is that open land cover does not represent habitat per se, rather land cover over which *P. smintheus* can more easily disperse. Because open land cover is linked to dispersal rates, it follows that its configuration should be more important than its total amount, as any effect it has on genetic structure should be linked to its effects on dispersal. Performing a similar analysis on land cover containing the host plant could lead to different conclusions, and potential support for the habitat amount hypothesis, if total habitat amount were to emerge as a significant predictor of genetic structure.

4.4.3 Early-winter weather

Winter weather conditions, particularly snow depth, had an important effect on both genetic differentiation and diversity. Populations that experienced less snow cover in November (both across all years of available data and for the year of lowest observed snow depth) were more genetically differentiated from other populations (Figure 4.1). And, expected heterozygosity was lower in populations with the lowest observed mean November snow depth across all years (Figure 4.1). Lower snow cover is expected to result in higher larval mortality, through freezing or early hatching (Matter et al., 2011; Roland & Matter, 2013, 2016), and therefore more frequent or severe bottlenecks. More frequent or severe bottlenecks in turn would result in higher rates of genetic drift that would drive up differentiation from other populations and reduce local heterozygosity. Snow depth is positively associated with survival in other insect species that overwinter

as larvae. In species such as the moth *Helicoverpa armigera* higher snow cover is believed to protect larvae from cold stress, as temperature under a snow pack tends to be higher than air temperature (Huang, 2016). Theoretically, snow pack can also buffer larvae against experiencing metabolic stress from warm spells; however, the warmer average temperature under snowpack results in higher metabolic rates and, in some species, higher mortality compared to larvae overwintering above the snowpack (Irwin & Jr, 2003). The positive effect of snow depth on genetic diversity, and its negative effect on genetic differentiation, in *P. smintheus* may reflect the greater importance of buffering against freezing than any cumulative effects of higher metabolic rates under snow cover. Furthermore, the relationship between lowest mean snow depth (an extreme weather condition variable) and both genetic differentiation and heterozygosity may indicate that there is a threshold minimum snow depth below which larval mortality greatly increases and a bottleneck (or otherwise severe population decline) occurs.

Cooler November daily low temperatures were unexpectedly associated with greater genetic diversity and lower genetic distances in *P. smintheus* populations. As eggs freeze below -28 °C (Matter et al., 2011), I anticipated that cooler early-winter temperatures would drive more frequent declines in population size resulting in higher rates of genetic drift, lower genetic diversity, and greater differentiation among populations, which I did not observe. It is possible that cooler overnight temperatures prevent premature larval development and hatching. However, in that case I would also expect that cooler average temperatures (not just the minimum temperature reached overnight) would also reduce premature hatching, and there was no relationship between mean November temperature and genetic metrics. As there is no clear mechanism by which colder minimum temperatures alone should promote genetic diversity, then the apparent effect of daily minimum temperature may reflect the effect of a correlated weather variable. Specifically, cooler overnight temperatures are predicted to be associated with greater snow cover, as the insulating snow prevents heat stored in the ground during the day to be transferred to the air resulting in cooler nighttime temperatures (Mote, 2008). Indeed, the variable of mean November daily minimum temperature was significantly negatively correlated with November snow depth in the year with the lowest snowfall – and it is snowfall in these years that is predicted to drive bottlenecks and reduce genetic diversity.

In addition to snow depth in November and mean daily minimum temperatures in November, equivalent weather variables in other months (snow depth in February and mean daily minimum temperatures in July) were also significant predictors of genetic diversity and differentiation. My focus on November weather conditions was based on Roland and Matter (2016), where November temperatures (including mean, maximum, and minimum temperatures) and snowfall emerged as important predictors of *P. smintheus* population growth relative to weather in all other months. Here, I was unable to differentiate the influence of early-winter weather from other correlated weather variables. My results support the importance of temperature and over-wintering snow depth in shaping genetic patterns, likely as a result of their effects on population growth rates. It is possible that July daily minimum temperatures reflect a different process than the assumed effects of winter temperature and snow depth on egg survival; however, given the field experiments on egg survival, and the previous models of population growth, it is more likely that the correlation of daily minimum temperatures throughout the year explains the relationship between July daily minimums and genetic diversity and differentiation. The differences between these results and those of Roland and Matter (2016) likely reflect the different spatial and temporal scales of the studies, as well as the differences between modelling population growth and genetic measures. My results reflect broader spatial patterns among populations at a single point in time, whereas previous models have considered a much smaller spatial scale with 20 years of population size data (Roland & Matter, 2016). Additionally, there is often a time lag between changes in population size and changes in patterns of genetic diversity (Epps & Keyghobadi, 2015); with changes in population size being more immediately affected by weather, the relationship between specific weather variables and population size would be more easily distinguished than between weather and genetic diversity and differentiation.

4.4.4 Conclusions

I show here that both weather and landscape play a role in shaping genetic diversity and differentiation in *P. smintheus* populations. This likely reflects the importance of winter weather, especially snow depth, on larval survival and consequently on annual variation

in population size, and of connectivity of open landscape patches on *P. smintheus* dispersal. These results suggest that changing weather and climate conditions can be as important as habitat and landscape in affecting populations and their genetic diversity. In the context of conservation or re-introduction of threatened species, while land cover may be improved through intervention, weather is out of our immediate control. Nonetheless, weather patterns should be assessed when setting conservation priorities; for example, sites that experience more frequent extreme weather events may not benefit from habitat restoration if any positive effects of restoration are offset by declines due to weather. Furthermore, if the effects of specific weather variables (e.g., temperature, precipitation) on populations can be distinguished, this information may be used to prioritize sites for conservation or restoration that match optimal values of those variables.

Landscape, specifically the connectivity of open landcover patches, was also a significant predictor of *P. smintheus* genetic diversity and differentiation. As *P. smintheus* disperses more easily through open than forested land cover (Roland, Keyghobadi, & Fownes, 2000), greater connectivity of open land covers at high altitudes likely reflects greater dispersal among populations. As has been previously concluded in landscape genetic studies and is also seen here, landscape connectivity should be an important component of habitat restoration separate from the total area of restored habitat. For example, if *P. smintheus* habitat were to be conserved, it may be more beneficial for genetic diversity to preserve connectivity among habitat patches rather than fewer, larger habitat patches, especially given the role of density-independent weather factors in driving local census and effective population sizes.

4.5 Literature cited

Beichman, A. C., Huerta-Sanchez, E., & Lohmueller, K. E. (2018). Using genomic data to infer historic population dynamics of nonmodel organisms. *Annual Review of Ecology, Evolution, and Systematics*, *49*(1), 433–456.

Bohonak, A. J. (1999). Dispersal, gene flow, and population structure. *The Quarterly Review of Biology*, *74*(1), 21–45.

Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: A practical information-theoretic approach. New York, NY: Springer.

Camargo, R. X. D., Boucher-Lalonde, V., & Currie, D. J. (2018). At the landscape level, birds respond strongly to habitat amount but weakly to fragmentation. *Diversity and Distributions*, *24*(5), 629–639. \

Caplins, S. A., Gilbert, K. J., Ciotir, C., Roland, J., Matter, S. F., & Keyghobadi, N. (2014). Landscape structure and the genetic effects of a population collapse. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1796), 20141798.

Castillo, J. A., Epps, C. W., Jeffress, M. R., Ray, C., Rodhouse, T. J., & Schwalm, D. (2016). Replicated landscape genetic and network analyses reveal wide variation in functional connectivity for American pikas. *Ecological Applications*, *26*(6), 1660–1676.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140.

Chase, M. K., Nur, N., & Geupel, G. R. (2005). Effects of weather and population density on reproductive success and population dynamics in a song sparrow (*Melospiza melodia()* population: A long-term study. *The Auk*, *122*(2), 571–592.

Cushman, S. A., Shirk, A., & Landguth, E. L. (2012). Separating the effects of habitat area, fragmentation and matrix resistance on genetic differentiation in complex landscapes. *Landscape Ecology*, *27*(3), 369–380.

Dupuis, J. R., Cullingham, C. I., Nielsen, S. E., & Sperling, F. A. H. (2019). Environmental effects on gene flow in a species complex of vagile, hilltopping butterflies. *Biological Journal of the Linnean Society*, *127*(2), 417–428.

Epps, C. W., & Keyghobadi, N. (2015). Landscape genetics in a changing world: Disentangling historical and contemporary influences and inferring change. *Molecular Ecology*, *24*(24), 6021–6040.

Evju, M., & Sverdrup-Thygeson, A. (2016). Spatial configuration matters: A test of the habitat amount hypothesis for plants in calcareous grasslands. *Landscape Ecology*, *31*(9), 1891–1902.

Frankham, R. (1995). Effective population size/adult population size ratios in wildlife: A review. *Genetics Research*, *66*(2), 95–107.

Funk, W. C., Blouin, M. S., Corn, P. S., Maxell, B. A., Pilliod, D. S., Amish, S., & Allendorf, F. W. (2005). Population structure of Columbia spotted frogs (*Rana luteiventris*) is strongly affected by the landscape. *Molecular Ecology*, *14*(2), 483–496.

Gardiner, R., Bain, G., Hamer, R., Jones, M. E., & Johnson, C. N. (2018). Habitat amount and quality, not patch size, determine persistence of a woodland-dependent mammal in an agricultural landscape. *Landscape Ecology*, *33*(11), 1837–1849.

Guppy, C. S., & Shepard, J. (2001). *Butterflies of British Columbia including western Alberta, southern Yukon, the Alaska Panhandle, Washington, northern Oregon, northern Idaho, northwestern Montana*. Vancouver [B.C.]: UBC Press.

Haddad, N. M., Gonzalez, A., Brudvig, L. A., Burt, M. A., Levey, D. J., & Damschen, E. I. (2017). Experimental evidence does not support the Habitat Amount Hypothesis. *Ecography*, *40*(1), 48–55.

Halkka, A., Halkka, L., Halkka, O., Roukka, K., & Pokki, J. (2006). Lagged effects of North Atlantic Oscillation on spittlebug *Philaenus spumarius* (Homoptera) abundance and survival. *Global Change Biology*, *12*(12), 2250–2262.

Hanski, I., & Ovaskainen, O. (2000). The metapopulation capacity of a fragmented landscape. *Nature*, *404*(6779), 755–758.

Henein, K., & Merriam, G. (1990). The elements of connectivity where corridor quality is variable. *Landscape Ecology*, *4*(2), 157–170.

Herzog, R., & Hadrys, H. (2017). Long-term genetic monitoring of a riverine dragonfly, *Orthetrum coerulescens* (Odonata: Libellulidae]: Direct anthropogenic impact versus climate change effects. *PLOS ONE*, *12*(5), e0178014.

Huang, J. (2016). Effects of soil temperature and snow cover on the mortality of overwintering pupae of the cotton bollworm, *Helicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae). *International Journal of Biometeorology*, *60*(7), 977–989.

Hutchinson, M. F., McKenney, D. W., Lawrence, K., Pedlar, J. H., Hopkinson, R. F., Milewska, E., & Papadopol, P. (2009). Development and testing of Canada-wide interpolated spatial models of daily minimum–maximum temperature and precipitation for 1961–2003. *Journal of Applied Meteorology and Climatology*, *48*(4), 725–741.

Irwin, J. T., & Lee, R. E. (2003). Cold winter microenvironments conserve energy and improve overwintering survival and potential fecundity of the goldenrod gall fly, *Eurosta solidaginis*. *Oikos*, *100*(1), 71–78.

Jangjoo, M., Matter, S. F., Roland, J., & Keyghobadi, N. (2016). Connectivity rescues genetic diversity after a demographic bottleneck in a butterfly population network. *Proceedings of the National Academy of Sciences*, *113*(39), 10914–10919.

Kerr, J. T., & Ostrovsky, M. (2003). From space to species: Ecological applications for remote sensing. *Trends in Ecology & Evolution*, *18*(6), 299–305.

Keyghobadi, N., Roland, J., & Strobeck, C. (1999). Influence of landscape on the population genetic structure of the alpine butterfly *Parnassius smintheus* (Papilionidae). *Molecular Ecology*, *8*(9), 1481–1495.

Keyghobadi, N., Roland, J., & Strobeck, C. (2005). Genetic differentiation and gene flow among populations of the alpine butterfly, *Parnassius smintheus*, vary with landscape connectivity. *Molecular Ecology*, *14*(7), 1897–1909.

Kuussaari, M., Rytteri, S., Heikkinen, R. K., Heliölä, J., & von Bagh, P. (2016). Weather explains high annual variation in butterfly dispersal. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1835), 20160413.

Lande, R. (1988). Genetics and demography in biological conservation. *Science*, *241*(4872), 1455–1460.

Lewellen, R. H., & Vessey, S. H. (1998). The Effect of Density Dependence and Weather on Population Size of a Polyvoltine Species. *Ecological Monographs*, *68*(4), 571–594.

Lindgren, J. P., & Cousins, S. A. O. (2017). Island biogeography theory outweighs habitat amount hypothesis in predicting plant species richness in small grassland remnants. *Landscape Ecology*, *32*(9), 1895–1906.

Livolsi, M. C., Williams, C. K., Coluccy, J. M., & Dibona, M. T. (2021). The effect of sea level rise on dabbling duck energetic carrying capacity. *The Journal of Wildlife Management*, *85*(4), 686–695.

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

Matter, S. F., Doyle, A., Illerbrun, K., Wheeler, J., & Roland, J. (2011). An assessment of direct and indirect effects of climate change for populations of the Rocky Mountain Apollo butterfly (*Parnassius smintheus* Doubleday). *Insect Science*, *18*(4), 385–392.

Meirmans, P. G., & Hedrick, P. W. (2011). Assessing population structure: F_{ST} and related measures. *Molecular Ecology Resources*, *11*(1), 5–18.

Melo, G. L., Sponchiado, J., Cáceres, N. C., & Fahrig, L. (2017). Testing the habitat amount hypothesis for South American small mammals. *Biological Conservation*, *209*, 304–314.

Mote, T. L. (2008). On the role of snow cover in depressing air temperature. *Journal of Applied Meteorology and Climatology*, *47*(7), 2008–2022.

Nei, M. (1972). Genetic distance between populations. *The American Naturalist*, *106*(949), 283–292.

Nei, M., Maruyama, T., & Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations. *Evolution*, *29*(1), 1–10.

Paradis, E. (2010). pegas: An R package for population genetics with an integrated–modular approach. *Bioinformatics*, *26*(3), 419–420.

Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics (Oxford, England)*, *35*(3), 526–528.

Perea, S., & Doadrio, I. (2015). Phylogeography, historical demography and habitat suitability modelling of freshwater fishes inhabiting seasonally fluctuating Mediterranean river systems: A case study using the Iberian cyprinid *Squalius valentinus*. *Molecular Ecology*, *24*(14), 3706–3722.

Pérez-Espona, S., McLeod, J. E., & Franks, N. R. (2012). Landscape genetics of a top neotropical predator. *Molecular Ecology*, *21*(24), 5969–5985.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & Deepayan, S. (2015). *nlme: Linear and Nonlinear Mixed Effects Models*.

Preisler, H. K., Hicke, J. A., Ager, A. A., & Hayes, J. L. (2012). Climate and weather influences on spatial temporal patterns of mountain pine beetle populations in Washington and Oregon. *Ecology*, *93*(11), 2421–2434.

Proft, K. M., Bateman, B. L., Johnson, C. N., Jones, M. E., Pauza, M., & Burridge, C. P. (2021). The effects of weather variability on patterns of genetic diversity in Tasmanian bettongs. *Molecular Ecology*, *30*(8), 1777–1790.

R Core Team. (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rašić, G., Filipović, I., Weeks, A. R., & Hoffmann, A. A. (2014). Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, *Aedes aegypti*. *BMC Genomics*, *15,* 275.

Robert, A. (2009). The effects of spatially correlated perturbations and habitat configuration on metapopulation persistence. *Oikos*, *118*(10), 1590–1600.

Roland, J., Keyghobadi, N., & Fownes, S. (2000). Alpine parnassius butterfly dispersal: Effects of landscape and population size. *Ecology*, *81*(6), 1642–1653.

Roland, J., & Matter, S. F. (2013). Variability in winter climate and winter extremes reduces population growth of an alpine butterfly. *Ecology*, *94*(1), 190–199.

Roland, J., & Matter, S. F. (2016). Pivotal effect of early-winter temperatures and snowfall on population growth of alpine *Parnassius smintheus* butterflies. *Ecological Monographs*, *86*(4), 412–428.

Roy, D. B., Rothery, P., Moss, D., Pollard, E., & Thomas, J. A. (2001). Butterfly numbers and weather: Predicting historical trends in abundance and the future effects of climate change. *Journal of Animal Ecology*, *70*(2), 201–217.

Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science*, *236*(4803), 787–792.

Solbreck, C. (1991). Unusual weather and insect population dynamics: *Lygaeus equestris* during an extinction and recovery period. *Oikos*, *60*(3), 343–350. JSTOR.

Store, R., & Jokimäki, J. (2003). A GIS-based multi-scale approach to habitat suitability modeling. *Ecological Modelling*, *169*(1), 1–15.

Vucetich, J. A., Waite, T. A., & Nunney, L. (1997). Fluctuating population size and the ratio of effective to census population size. *Evolution*, *51*(6), 2017–2021. JSTOR.

Wang, I. J., Johnson, J. R., Johnson, B. B., & Shaffer, H. B. (2011). Effective population size is strongly correlated with breeding pond size in the endangered California tiger salamander, *Ambystoma californiense*. *Conservation Genetics*, *12*(4), 911–920.

Watling, J. I., Arroyo-Rodríguez, V., Pfeifer, M., Baeten, L., Banks-Leite, C., Cisneros, L. M., … Fahrig, L. (2020). Support for the habitat amount hypothesis from a global synthesis of species density studies. *Ecology Letters*, *23*(4), 674–681.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, *16*(2), 97–159.

Wright, S. (1943). Isolation by distance. *Genetics*, *28*(2), 114–138.

Zhang, W., Lu, Y., van der Werf, W., Huang, J., Wu, F., Zhou, K., … Rosegrant, M. W. (2018). Multidecadal, county-level analysis of the effects of land use, Bt cotton, and weather on cotton pests in China. *Proceedings of the National Academy of Sciences*, *115*(33), E7700–E7709.

Chapter 5

5 Repeated bottlenecks in a butterfly population network temporarily disrupt patterns of genomic diversity and differentiation

5.1 Introduction

Natural populations experience fluctuations in population size due to both density dependent and independent factors (Nicholson, 1954). A bottleneck is said to occur when there is a sudden and extreme reduction in population size, for example due to increased mortality or the founding of a new population from a small number of individuals (Nei, Maruyama, & Chakraborty, 1975). Bottlenecks are expected to result in a loss of genetic diversity. First, a loss of genetic diversity (specifically allelic richness) is expected to occur immediately after a bottleneck as only alleles retained by the surviving individuals remain in the population (Allendorf, 1986). Until the population grows in size and recovers from the bottleneck, genetic diversity will continue to be lost at a higher rate given the stronger effects of genetic drift in small populations (Nei et al., 1975). Not all populations are isolated, however, and when migrants are exchanged bottlenecks can also affect spatial patterns of genetic differentiation (e.g., isolation by distance; IBD) in the network of connected populations. Populations are expected to become more differentiated immediately after a bottleneck as different alleles are lost by chance in each population, unrelated to their location and connectivity to other populations (Chakraborty & Nei, 1977).

When testing these predictions in natural populations, different approaches have been taken depending on when the sampling occurs relative to the occurrence of the bottleneck. When sampling for genetic data occurs significantly later than a demographic bottleneck is known or hypothesized to have occurred, studies may focus on confirming or refuting historical bottlenecks and describing long-term population trends (e.g. Gattepaille, Jakobsson, & Blum, 2013; Pilot et al., 2014; Stoffel et al., 2018). When a bottleneck is known to have recently occurred, or there is sampling during and after the bottleneck, questions of how genetic diversity and differentiation are changing can be

more directly addressed. In some of these cases, the expected loss of allelic richness (Fauvelot, Cleary, & Menken, 2006), increase in differentiation (Kekkonen, Hanski, Jensen, Väisänen, & Brommer, 2011), and loss of population structure has been observed. However, severe declines in population size do not always result in decreased genetic diversity (Le Gouar et al., 2009; Suárez, Betancor, Fregel, Rodríguez, & Pestano, 2012) or increased differentiation among populations; it is therefore important to empirically demonstrate the genetic consequences of a demographic bottleneck instead of relying solely on theoretical expectations. In some systems bottlenecks may not result in loss of diversity because immigration restores allelic diversity post-bottleneck. Even in systems where there is known to be no or very low levels of immigration, a bottleneck may not affect the measured allelic richness if the population size recovers quickly and the effects of the bottleneck are limited to any alleles lost in a single round of high mortality (Le Gouar et al., 2009; Suárez et al., 2012).

In addition to their effects on overall genetic diversity and population structure, bottlenecks can affect genetic diversity at expressed, functional loci (i.e., non-neutral loci) in several ways. In the absence of selection, bottlenecks reduce functional genetic diversity in the same way that neutral diversity is lost (i.e., through the effect of drift). This occurs when the genotype at a functional locus does not affect whether an individual survives and reproduces in whatever conditions caused the bottleneck (whether the bottleneck occurred because of increased mortality or through a founder event). The reduction of genetic diversity at specific functional loci is often of concern in the conservation of populations at risk. A broad reduction in functional genetic diversity is predicted to reduce the evolutionary potential of a population (Frankham et al., 1999), and the loss of diversity both genome-wide (Reed & Frankham, 2003) and at specific loci (e.g., the major histocompatibility complex; Agudo et al., 2012; Wegner, Kalbe, Milinski, & Reusch, 2008) is associated with lower fitness. Studies concerned with specific loci often focus on whether balancing (e.g., Gos, Slotte, & Wright, 2012; Sutton, Nakagawa, Robertson, & Jamieson, 2011) or directional (e.g., Windig, Veerkamp, & Nylin, 2004) selection pressures unrelated to the cause of a bottleneck can maintain the presence of minor alleles despite genetic drift. In addition to allele frequencies at functional loci changing via drift, they may also change via selection if certain phenotypes are better

able to survive the conditions that caused the bottleneck. If the bottleneck is driven by a factor that acts relatively suddenly (e.g., extreme weather, disease outbreak), the rapid loss of less fit individuals should be associated with abrupt changes in allele frequencies. This pattern is similar to the artificial selection in animal and crop domestication, where population size is repeatedly reduced according to which individuals carry a desired phenotype.

Many studies of genetic change at functional loci during bottlenecks are concerned with the implications of lost genetic diversity on population survival and evolutionary potential, whereas the actual selection pressures that may be present during the bottleneck are less frequently considered (Frankham, 2005; Ørsted, Hoffmann, Sverrisdóttir, Nielsen, & Kristensen, 2019). This may be due to the stochastic nature of many bottlenecks, where disease outbreaks or extreme weather conditions are considered to be unusual, one-time occurrences and any novel selective pressures would be of little long-term interest. However, there are prominent examples that come from studying changes in the morphology of birds before and after extreme weather events, although these studies typically do not try to establish that a bottleneck had occurred in the strict genetic sense (but rather are based on direct observation of an extreme decline in population size). Bumpus (1899) famously presented hypotheses regarding the survival of house sparrows (*Passer domesticus*) after a winter storm; further analyses of his observations provide support for sex-based selection for larger body size in males and intermediate size in females (Johnston, Niles, & Rohwer, 1972). A drought on Isla Genovesa in the Galápagos drove down population numbers of the cactus finch *Geospiza conirostri* and selected for individuals with beaks shaped appropriately for foraging on arthropods (Grant & Grant, 1989). Long-term population monitoring has been crucial to the study of selection on finches in the Galápagos, as it provides data from before and after extreme weather events and allows perturbations in population size to be detected (Boag & Grant, 1981; Grant & Grant, 1989).

Genetic changes during bottlenecks have historically been characterized using both non-functional and functional loci. Early work used allozymes to assess the loss of allelic richness versus others measures of diversity such as heterozygosity, finding, as predicted,

that allelic richness was lost more easily and was a more reliable genetic indicator of a recent bottleneck (Leberg, 1992). As microsatellites were discovered and developed as genetic markers, their greater variability gave them more power to detect the loss of allelic richness and the change in the distribution of minor allele frequencies (i.e., a shift to more loci with intermediate frequencies as rare alleles are lost) (Luikart, Sherwin, Steele, & Allendorf, 1998). Targeted sequencing of functional loci was combined with microsatellite genotyping to compare changes in functional allele frequencies to a non-functional (i.e., neutral) baseline (e.g., Oliver & Piertney, 2012; Sutton et al., 2011).

Single nucleotide polymorphisms have been increasingly used in population genetic studies in general as an alternative to microsatellites as new sequencing and genotyping technologies have been developed. Initially relatively few SNPs (<100) were used in population genetic studies; as whole genome sequencing and related techniques that subsampled the genome (e.g., restriction site associated DNA sequencing) were developed, it became common to see large SNP (>1000-10 000) datasets being used to answer population genetic questions (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016). Large, genome-wide panels of single nucleotide polymorphisms are of particular interest when disentangling long-term demographic history, as patterns of linkage allow the identification of bottlenecks that occurred too long ago for any changes to allelic richness to be detectable (Achaz, 2009; Ramos-Onsins & Rozas, 2002). However, whole genome sequencing remains expensive, and in some cases low quantity or quality DNA prevents high quality RAD-seq from succeeding. Smaller panels of SNPs are again being developed for either higher throughput analysis (Jardine et al., 2016; Tabuloc et al., 2019), or for use with poorer quality samples (Natesh et al., 2019) in population genetic studies. As these smaller panels are used as a replacement for microsatellites in assessing genetic diversity, they should also be useful in the study of genetic changes during bottlenecks. Compared to largely neutral microsatellites, these SNP panels have the added benefit of assessing changes in allele frequency of functional loci in addition to overall changes in genetic diversity, in a single genotyping run.

Smaller SNP panels are often developed from a larger existing dataset, such as from whole genome sequencing or RAD-sequencing. Criteria for narrowing down typically

thousands of potential SNPs to a panel of a few hundred depend on the anticipated uses of the panel. In some cases, SNP panels are intended to answer a specific question, such as diagnosing the population of origin of an individual (e.g., Muñoz et al., 2015); in this case, SNPs are chosen to be highly informative at distinguishing populations. In other cases, SNP panels are intended to be a replacement for larger RAD-seq datasets, which are assumed to represent both neutral and non-neutral variability at randomly distributed locations across the genome. To develop a representative SNP panel, criteria include ensuring that loci have a sufficiently high minor allele frequency that the anticipated sample size will display polymorphism, and that loci are not statistically linked. Including a set of known functional loci (if a transcriptome is available to identify expressed sequences) is useful when both neutral and selective processes are under study. Selecting functional loci can be difficult in the absence of candidate loci or a specific hypothesis about how selection is acting. Choosing loci that contribute to a variety of cellular functions gives the most flexibility for future study and, arguably, better represents the genome (an important consideration if the functional loci will also be included in general population genetic analyses).

To study the effect of fluctuating population size over time on genetic diversity and differentiation, I moved from studying *Parnassius smintheus* populations separated by tens of kilometers to a population network located on a single ridge in the East Kananaskis region, where populations are separated by a maximum of ~9 km. Populations on Jumpingpound Ridge have been studied since 1995, using both genetic and mark-recapture data. An index of adult population size in each meadow each year has been estimated using Craig's method (Craig, 1953). During this time, two network-wide bottleneck events have been observed, where the index of adult population size across the network declined by 60-100% from one year to the next (Caplins et al., 2014), but subsequently increased again within one to two years. The first population bottleneck occurred in 2003 and the second spanned 2010 to 2011, with population sizes remaining very low in both years. The mechanism underlying these bottlenecks has not been definitively proven, but there is evidence that unfavourable early winter conditions can result in high egg mortality (Roland & Matter, 2013, 2016). Specifically, warm early winter conditions resulting in low snow cover, as well as extreme cold snaps before

sufficient snow cover has accumulated, are both predicted to cause egg mortality by freezing (Roland & Matter, 2013, 2016). Very low population numbers were observed again in 2019, although it is unclear yet whether this was driven by early winter conditions or unusual spring weather. This system allows for the study of changes in both neutral genetic patterns (e.g., genetic diversity and spatial genetic structure) and potentially adaptive genetic variation over two periods of population size collapse and recovery. Key to the study of adaptive change in this system is that the mechanism driving collapse is likely consistent – overwintering egg mortality as a result of weather conditions – such that selective pressures should be similar during both collapses.

Previous work on the Jumpingpound Ridge bottlenecks has used microsatellites to track changes in allelic richness and patterns of IBD across those events. Populations tended to lose allelic richness only after the second, more protracted bottleneck, and the rapid recovery of allelic richness after that bottleneck was mediated by connectivity to neighbouring populations (i.e., populations with greater potential for immigration recovered allelic richness more quickly) (Caplins et al., 2014; Jangjoo, Matter, Roland, & Keyghobadi, 2016). Patterns of IBD broke down during both bottlenecks but were present before the first bottleneck and were quickly restored before the second bottleneck. Here I use SNPs to assess how genetic diversity, including at functional loci, changes over these two bottlenecks in *P. smintheus* populations on Jumpingpound Ridge. I develop a moderate sized SNP panel suitable for genotyping DNA extracted from wing clips, using a selection of loci initially sequenced in a RADseq library (see Chapter 2). I expect that some alleles will be lost by chance as population size drops during bottlenecks, so that overall (functional and non-functional) allelic richness will be lower in years immediately after bottlenecks. Populations will lose different alleles by chance, so patterns of IBD that are present prior to the bottlenecks will be lost as populations become more differentiated regardless of their location or connectivity. As immigration reintroduces alleles, I expect that allelic richness will increase several generations after the bottleneck and that patterns of IBD will be reestablished. I also expect that a small number of functional loci may consistently oscillate in allele frequency over bottlenecks as a result of fluctuating selection pressures, and that their function (or the function of

loci linked to them by their proximity in the genome) is related to surviving the overwintering conditions that drive up mortality and underlie the bottleneck events.

5.2 Methods

5.2.1 Sampling location

Parnassius smintheus populations on Jumpingpound Ridge, Alberta, have been monitored using mark-recapture since 1995 (Goff, Yerke, Keyghobadi, & Matter, 2018; Keyghobadi, Roland, & Strobeck, 1999; Roland, Keyghobadi, & Fownes, 2000). A population has been defined as all individuals sampled within a distinct meadow along the ridgeline, where such meadows are often but not always separated by non-habitat forest matrix (Figure 5.1). While true census population sizes are not known, Craig's method (Craig, 1953) is used to provide indices of population size in each meadow, and to ascertain how populations sizes are changing from year to year (Matter, Keyghobadi & Roland 2014). These indices show that population size fluctuates temporally and that these fluctuations are not necessarily synchronous among the different populations, except for in the years of the network-wide bottlenecks (Figure 5.2).

The centroids of meadows are separated by as few as several hundred meters, to as many as nine kilometers for the most distant pair of populations examined here, as measured along the ridge-top. Individuals are observed dispersing along the top of the ridge and are not expected to frequently move through the lower elevation forest on the sides of the ridge, so the distances between pairs of meadows are measured along the ridgeline (Keyghobadi et al., 1999). Individuals are occasionally observed dispersing between neighbouring meadows, and more rarely between non-neighbouring meadows and meadows separated by forest (Roland et al., 2000).

5.2.2 Sample collection

Populations are surveyed by mark-recapture during the adult flight season, which is typically a five to six week period in July to early August. Meadows are typically visited three to four times each season; individuals are hand netted and each is marked with a

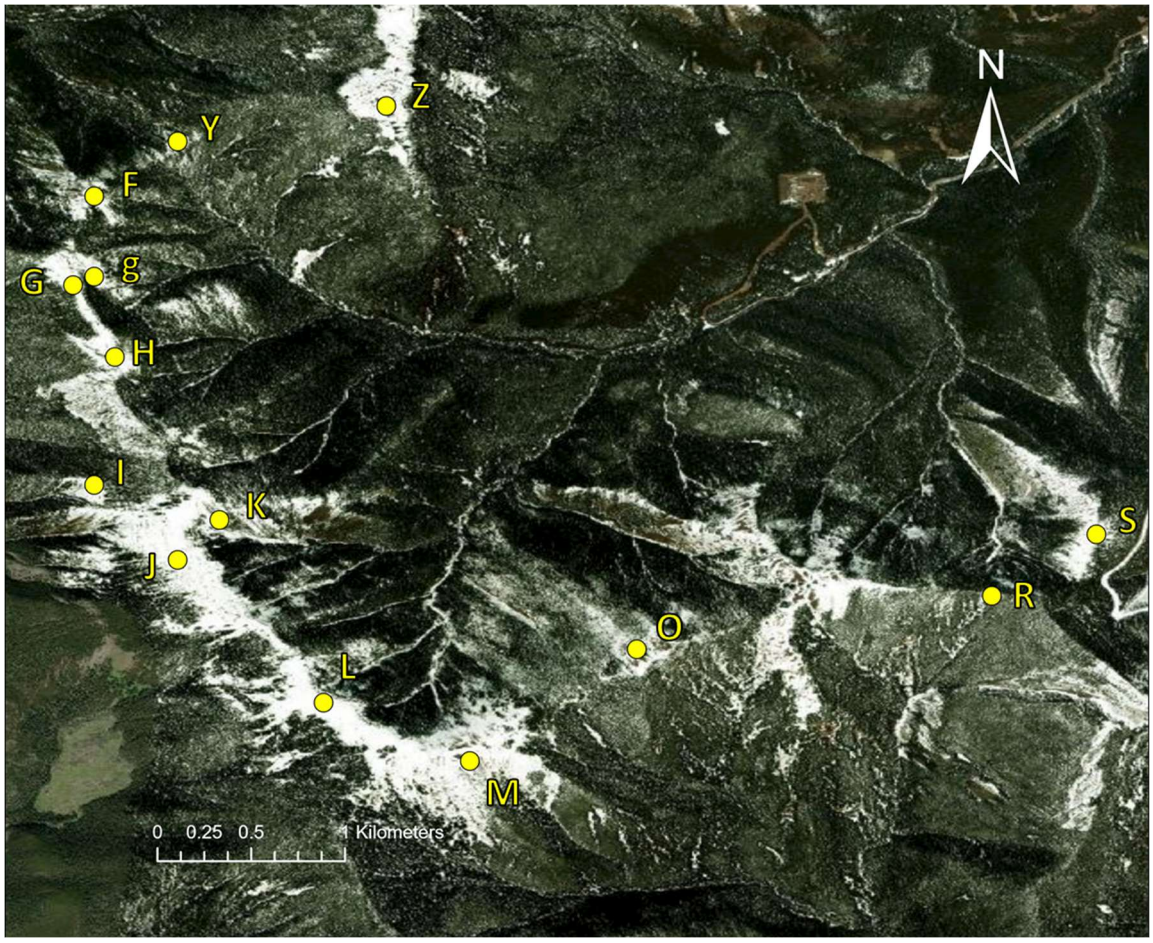

Figure 5.1 *Parnassius smintheus* individuals were sampled from 14 meadows on Jumpingpound Ridge. Light coloured areas indicate unforested areas, including meadows above tree-line. Each sample site is identified by a letter, and yellow points indicate the centroids of the sampled meadows. Map data: Maxar Technologies.

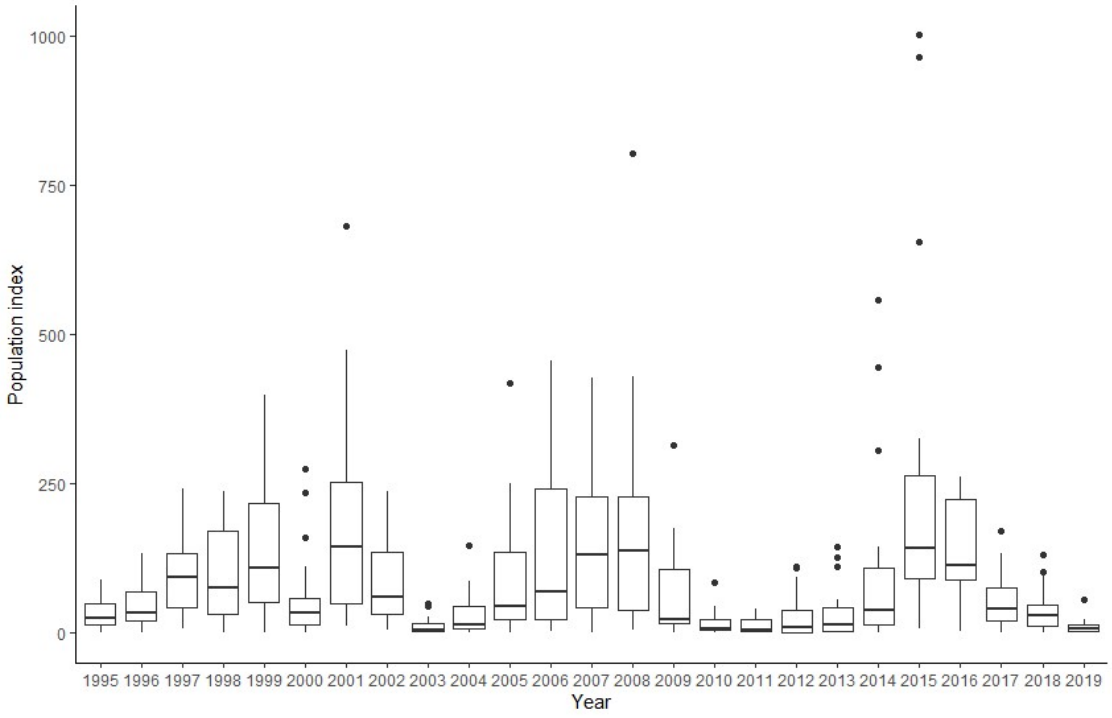

Figure 5.2 *Parnassius smintheus* population size indices for 17 populations on Jumpingpound Ridge, Alberta across years. Some populations represented here were excluded from my genetic analysis due to low sample size. An index of the size of each population was estimated using Craig's method based on mark-recapture data for all years except for 1997-2000, where population size was estimated based on Pollard transects and converted to be comparable to Craig's method estimates (as in Roland & Matter, 2013). Boxes show central 50% of values and the median, across populations. Tails represent values within 1.5 times the interquartile range, and points represent values outside 1.5 times the interquartile range. Data are from S. Matter (pers. comm.)

unique ID code indicating the population of origin. From approximately every third individual marked, a wing clip (small piece of wing tissue, approximately 0.25 cm^2) is sampled (Roland et al., 2000) as a source of DNA. I used wing clips collected in 1995, 2005, 2008, and 2013, from 16 *P. smintheus* populations on Jumpingpound Ridge. The first three years bracket a population bottleneck that occurred in 2003, including a pre-bottleneck year (1995), an immediate post-bottleneck year (2005), and a recovery year (2008). The years 2008 and 2013, in turn, bracket the second population bottleneck that began in 2010 and extended into 2011. Some populations had a low number of sampled individuals available for genotyping in some years; I retained only those samples with five or more individuals (see Chapter 2 for the effects of low sample size on population genetic analyses).

5.2.3 SNP panel development and genotyping

I extracted DNA from wing clips using a DNeasy Blood and Tissue Kit (Qiagen, Germantown, MD), and genotyped DNA at 171 SNP loci on a MassARRAY iPlex Gold platform (Sequenom, CITY). The iPlex platform uses a PCR reaction to first amplify the region flanking each SNP (~40-60 bp in each direction). After PCR, primers bind directly adjacent to the SNP of interest and are extended with mass-modified dideoxynucleotides. Different alleles are differentiated by the weight differences from the mass-modified nucleotide added at the polymorphic site, as detected using MALDI-TOF mass spectrometry.

I developed the 171 SNP panel from an initial restriction site associated DNA sequencing (RADseq) library (see Chapter 1). For iPlex primers to bind, only RADseq loci with at least 30bp of flanking region around a SNP, and without a second SNP in the flanking region, could be used. I filtered loci that met these conditions for statistical linkage, using the software PLINK to remove the fewest SNPs necessary to generate an unlinked SNP dataset. I used Magic-BLAST (Boratyn, Thierry-Mieg, Thierry-Mieg, Busby, & Madden, 2019) to match these loci against a transcriptome that was previously generated from RNA isolated from the thorax of adult butterflies captured during flight (Jangjoo, 2018). I defined loci that had at least a 90% match as expressed, functional loci. I identified Putative functions for functional loci using Trinotate (Bryant et al., 2017) (Table B1). I

chose functional loci for the final SNP panel to represent a range of cellular functions, in rough proportion to their representation among the available functional loci. Non-functional loci included all loci that had less than 10% match against the transcriptome, but also a greater than 90% percent match against a *P. smintheus* short read whole genome library (Allio et al., 2020). It is possible that some loci classified as "non-functional" using these criteria may in fact be expressed in body tissues outside of the thorax, or at an earlier life stage.

In addition to the functional loci identified from the RADseq library, I included six variable sites from the phosphoglucose isomerase (*Pgi*) locus identified by Jangjoo et al (2019). Phosphoglucose isomerase is a metabolic enzyme responsible for one of the early steps in glycolysis, the conversion of glucose-6-phosphate to fructose-6-phosphate. Phosphoglucose isomerase genotype is associated with phenotypic variation in a number of arthropods, including movement speed and thermal stress response in the beetle *Chrysomela aeneicollis* (Rank, Bruce, McMillan, Barclay, & Dahlhoff, 2007) and mating success, flight performance, and lifespan in several Colias butterflies (W. B. Watt, Wheat, Meyer, & Martin, 2003; Ward B. Watt, 1977; Ward B. Watt, Carter, & Blower, 1985). In *Melitaea cinxia*, the Glanville fritillary, *Pgi* genotype is a predictor of flight metabolism and dispersal ability (Mitikka & Hanski, 2010; Niitepõld et al., 2009). All six *Pgi* SNPs included here are non-synonymous. Two of the *Pgi* SNPs (1018 and 1129) were included because preliminary analyses showed some fluctuation in allele frequency with population bottlenecks (Jangjoo, pers. comm.). SNP 1018 codes for a non-polar amino acid at the major allele (Ala) and polar amino acid at the minor allele (Thr), while SNP 1129 codes for polar amino acids at both alleles (Ser and Thr). I also included two SNPs that had variable sites that translated to a polar amino acid at one allele and a non-polar amino acid at the other (626: Phe and Ser; 1241: Gln and Leu), as well as two *Pgi* SNPs that each had a variable site that coded for either both polar (28: Asp and Tyr) or both non-polar (1612: Val and Ile) amino acids at the alternate alleles (Jangjoo, 2018). The final SNP panel was composed of six *Pgi* loci, 35 functional loci, and 130 putatively non-functional loci.

5.2.4 Neutral population genetic analyses

I examined how measures of genetic diversity and differentiation change across the cycles of population bottlenecks. I used all SNPs to estimate these basic population genetic parameters.

I calculated the proportion of SNPs out of Hardy-Weinberg equilibrium in each population per year using the hw.test function in the pegas (Paradis, 2010) package in the statistical software R (R Core Team, 2017). I calculated allelic richness per locus in each population each year using the allelic.richness function in the R package hierfstat (Goudet, 2005). I rarified allelic richness to 10 alleles, as the smallest sample size per population was 5 diploid individuals. I averaged allelic richness across all loci to get a population estimate of mean allelic richness. I calculated expected heterozygosity per population per year using the Hs function in the adegenet package (Jombart, 2008). I examined whether allelic richness and expected heterozygosity differed significantly among years using linear mixed effects models implemented in the package nlme (Pinheiro et al., 2015), with year as a predictor and population as a random effect. I used contrasts (implemented in the R package lsmeans; Lenth, 2016) to compare allelic richness and expected heterozygosity between all pairs of years, as well as between pre-bottleneck and post-bottleneck years combined. I used a Bonferroni correction to account for multiple comparisons.

I calculated global genetic differentiation among all populations in each year as Weir and Cockerham's estimate of global F_{ST} using the wc function in the R package hierfstat. Genetic differentiation between population pairs was estimated as Nei's corrected genetic distance (Nei, 1978) using the R package gstudio (Dyer, 2014). In this system, pairwise genetic distance has previously been measured with microsatellite data using both Nei's genetic distance (Keyghobadi et al., 1999) and pairwise F_{ST} (Caplins et al., 2014); for this SNP dataset, Nei's genetic distance had a stronger linear relationship to geographic distance and so was used. I tested the significance of the relationship between genetic distance and geographic distance along the ridgeline (i.e., IBD) using two complementary statistical methods: Mantel tests and maximum likelihood population effects (MLPE) models (Clarke, Rothery, & Raybould, 2002; Mantel, 1967). I implemented Mantel tests

using the mantel function from the package vegan (Oksanen et al., 2019), using the Spearman correlation method. The Mantel statistic represents the correlation between two pairwise distance matrices (here, a matrix of pairwise distances between meadow centroids, and a matrix of pairwise Nei's D between populations), and its statistical significance is assessed via matrix permutation. For MLPE models, I used the R package nlme (Pinheiro et al., 2015) to run generalized least square models fitting Nei's genetic distance to geographic distance between sites; the model included a correlation structure accounting for the pairwise nature of my data, as implemented in the R package corMLPE (Pope, 2018).

5.2.5 Analyses of signatures of selection

Population bottlenecks may be associated with temporally fluctuating selection. When environmental conditions change across a bottleneck (either as the driver of the bottleneck or in response to it), selection pressures will also change. I examined allele frequencies of all loci, including both functional and putatively non-functional, across all four sampled years for signatures of selection. Putatively non-functional loci were included because they may be unexpressed but still under selection (e.g., a promotor region), or they may be linked to nearby loci that are under selection. Furthermore, although loci defined as non-functional are not found in the available *P. smintheus* transcriptome, the transcriptome is only of adults caught in flight during the day. It is still possible that loci classified as non-functional could be expressed (and experiencing selection) at other life stages, or under other conditions. As evidence of fluctuating selection, I looked for fluctuations in allele frequencies, which would be expected if divergent selection pressures are experienced in population collapse years versus recovery/stable years. I also looked for signatures of directional selection as loci whose minor allele frequency either consistently increased or decreased across the entire sampling period (i.e., from 1995 to 2013).

I used linear mixed effect models to identify loci whose minor allele frequency differed significantly and consistently between pre-bottleneck and post-bottleneck years, indicating possible fluctuating selection. The minor allele frequency was calculated across all individuals, separately for each population each year. I linearized minor allele

frequency at the level of the population using a logit transformation of: minor allele frequency + 0.001. The offset of 0.001 was necessary because some populations had minor allele frequencies of zero (where the minor allele was not seen) or one (where only the minor allele was seen). The minor allele was defined globally, as the less common allele across all genotyped individuals (pooled across populations), so that in some populations the globally defined minor allele was actually the more common allele. I used pre-bottleneck (1995 and 2008) versus post-bottleneck (2005 and 2013) as a single factor with two levels to predict the transformed minor allele frequency. Population was included as a random factor.

I also used linear mixed effect models to identify loci whose minor allele frequency showed a significant increasing or decreasing trend from 1995 to 2013, indicating potential directional selection. In these analyses, year was coded as a numerical value, with the first year of sampling as the origin (i.e., from 0 in 1995 to 18 in 2013). Allele frequency was again logit transformed with an offset of 0.001, and population was included as a random factor.

For both analyses, I considered loci to be potentially experiencing fluctuating or directional selection if the p-value for the predictor term was less than 0.05. Data for loci that were identified as such were then plotted as boxplots and examined for consistent trends. Loci where an oscillating or directional trend was observed over at least three of the four sampled years were retained; this eliminated loci where a single outlying year drove significance in the linear models.

5.3 Results

5.3.1 SNP dataset

After excluding populations with fewer than five individuals, there were between seven and 12 populations used for analysis in each year (1995:11, 2005:12, 2008:11, 2013:7) for a total of 567 individuals (Table 5.1). Neutral population genetic analyses were conducted using the data from all successfully genotyped SNPs, while analyses for signatures of selection were conducted using only those SNPs that were successfully genotyped and polymorphic in the datasets across all four years, for a total of 144 SNPs.

Table 5.1 Sample sizes and basic population genetic statistics (N: sample size, AR: allelic richness, H_E expected heterozygosity, HWE: proportion of loci out of Hardy-Weinberg equilibrium) for 14 populations of *Parnassius smintheus*, characterized by a panel of 144 SNPs, in each of four different years. Two years (1995 and 2008) were each before an observed demographic bottleneck, and two years (2005 and 2013) were each after an observed demographic bottleneck.

Pop	N - 1995	N - 2005	N - 2008	N - 2013	AR 1995	AR 2005	AR 2008	AR 2013	H_E 1995	H_E 2005	H_E 2008	H_E 2013	HWE 1995	HWE 2005	HWE 2008	HWE 2013
F	14	9	17	-	1.66	1.64	1.69	-	0.21	0.20	0.23	-	0.05	0.11	0.07	-
g	15	-	17	12	1.67	-	1.69	1.66	0.22	-	0.22	0.21	0.09	-	0.08	0.10
G	-	20	16	13	-	1.65	1.68	1.62	-	0.21	0.22	0.23	-	0.11	0.09	0.08
H	-	6	14	-	-	1.74	1.71	-	-	0.20	0.23	-	-	0.09	0.12	-
I	7	9	-	11	1.70	1.65	-	1.62	0.21	0.22	-	0.21	0.07	0.06	-	0.06
J	10	12	18	20	1.67	1.69	1.68	1.64	0.22	0.22	0.22	0.21	0.06	0.13	0.16	0.09
K	15	8	18	19	1.67	1.63	1.67	1.64	0.22	0.18	0.22	0.21	0.14	0.04	0.09	0.16
L	15	16	18	14	1.71	1.68	1.68	1.67	0.23	0.21	0.22	0.23	0.11	0.19	0.12	0.09
M	15	15	18	39	1.68	1.68	1.69	1.70	0.22	0.22	0.23	0.23	0.10	0.09	0.09	0.19
O	-	8	18	-	-	1.59	1.69	-	-	0.20	0.22	-	-	0.08	0.14	-
R	8	-	10	-	1.77	-	1.71	-	0.21	-	0.23	-	0.05	-	0.08	-
S	-	12	13	-	-	1.61	1.66	-	-	0.20	0.23	-	-	0.09	0.09	-
Y	13	-	-	-	1.71	-	-	-	0.23	-	-	-	0.09	-	-	-
Z	8	17	10	-	1.67	1.64	1.69	-	0.21	0.20	0.22	-	0.04	0.12	0.07	-

Most SNPs were in Hardy-Weinberg equilibrium within populations in a given year; on average, 10% of SNPs were out of HWE (Table 5.1).

The sample from population S in 1995 was an outlier in both allelic richness and expected heterozygosity; its allelic richness was above and expected heterozygosity was below the median for all samples by a factor greater than 1.5 times the interquartile range. This sample was also influential in analyses as it drove significant differences in allelic richness between 1995 and 2005. Therefore, I removed data from population S for the year 1995 from all analyses; data from population S were included for years 2005 and 2008, and there were insufficient individuals sampled in 2013. Allelic richness for population R in 1995 was just marginally above the median for all samples by a factor of 1.5 times the interquartile range, but since this sample was not an outlier in any other variables or highly influential, it was included in analyses.

5.3.2 Neutral population genetic analyses

Global F_{ST} was higher in both post-bottleneck years than in pre-bottleneck years, and was highest in 2013 (Table 5.2). Allelic richness fluctuated over time and in concert with the observed demographic bottlenecks. Mean allelic richness was significantly lower in the post-bottleneck year 2013 than in both pre-bottleneck years, 1995 and 2008 (Figure 5.3, Table 5.3). Mean allelic richness in 2005 was also lower than in 1995 and 2008, although the difference was marginally non-significant ($0.05<p<0.1$). Pre-bottleneck years (1995 and 2008) did not differ in allelic richness from each other, and post-bottleneck years (2005 and 2013) did not differ from each other.

Mean expected heterozygosity was significantly lower in 2005 than in other years, including after the second bottleneck in 2013 (Figure 5.4, Table 5.3). Mean expected heterozygosity did not differ significantly among any other pair of years (i.e., expected heterozygosity did not significantly decrease in 2013 after the second bottleneck).

Patterns of IBD among populations also appeared to fluctuate over time, although Mantel tests and MLPE models detected IBD with differing sensitivity. Both methods detected

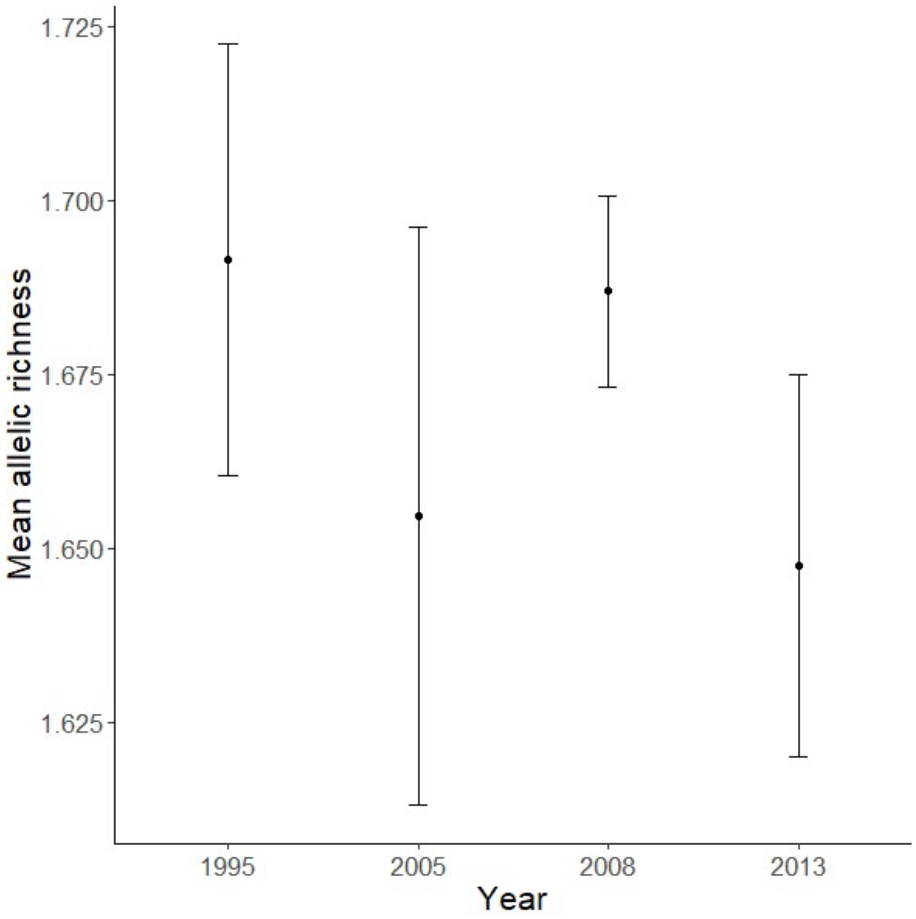

Figure 5.3 Mean allelic richness (rarefied to 10 alleles and averaged across 144 SNP loci) for 14 populations of *Parnassius smintheus* across four years. The years 1995 and 2008 were each before an observed demographic bottleneck, and the years 2005 and 2013 were each after an observed demographic bottleneck. Bars indicate standard deviation among populations in each year.

Table 5.2 Genetic differentiation and diversity metrics, averaged across 14 populations of *Parnassius smintheus*, using a panel of 144 SNPs, in each of four years. Two years (1995 and 2008) were each before an observed demographic bottleneck, and two years (2005 and 2013) were each after an observed demographic bottleneck. Metrics include global F_{ST} (estimated as in Weir & Cockerham, 1984), allelic richness (AR; rarefied to 10 alleles), and expected heterozygosity (H_E).

Year	Global F_{ST}	AR		H_E	
	[95% CI]	Mean	S.D.	Mean	S.D.
1995	0.013 [0.007, 0.020]	1.69	0.031	0.22	0.01
2005	0.020 [0.013, 0.027]	1.65	0.042	0.21	0.011
2008	0.012 [0.008, 0.016]	1.69	0.014	0.22	0.005
2013	0.036 [0.030, 0.043]	1.65	0.027	0.22	0.008

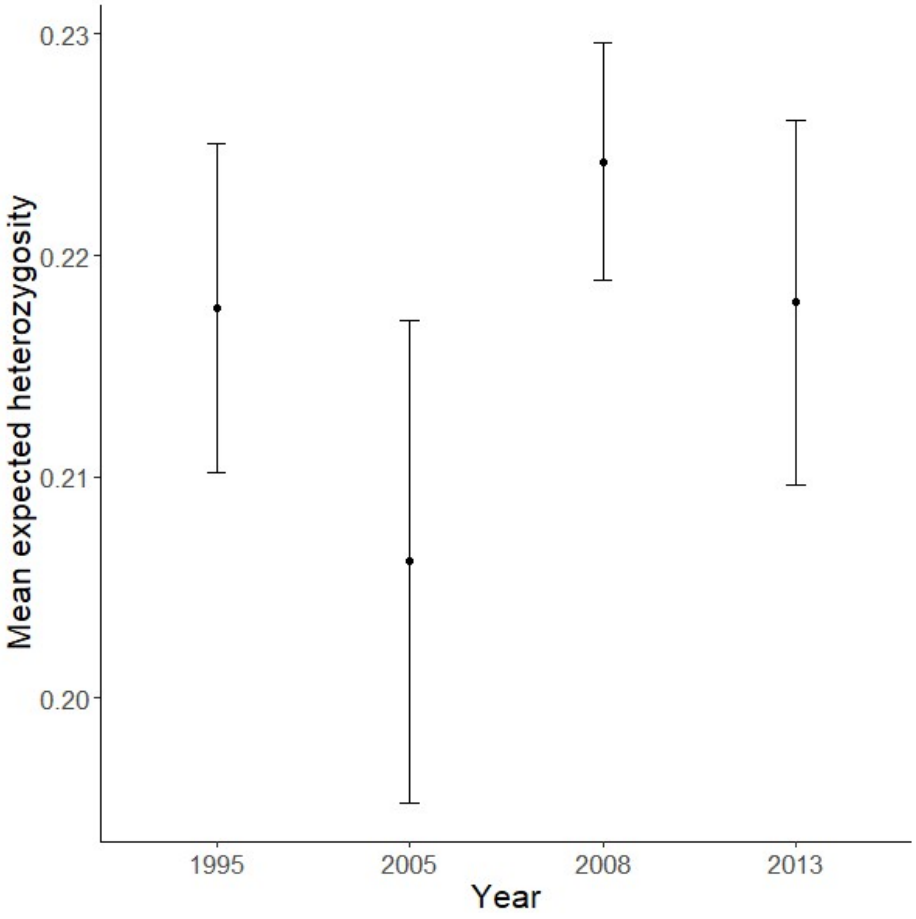

Figure 5.4 Mean expected heterozygosity (averaged across 144 SNP loci) for 14 populations of *Parnassius smintheus* across four years. The years 1995 and 2008 were each before an observed demographic bottleneck, and the years 2005 and 2013 were each after an observed demographic bottleneck. Bars indicate standard deviation among populations in each year.

Table 5.3 The relationship between year and two measures of genetic diversity (AR: allelic richness; H_E: expected heterozygosity) for 14 populations of *Parnassius smintheus,* estimated using linear mixed effects models. The years 1995 and 2008 were each before an observed demographic bottleneck, and the years 2005 and 2013 were each after an observed demographic bottleneck. Differences in genetic diversity between each pair of years were characterized using contrasts, with the associated p-values corrected for multiple comparisons using a Bonferonni correction. 'Coef" represents the estimated model coefficient for the preditor 'year', for each level of contrast. 'Pre vs post bottleneck' represents a contrast of both pre-bottleneck years versus both post-bottleneck years.

Contrasts	AR		H_E	
	coef	p-value	coef	p-value
1995 vs 2005	0.0317	0.05	0.0115	0.024
1995 vs 2008	0.0025	1	-0.0066	0.47
1995 vs 2013	0.0407	0.019	-0.0002	1
2005 vs 2008	-0.0292	0.054	-0.0181	0.0001
2005 vs 2013	0.0090	1	-0.0116	0.046
2008 vs 2013	0.0382	0.024	0.0065	0.74
Pre vs post bottleneck	0.0699	0.0014	0.0179	0.74

significant IBD among populations in 1995 and no IBD in 2005 (Table 5.4). IBD was only detected in 2008 using a Mantel test, and in 2013 using a MLPE model. For the two pre-bottleneck years where IBD was expected, the pattern of IBD as detected by Mantel tests was weaker in 2008 than 1995 (Figure 5.5).

5.3.3 Analyses of signatures of selection

Fifteen loci had minor allele frequencies that differed significantly between pre- and post-bottleneck years, consistent with fluctuating selection. Minor allele frequency showed a significant increase or decrease with year at 18 other loci, consistent with directional selection. After controlling for the false discovery rate using the Benjamini-Hochberg method, no loci had a significant fluctuating allele frequency (i.e., differed significantly between pre- and post-bottleneck years) and three loci showed a significant directional change in minor allele frequency. All three of these loci showed a consistent directional change in allele frequency for at least 3 of the 4 years sampled. The mean frequency of the minor allele at locus PS_1032489 declined over time from 0.50±0 in 1995 to 0.40±0.034 in 2013 (averaged across all populations; Figure 5.6). All individuals were heterozygous at locus PS_1032489 in 1995; the majority of individuals remained heterozygous in other years, but the appearance of major allele homozygotes drove the reduction in minor allele frequency. Minor allele homozygotes at this locus were only observed in 2005 (n=8). The mean frequency of the minor allele at locus PS_9044 increased over time from 0.03±0.04 in 1995 to 0.18±0.09 in 2013, and that at locus PS_77233 increased in frequency from 0 to 0.08±0.06 (Figure 5.6).

Table 5.4 The strength and significance of isolation by distance among 14 *Parnassius smintheus* populations, in each of four different years, characterized using Mantel tests and maximum likelihood population effects (MLPE) models. Pairwise genetic distances were calculated as Nei's corrected genetic distance (Nei, 1978), and geographic distances as the distance along the ridgeline connecting the centroids of meadows. 'r' is the Mantel correlation coefficient, and 'coef' is the estimated coefficient from the MLPE mixed model.

Year	Mantel		MLPE	
	r	p-value	coef	p-value
1995	0.67	0.001	0.0062	0
2005	0.057	0.4	0.00038	0.66
2008	0.44	0.006	0.00042	0.22
2013	0.22	0.19	0.006	0.04

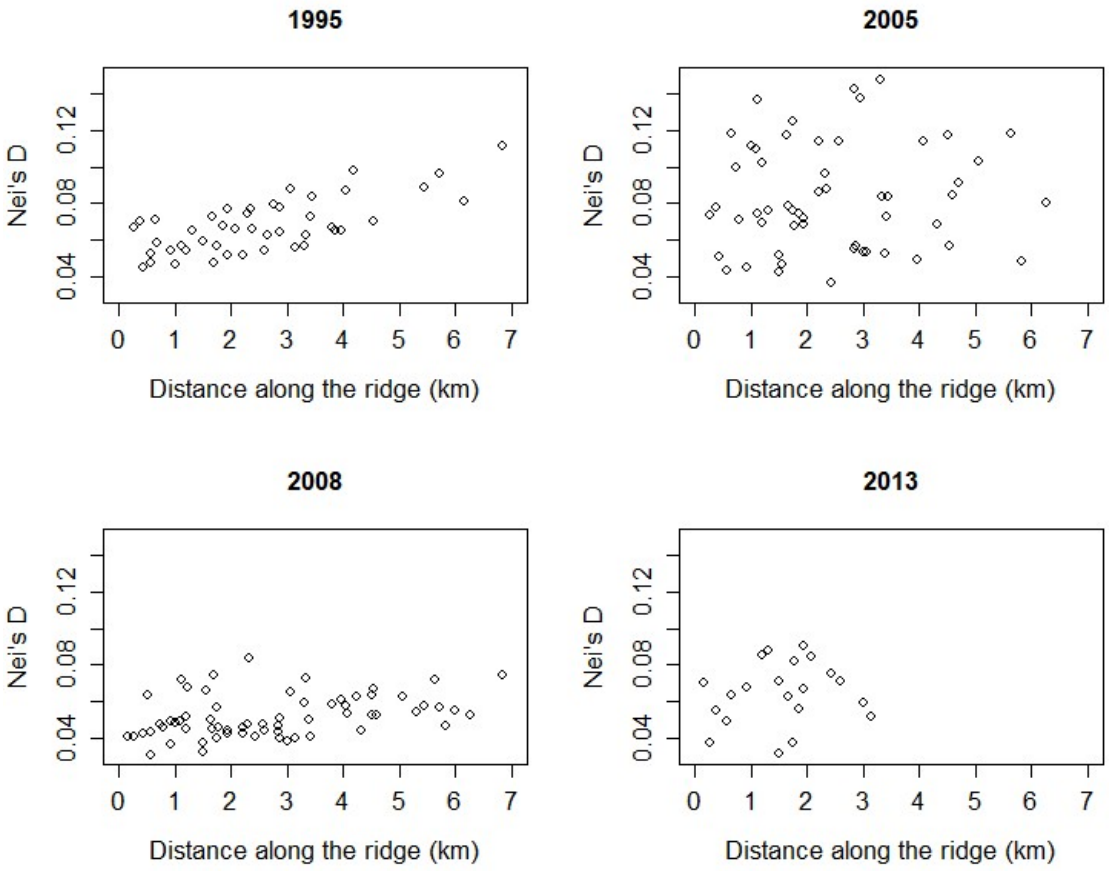

Figure 5.5 The relationship between Nei's corrected genetic distance (Nei, 1978) and geographic distance across 14 *Parnassius smintheus* populations, in each of four different years. Geographic distance was measured along the ridgeline and between the centroids of each populated meadow. The years 1995 and 2008 were each before an observed demographic bottleneck, and the years 2005 and 2013 were each after an observed demographic bottleneck.

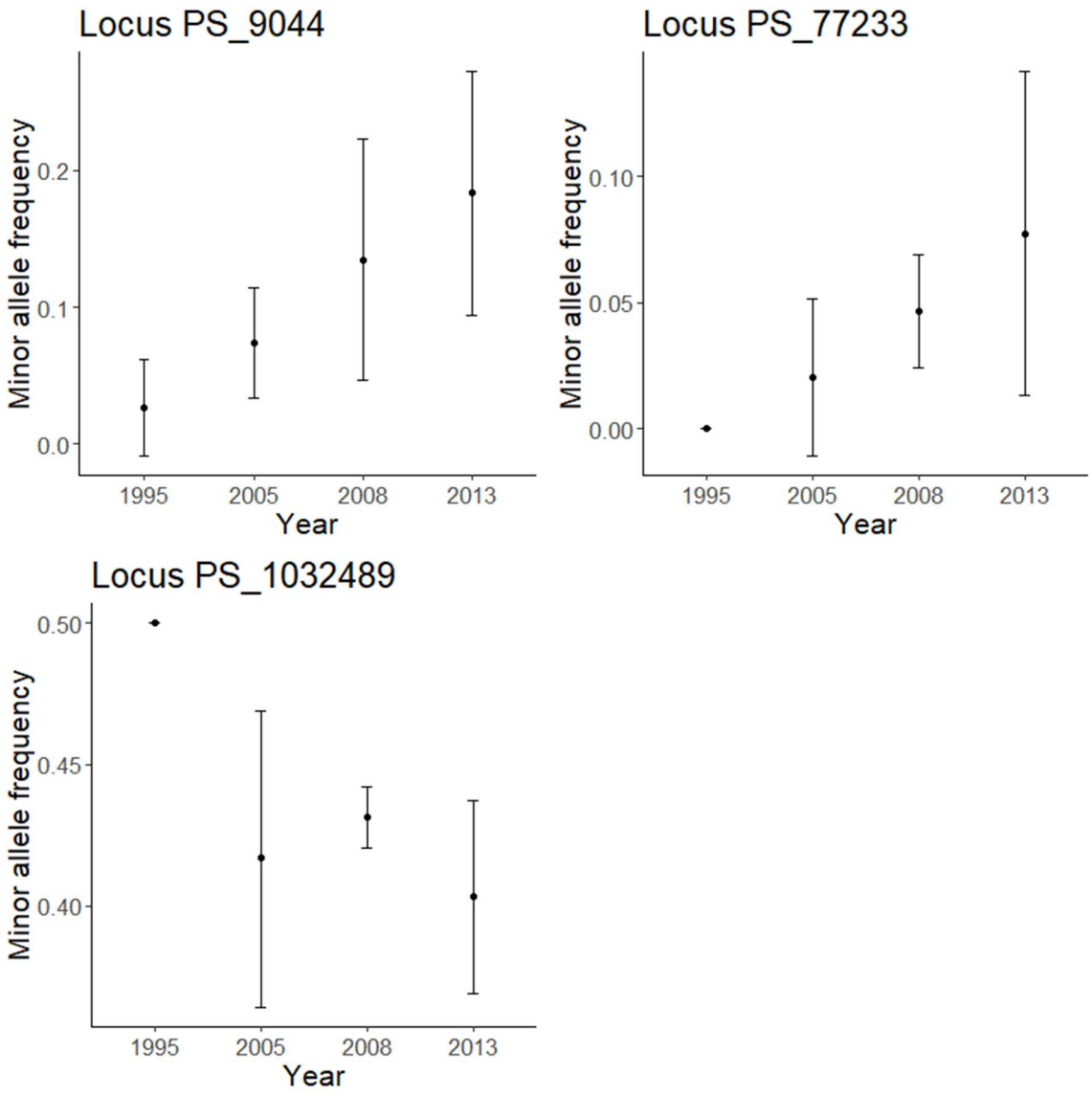

Figure 5.6 Mean minor allele frequency of three SNP loci (PS_9044, PS_77233, and PS_1032489) whose minor allele frequency changed significantly from 1995 to 2013, as measured across 14 *Parnassius smintheus* populations using linear mixed effects models. Bars indicate standard deviation among populations in each year.

5.4 Discussion

5.4.1 Changes in genetic diversity over two bottlenecks

Genetic diversity is expected to be lost during population bottlenecks. Only alleles represented in the surviving individuals will remain (Allendorf, 1986), and additional alleles may continue to be lost rapidly through genetic drift as long as population sizes remain low (Nei et al., 1975). Both the minimum population size reached, and the number of generations at low population size, are expected to contribute to greater genetic loss of genetic diversity. However, this expectation has not often been tested empirically. Considering only the minimum population size reached, among pinniped species the modelled likelihood to have experienced very severe past declines in population size is associated with lower contemporary allelic richness (Stoffel et al., 2018). There is also some evidence that in laboratory populations of *D. melanogaster* the minimum population size reached has a stronger impact on the loss of allelic richness than the length of the bottleneck, but only based on comparison of a short, severe bottleneck to a long bottleneck with a higher minimum population size (England et al., 2003).

I found that allelic richness in *P. smintheus* populations decreased during bottlenecks but rebounded fairly quickly as population sizes recovered within a few generations. The duration of the bottleneck did appear to be important in affecting the loss of allelic diversity. The 2010-2011 bottleneck lasted longer than the 2003 bottleneck, with population numbers remaining very low over two generations (i.e., years) rather than for a single generation (Jangjoo, Matter, Roland, & Keyghobadi, 2016). Previous studies in this system using microsatellites have found that allelic richness declined over the 2010-2011 bottleneck but showed no trend over the 2003 bottleneck (Caplins et al., 2014; Jangjoo et al., 2016). Here, allelic richness estimated using SNPs similarly declined significantly over the longer 2010-2011 bottleneck, and trended downwards, though not significantly so, over the shorter 2003 bottleneck. A second metric of genetic diversity, expected heterozygosity, does not follow the same pattern as allelic richness. Expected

heterozygosity estimated with microsatellites does not fluctuate across bottlenecks (Caplins et al., 2014; Jangjoo et al., 2016), and I found that expected heterozygosity at SNPs was only reduced across the 2003 bottleneck. Expected heterozygosity is expected to be less sensitive to bottlenecks than allelic richness, especially in markers such as microsatellites where rare alleles may be lost with little impact on mean heterozygosity (Allendorf, 1986; Spencer, Neigel, & Leberg, 2000). A loss of allelic richness, but not necessarily of expected heterozygosity, is therefore expected during bottlenecks, as seen for example in populations of the squirrel *Spermophilus lateralis* (McEachern, Van Vuren, Floyd, May, & Eadie, 2011).

While the loss of allelic richness during bottlenecks is ultimately the result of genetic drift, its recovery as population sizes increase again is likely due to gene flow mediated by dispersal among neighbouring populations. That is, novel alleles are re-introduced to populations via immigration from other nearby populations. Across both bottlenecks in this system, there is evidence that allelic diversity at microsatellites recovered most quickly in those populations having higher connectivity to other populations in the network, and therefore higher potential for incoming gene flow (Caplins et al., 2014,.Jangjoo et al., 2016). The primary evolutionary process determining allelic diversity in this system therefore shifts between genetic drift during periods of population decline and gene flow during periods of population recovery (Jangjoo et al. 2020). Furthermore, this shift between gene flow and drift as the dominant influence occurs very rapidly, as seen after the 2003 bottleneck where allelic richness was recovered extremely quickly. For the populations where allelic richness at SNPs decreased between 1995 and 2005, that allelic richness was recovered again by 2008. Allelic richness was thus recovered within three generations of the samples taken in 2005, and five generations of the beginning of the bottleneck in 2003.

In populations where recovery of genetic diversity is driven by immigration, rather than being limited to novel mutations over a longer time span, differences in immigration rates, generation time, effective population size, and population spatial configurations make it difficult to generalize how long recovery of allelic richness should take. In the ground squirrel *Spermophilus lateralis*, recovery of allelic richness post bottleneck was

achieved within just over 1 generation due to high rates of immigration (McEachern, Van Vuren, Floyd, May, & Eadie, 2011). Organisms with more limited dispersal opportunities such as the parasitic louse *Geomydoecus aurei* take at least 45 generations to approach a stable plateau of allelic richness after a founder event (Demastes, Hafner, Hafner, Light, & Spradling, 2019). For *P. smintheus* populations on Jumpingpound ridge, the number of generations required for the recovery of allelic richness may be around five generations, based on the time between the bottleneck in 2003 and the recovery of allelic richness in 2008. This relatively rapid response is likely a result of a combination of factors in particular a moderately high immigration rate among neighbouring populations (Roland et al. 2000) and a short generation time (Epps & Keyghobadi 2015). Genotyping individuals from additional years after 2013 will provide further insight into the temporal dynamics of genetic diversity in this system. Importantly, without an observed plateau of allelic diversity across several years or knowledge of the demographic history prior to 1995, it is unclear whether the level of allelic richness observed in 2008 is what would be expected at near equilibrium conditions. The samples collected between 2013 and the most recent bottleneck in 2019 may shed light on this either in the form of an observable plateau or a continued recovery of allelic richness.

5.4.2 Changes in genetic differentiation over two bottlenecks

As population sizes decrease and genetic drift becomes stronger during bottlenecks, different alleles are lost by chance in different populations. As a result, those populations become more genetically differentiated. Spatial patterns of genetic structure among populations, particularly patterns of IBD, are also likely to be disrupted as the random genetic differentiation of populations is independent of their spatial distance. Subsequently, as populations recover in size, gene flow among them allows nearby populations to again share common alleles, reducing differentiation overall and re-establishing IBD. Overall, and consistent with microsatellite data (Caplins et al. 2014; Jangjoo et al. 2020) I saw the expected increase in genetic differentiation and loss of IBD across the 2003 bottleneck, a decrease in differentiation and partial recovery of IBD by 2008, and another increase in differentiation with some loss of IBD again in 2013 (Figure 5.5, Table 5.4). These patterns provide further support for the shifting importance of drift

and gene flow in this population network as population sizes fluctuate. There are few studies outside of the work on *P. smintheus* examining how genetic differentiation and spatial population genetic structure change across a bottleneck. This requires genetic sampling both before and after what is often an unpredictable event, and in a set of interconnected populations. It can often be simpler to compare population networks with different demographic histories; for example, in sockeye salmon, populations that spawned in the tributaries of one lake believed to have undergone a founding bottleneck lack the IBD seen in other, nearby lakes (Ramstad, Woody, Sage, & Allendorf, 2004).

Detection of significant IBD in 2013 using MLPE models was unexpected and had not been observed in analyses based on microsatellites (Jangjoo et al., 2020). It is worth noting that although I detected significant IBD in 2013 but not in 2008 using MLPE models, when plotted the 2008 data look more consistent with typical IBD patterns than the 2013 data (Figure 5.5). The 2013 data presented here differ somewhat from those analyzed using microsatellites; due to restricted available samples, I analyzed fewer populations (six versus nine). Among the populations I could not include were those located at the ends of Jumpingpound Ridge, resulting in a smaller spatial extent of my analyses compared to the previous microsatellite-based analyses. Additionally, a higher proportion of large, well-connected populations met my minimum sample size requirements in 2013 than in other years. Meadows like L and M that maintain large populations and are centrally located with more opportunities for dispersal tend to lose less microsatellite allelic diversity and regain it faster after bottlenecks (Caplins et al., 2014; Jangjoo et al., 2016). It is possible that these populations with higher connectivity that were disproportionately included in 2013 exchanged more individuals and started to restore a pattern of IBD, which would otherwise not have been observed if smaller, less connected populations had been included.

5.4.3 Signatures of oscillating and directional selection

Superimposed on the shifting relative influence of drift and gene flow among populations through bottleneck events are also potential selection pressures that I was able to explore using my SNP data. Among the loci whose allele frequencies changed directionally and significantly across bottlenecks, one locus (PS_1032489) was found in the adult *P.*

smintheus transcriptome. Using the Basic Local Alignment Search Tool from NCBI, the closest related and sequenced locus is from the diamondback moth, *Plutella xylostella*, and is predicted to encode an RNA-directed DNA polymerase whose function is predicted based on its similarities to the mobile element jockey found in *Drosophila melanogaster* (You et al., 2013). RNA-directed DNA polymerases, also called reverse transcriptases, use an RNA template to encode double stranded DNA. The mobile element jockey may have been introduced into *Drosophila* species from retroviruses (Priimägi, Mizrokhi, & Ilyin, 1988), and is now present at 67 copies in the *D. melanogaster* genome (Kaminker et al., 2002). While mobile elements do not inherently have a function for the host organism besides their own self-replication, some mobile elements have been coopted by the host. For example, the Iris gene in *Drosophila* species was derived from an insect retrovirus, and putatively functions to defend against infection from other retroviruses (Malik & Henikoff, 2005). Other mobile elements can affect an organism's phenotype directly when insertion is close to an expressed region. In *D. melanogaster*, a spontaneous jockey insertion decreased alcohol dehydrogenase activity, possibly affecting the ability of a nearby enhancer to bind to its promoter (White & Jacobson, 1996).

The nature of the changes in genotype at this locus are interesting. In 1995 all individuals sampled were heterozygous at this locus; in following years, the decrease in minor allele frequency is driven by the appearance of individuals that were homozygous at either the major or minor allele. Although the low number of minor allele homozygotes may indicate selection for heterozygotes and against minor allele homozygotes, the presence of such adult homozygotes in 2005 at least indicates that the genotype is not lethal. This deviation from Hardy-Weinberg equilibrium may be driven by a heterozygote advantage at PS_1032489 or another linked locus prior to 1995, after which the potential fitness advantage of the heterozygote appears to have declined. The long-term (i.e., found post-2005) appearance of only the major allele homozygote indicates that there is more than a relaxation of heterozygote advantage, otherwise the minor allele homozygote would also have increased in frequency as the locus moved towards Hardy-Weinberg equilibrium. Furthermore, the proportion of major allele homozygotes decreases somewhat in 2008 relative to 2005 and 2013. Taken together this may indicate the appearance of selection

for the major allele or a linked locus during subsequent bottlenecks, which is driving an overall pattern of directional selection.

For loci whose allele frequencies changed significantly and directionally, such changes in allele frequency may be driven by environmental variables that have been changing over this time span. One such variable is summer temperature, which has been increasing over the time period studied here. The average temperature in July has risen by about 1.3°C since 1990 in the area around Jumpingpound Ridge (based on 5 year averages, 1988-1992 and 2010-2014, from Natural Resources Canada data presented in Chapter 3). Ambient temperature controls and can exert selection on many biochemical, physiological and behavioural traits in insects (Bing & Le, 2005; Sinclair, Addo-Bediako, & Chown, 2003). For example, allozymes may have different optimal temperatures, such that the efficiency of metabolic processes changes with ambient temperature; in the butterfly *Melitaea cinxia*, optimum flight temperature changes depending on the genotype at the phosphoglucose isomerase locus (Niitepõld et al., 2009). *Parnassius smintheus* adults fly and mate in July, so an increase in average temperature could select for individuals that function better at higher temperatures. A visual examination of historical July temperature data (Hutchinson et al., 2009) shows that temperature only begins to increase around 1990. The relatively recent changes in temperature could explain why changes in allele frequency continue to be seen; if the selection pressure had been operating over a longer period, alleles may have already been fixed or reached an equilibrium between any competing selection pressures.

Another relevant environmental change is the decrease in snowfall, over both the entire winter and in November specifically. Reduced snowfall results in a diminished blanket of insulating snow cover, which increases the risk of mortality by exposing overwintering eggs to ambient air temperatures. First, reduced snow cover can increase the probability of premature emergence or increased metabolic activity (thus reducing energy stores) in early winter as a result of exposure to warm air temperatures (Roland & Matter, 2013, 2016). Conversely, it increases the risk of eggs freezing through exposure to temperatures below the lower lethal limit during particularly cold periods. Snow cover on Jumpingpound is decreasing over time (Filazzola, pers. comm.), but year to year variation

in snow cover superimposed on that trend is also likely part of what drives periodic bottlenecks (i.e., low snow cover coupled with unusually high or low November temperature; Roland & Matter, 2013, 2016). Although no locus had significant fluctuations in allele frequency, this does not rule out the possibility that these bottleneck conditions are driving observed directional patterns. If there is no temporally varying balancing selection (i.e., alleles selected for during bottlenecks are not selected against during recovery periods), or if the opposing selection pressure is weak relative to the selection pressure experienced during the bottleneck, then the occurrence of two bottlenecks in a relatively short time span could lend the appearance of consistent directional selection.

Regardless of whether winter conditions might be a source of continuous, directional selection or periodic, fluctuating selection, these sources of egg mortality may select for individuals that have a higher temperature threshold for emergence, greater energy storage, or mechanisms that increase freeze avoidance. One butterfly tolerant of warm winters is *Papilio glaucus,* as compared to its congener *P. canadensis* (a species with which it occasionally hybridizes). Unlike *P. glaucus* pupae, *P. canadensis* pupae lose body weight, likely as a result of increased metabolism, when overwintering temperatures increase (Mercader & Scriber, 2008). Given the variation observed between these are different but very closely related, species, it is plausible that there is also intraspecific variation in temperature-dependent metabolic enzyme activity upon which selection could act. Furthermore, in the case of selection driven by egg freezing, *P. smintheus* individuals with greater expression levels of glycerol, mannitol, and trehalose (known cryoprotectants in the related *P bremeri*; Park, Kim, Park, Lee, & Lee, 2017) may survive better. In the context of freeze avoidant species, cryoprotectant expression levels may be heritable and shift the lower lethal temperature downward (Morey, Venette, & Hutchison, 2013). Alternately, the lack of significant fluctuations at any SNP locus may indicate that surviving a bottleneck event has no genetic basis. Snow cover varies spatially within Jumpingpound meadows (Roland, Filazzola, & Matter, 2021); the eggs that survive bottlenecks may simply be those laid in areas that happened to accumulate greater snow cover, rather than representing individuals with greater cold tolerance.

5.4.4 Conclusions

Repeated, extreme fluctuations in the size of natural populations provide an empirical system in which to test theories of how bottlenecks affect neutral genetic diversity and differentiation. When different cycles of population size increase and decrease are driven by consistent processes, and are associated with similar selective pressures, there is an additional opportunity to study changes at functional loci potentially driven by selection. I found that in a network of populations of *P. smintheus*, the relative importance of genetic drift and gene flow appeared to shift across two bottlenecks. Consistent with previous data from microsatellite markers, genetic diversity and population structure were lost through drift at low population sizes, and then rapidly recovered (within five generations) as dispersing individuals reintroduced alleles among local populations. By using a SNP panel rather than neutral markers, I also showed that for a small proportion of loci, allele frequency changed directionally over the period of study, indicating that those loci are potentially under directional selection. The ongoing development of a *P. smintheus* genome will be important for exploring selection pressures further, as it may expose potential functions of, or other loci linked to, those SNPs where significant directional change was observed. A recent bottleneck in 2019 is also an exciting opportunity to further test the conclusions of this study, both to test for the repeated loss of overall genetic diversity and spatial genetic structure, and to examine whether directional patterns of change at putatively selected SNPs continue.

5.5 Literature cited

Achaz, G. (2009). Frequency spectrum neutrality tests: One for all and all for one. *Genetics*, *183*(1), 249–258.

Agudo, R., Carrete, M., Alcaide, M., Rico, C., Hiraldo, F., & Donázar, J. A. (2012). Genetic diversity at neutral and adaptive loci determines individual fitness in a long-lived territorial bird. *Proceedings: Biological Sciences*, *279*(1741), 3241–3249.

Allendorf, F. W. (1986). Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology*, *5*(2), 181–190.

Allio, R., Scornavacca, C., Nabholz, B., Clamens, A.-L., Sperling, F. A., & Condamine, F. L. (2020). Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Systematic Biology*, *69*(1), 38–60.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81–92.

Bing, C., & Le, K. (2005). Insect population differentiation in response to enviromental thermal stress. *Progress in Natural Science*, *15*(4), 289–296.

Boag, P. T., & Grant, P. R. (1981). Intense natural selection in a population of Darwin's finches (Geospizinae) in the Galápagos. *Science*, *214*(4516), 82–85.

Boratyn, G. M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B., & Madden, T. L. (2019). Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinformatics*, *20*(1), 405.

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., … Whited, J. L. (2017). A tissue-mapped axolotl *de novo* transcriptome enables identification of limb regeneration factors. *Cell Reports*, *18*(3), 762–776.

Caplins, S. A., Gilbert, K. J., Ciotir, C., Roland, J., Matter, S. F., & Keyghobadi, N. (2014). Landscape structure and the genetic effects of a population collapse. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1796), 20141798.

Chakraborty, R., & Nei, M. (1977). Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. *Evolution*, *31*(2), 347–356.

Clarke, R. T., Rothery, P., & Raybould, A. F. (2002). Confidence limits for regression relationships between distance matrices: Estimating gene flow with distance. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*(3), 361.

Craig, C. C. (1953). On the utilization of marked specimens in estimating populations of flying insects. *Biometrika*, *40*(1/2), 170–176.

Demastes, J. W., Hafner, D. J., Hafner, M. S., Light, J. E., & Spradling, T. A. (2019). Loss of genetic diversity, recovery and allele surfing in a colonizing parasite, *Geomydoecus aurei*. *Molecular Ecology*, *28*(4), 703–720.

Dyer, R. J. (2021). *gstudio: An package for the spatial analysis of population genetic marker data*.

England, P. R., Osler, G. H. R., Woodworth, L. M., Montgomery, M. E., Briscoe, D. A., & Frankham, R. (2003). Effects of intense versus diffuse population bottlenecks on microsatellite genetic diversity and evolutionary potential. *Conservation Genetics*, *4*(5), 595–604.

Epps, C. W., & Keyghobadi, N. (2015). Landscape genetics in a changing world: Disentangling historical and contemporary influences and inferring change. *Molecular Ecology*, *24*(24), 6021–6040.

Fauvelot, C., Cleary, D. F. R., & Menken, S. B. J. (2006). Short-term impact of disturbance on genetic diversity and structure of Indonesian populations of the butterfly *Drupadia theda* in East Kalimantan. *Molecular Ecology*, *15*(8), 2069–2081.

Frankham, R. (2005). Genetics and extinction. *Biological Conservation*, *126*(2), 131–140.

Frankham, R., Lees, K., Montgomery, M. E., England, P. R., Lowe, E. H., & Briscoe, D. A. (1999). Do population size bottlenecks reduce evolutionary potential? *Animal Conservation Forum*, *2*(4), 255–260.

Gattepaille, L. M., Jakobsson, M., & Blum, M. G. (2013). Inferring population size changes with sequence and SNP data: Lessons from human bottlenecks. *Heredity*, *110*(5), 409–419.

Goff, J., Yerke, C., Keyghobadi, N., & Matter, S. F. (2018). Dispersing male *Parnassius smintheus* butterflies are more strongly affected by forest matrix than are females. *Insect Science*, *26*(5), 932-944.

Gos, G., Slotte, T., & Wright, S. I. (2012). Signatures of balancing selection are maintained at disease resistance loci following mating system evolution and a population bottleneck in the genus Capsella. *BMC Evolutionary Biology*, *12*(1), 1–10.

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, *5*(1), 184–186.

Grant, B. R., & Grant, P. R. (1989). Natural Selection in a Population of Darwin's Finches. *The American Naturalist*, *133*(3), 377–393.

Hutchinson, M. F., McKenney, D. W., Lawrence, K., Pedlar, J. H., Hopkinson, R. F., Milewska, E., & Papadopol, P. (2009). Development and testing of Canada-wide interpolated spatial models of daily minimum–maximum temperature and precipitation for 1961–2003. *Journal of Applied Meteorology and Climatology*, *48*(4), 725–741.

Jangjoo, M. (2018). Spatial and temporal patterns of neutral and adaptive genetic variation in the alpine butterfly, *Parnassius smintheus*. *Electronic Thesis and Dissertation Repository*.

Jangjoo, M., Matter, S. F., Roland, J., & Keyghobadi, N. (2016). Connectivity rescues genetic diversity after a demographic bottleneck in a butterfly population network. *Proceedings of the National Academy of Sciences*, *113*(39), 10914–10919.

Jangjoo, M., Matter, S. F., Roland, J., & Keyghobadi, N. (2020). Demographic fluctuations lead to rapid and cyclic shifts in genetic structure among populations of an alpine butterfly, *Parnassius smintheus*. *Journal of Evolutionary Biology*, *33*(5), 668–681.

Jardine, D. I., Blanc-Jolivet, C., Dixon, R. R. M., Dormontt, E. E., Dunker, B., Gerlach, J., … Lowe, A. J. (2016). Development of SNP markers for Ayous (*Triplochiton scleroxylon* K. Schum) an economically important tree species from tropical West and Central Africa. *Conservation Genetics Resources*, *8*(2), 129–139.

Johnston, R. F., Niles, D. M., & Rohwer, S. A. (1972). Hermon Bumpus and natural selection in the house sparrow *Passer domesticus*. *Evolution*, *26*(1), 20–31.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070–3071.

Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., … Celniker, S. E. (2002). The transposable elements of the *Drosophila melanogaster* euchromatin: A genomics perspective. *Genome Biology*, *3*(12), research0084.1-84.2.

Kekkonen, J., Hanski, I. K., Jensen, H., Väisänen, R. A., & Brommer, J. E. (2011). Increased genetic differentiation in house sparrows after a strong population decline: From panmixia towards structure in a common bird. *Biological Conservation*, *144*(12), 2931–2940.

Keyghobadi, N., Roland, J., & Strobeck, C. (1999). Influence of landscape on the population genetic structure of the alpine butterfly *Parnassius smintheus* (Papilionidae). *Molecular Ecology*, *8*(9), 1481–1495.

Le Gouar, P. J., Vallet, D., David, L., Bermejo, M., Gatti, S., Levréro, F., … Ménard, N. (2009). How Ebola impacts genetics of western lowland gorilla populations. *PLoS ONE*, *4*(12).

Leberg, P. L. (1992). Effects of population bottlenecks on genetic diversity as measured by allozyme electrophoresis. *Evolution*, *46*(2), 477–494.

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33.

Luikart, G., Sherwin, W. B., Steele, B. M., & Allendorf, F. W. (1998). Usefulness of molecular markers for detecting population bottlenecks via monitoring genetic change. *Molecular Ecology*, *7*(8), 963–974.

Malik, H. S., & Henikoff, S. (2005). Positive selection of *Iris*, a retroviral *Envelope*–derived host gene in *Drosophila melanogaster*. *PLOS Genetics*, *1*(4), e44.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, *27*(2 Part 1), 209–220.

Matter, S. F., Keyghobadi, N., & Roland, J. (2014). Ten years of abundance data within a spatial population network of the alpine butterfly, *Parnassius smintheus*. *Ecology*, *95*(10), 2985–2985.

McEachern, M. B., Van Vuren, D. H., Floyd, C. H., May, B., & Eadie, J. M. (2011). Bottlenecks and rescue effects in a fluctuating population of golden-mantled ground squirrels (*Spermophilus lateralis*). *Conservation Genetics*, *12*(1), 285–296.

Mercader, R. J., & Scriber, J. M. (2008). Asymmetrical thermal constraints on the parapatric species boundaries of two widespread generalist butterflies. *Ecological Entomology*, *33*(4), 537–545.

Mitikka, V., & Hanski, I. (2010). Pgi genotype influences flight metabolism at the expanding range margin of the European map butterfly. *Annales Zoologici Fennici*, *47*(1), 1–14.

Morey, A., Venette, R., & Hutchison, W. (2013). Could natural selection change the geographic range limits of light brown apple moth (Lepidoptera, Tortricidae) in North America? *NeoBiota*, *18*, 151–156.

Muñoz, I., Henriques, D., Johnston, J. S., Chávez-Galarza, J., Kryger, P., & Pinto, M. A. (2015). Reduced SNP panels for genetic identification and introgression analysis in the dark honey bee (*Apis mellifera mellifera*). *PLOS ONE*, *10*(4), e0124365.

Natesh, M., Taylor, R. W., Truelove, N. K., Hadly, E. A., Palumbi, S. R., Petrov, D. A., & Ramakrishnan, U. (2019). Empowering conservation practice with efficient and economical genotyping from poor quality samples. *Methods in Ecology and Evolution*, *10*(6), 853–859.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, *89*(3), 583–590.

Nei, M., Maruyama, T., & Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations. *Evolution*, *29*(1), 1–10.

Nicholson, A. J. (1954). An outline of the dynamics of animal populations. *Australian Journal of Zoology*, *2*(1), 9–65.

Niitepõld, K., Smith, A. D., Osborne, J. L., Reynolds, D. R., Carreck, N. L., Martin, A. P., … Hanski, I. (2009). Flight metabolic rate and Pgi genotype influence butterfly dispersal rate in the field. *Ecology*, *90*(8), 2223–2232.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2019). *vegan: Community Ecology Package (Version 2.5-6)*.

Oliver, M. K., & Piertney, S. B. (2012). Selection maintains MHC diversity through a natural population bottleneck. *Molecular Biology and Evolution*, *29*(7), 1713–1720.

Ørsted, M., Hoffmann, A. A., Sverrisdóttir, E., Nielsen, K. L., & Kristensen, T. N. (2019). Genomic variation predicts adaptive evolutionary responses better than population bottleneck history. *PLOS Genetics*, *15*(6), e1008205.

Paradis, E. (2010). pegas: An R package for population genetics with an integrated–modular approach. *Bioinformatics*, *26*(3), 419–420.

Park, Y., Kim, Y., Park, G.-W., Lee, J.-O., & Lee, K.-W. (2017). Supercooling capacity along with up-regulation of glycerol content in an overwintering butterfly, *Parnassius bremeri*. *Journal of Asia-Pacific Entomology*, *20*(3), 949–954.

Pilot, M., Greco, C., vonHoldt, B. M., Jędrzejewska, B., Randi, E., Jędrzejewski, W., … Wayne, R. K. (2014). Genome-wide signatures of population bottlenecks and diversifying selection in European wolves. *Heredity*, *112*(4), 428–442.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & Deepayan, S. (2015). *nlme: Linear and nonlinear mixed effects models*.

Pope, N. (2020). *CorMLPE: A correlation structure for symmetric relational data*.

Priimägi, A. F., Mizrokhi, L. J., & Ilyin, Y. V. (1988). The Drosophila mobile element jockey belongs to LINEs and contains coding sequences homologous to some retroviral proteins. *Gene*, *70*(2), 253–262.

R Core Team. (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ramos-Onsins, S. E., & Rozas, J. (2002). Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution*, *19*(12), 2092–2100.

Ramstad, K. M., Woody, C. A., Sage, G. K., & Allendorf, F. W. (2004). Founding events influence genetic population structure of sockeye salmon (*Oncorhynchus nerka*) in Lake Clark, Alaska. *Molecular Ecology*, *13*(2), 277–290.

Rank, N. E., Bruce, D. A., McMillan, D. M., Barclay, C., & Dahlhoff, E. P. (2007). Phosphoglucose isomerase genotype affects running speed and heat shock protein expression after exposure to extreme temperatures in a montane willow beetle. *Journal of Experimental Biology*, *210*(5), 750–764.

Reed, D. H., & Frankham, R. (2003). Correlation between fitness and genetic diversity. *Conservation Biology*, *17*(1), 230–237.

Roland, J., Filazzola, A., & Matter, S. F. (2021). Spatial variation in early-winter snow cover determines local dynamics in a network of alpine butterfly populations. Ecography, 44(2), 334–343.

Roland, J., Keyghobadi, N., & Fownes, S. (2000). Alpine parnassius butterfly dispersal: Effects of landscape and population size. *Ecology*, *81*(6), 1642–1653.

Roland, J., & Matter, S. F. (2013). Variability in winter climate and winter extremes reduces population growth of an alpine butterfly. *Ecology*, *94*(1), 190–199.

Roland, J., & Matter, S. F. (2016). Pivotal effect of early-winter temperatures and snowfall on population growth of alpine *Parnassius smintheus* butterflies. *Ecological Monographs*, *86*(4), 412–428.

Sinclair, B. J., Addo-Bediako, A., & Chown, S. L. (2003). Climatic variability and the evolution of insect freeze tolerance. *Biological Reviews*, *78*(2), 181–195.

Spencer, C. C., Neigel, J. E., & Leberg, P. L. (2000). Experimental evaluation of the usefulness of microsatellite DNA for detecting demographic bottlenecks. *Molecular Ecology*, *9*(10), 1517–1528.

Stoffel, M. A., Humble, E., Paijmans, A. J., Acevedo-Whitehouse, K., Chilvers, B. L., Dickerson, B., … Hoffman, J. I. (2018). Demographic histories and genetic diversity across pinnipeds are shaped by human exploitation, ecology and life-history. *Nature Communications*, *9*(1), 4836.

Suárez, N. M., Betancor, E., Fregel, R., Rodríguez, F., & Pestano, J. (2012). Genetic signature of a severe forest fire on the endangered Gran Canaria blue chaffinch (*Fringilla teydea polatzeki*). *Conservation Genetics*, *13*(2), 499–507.

Sutton, J. T., Nakagawa, S., Robertson, B. C., & Jamieson, I. G. (2011). Disentangling the roles of natural selection and genetic drift in shaping variation at MHC immunity genes. *Molecular Ecology*, *20*(21), 4408–4420.

Tabuloc, C. A., Lewald, K. M., Conner, W. R., Lee, Y., Lee, E. K., Cain, A. B., … Chiu, J. C. (2019). Sequencing of *Tuta absoluta* genome to develop SNP genotyping assays for species identification. *Journal of Pest Science*, *92*(4), 1397–1407.

Watt, W. B., Wheat, C. W., Meyer, E. H., & Martin, J.-F. (2003). Adaptation at specific loci. VII. Natural selection, dispersal and the diversity of molecular–functional variation patterns among butterfly species complexes (Colias: Lepidoptera, Pieridae). *Molecular Ecology*, *12*(5), 1265–1275.

Watt, Ward B. (1977). Adaptation at specific loci. I. Natural selection on phosphoglucose isomerase of Colias butterflies: Biochemical and population aspects. *Genetics*, *87*(1), 177–194.

Watt, Ward B., Carter, P. A., & Blower, S. M. (1985). Adaptation at specific loci. IV. Differential mating success among glycolytic allozyme genotypes of Colias butterflies. *Genetics*, *109*(1), 157–175.

Wegner, K. M., Kalbe, M., Milinski, M., & Reusch, T. B. (2008). Mortality selection during the 2003 European heat wave in three-spined sticklebacks: Effects of parasites and MHC genotype. *BMC Evolutionary Biology*, *8*(1), 1–12.

White, L. D., & Jacobson, J. W. (1996). Insertion of the retroposable element, jockey, near the Adh gene of *Drosophila melanogaster* is associated with altered gene expression. *Genetics Research*, *68*(3), 203–209.

Windig, J. J., Veerkamp, R. F., & Nylin, S. (2004). Quantitative genetic variation in an island population of the speckled wood butterfly (*Pararge aegeria*). *Heredity*, *93*(5), 450–454.

You, M., Yue, Z., He, W., Yang, X., Yang, G., Xie, M., … Wang, J. (2013). A heterozygous moth genome provides insights into herbivory and detoxification. *Nature Genetics*, *45*(2), 220–225.

Chapter 6

6 General Discussion

6.1 Overview

My thesis adds to the growing body of work that takes advantage of advances in DNA sequencing technology to investigate the processes that shape genetic diversity in natural populations. I used single nucleotide polymorphisms identified through a reduced representation sequencing approach to investigate how genetic diversity is shaped by environmental factors across two spatial and temporal scales in the alpine butterfly *Parnassius smintheus*. My first objective was to develop and test a reduced representation sequencing protocol for *P. smintheus*, which I used to inform decisions around how to best process data for different population genetic analyses (Chapter 2) and to make recommendations for the minimum sample sizes for future surveys (Chapter 3). My second objective was to apply this dataset to two sets of natural *P. smintheus* populations sampled at different spatial and temporal scales (i.e., a broader spatial extent at a single time point, and a finer spatial extent over multiple years), to assess what factors shaped patterns of genetic diversity through the processes of gene flow, genetic drift, and selection (Chapters 4 and 5). Throughout my thesis, my major contributions include making valuable recommendations for data processing and sampling for future studies using a reduced representation approach, and providing evidence for how landscape, and variability in weather, affect patterns of genetic diversity.

6.2 The future of molecular markers in population genetics

The most popular molecular marker at any given time in the history of population genetics has depended on the technology of the era. With each new advance, old molecular markers have largely become outdated (Schlötterer, 2004). Large datasets of single nucleotide polymorphisms, as genotyped by reduced representation or whole genome sequencing, are among the newest of these markers (Casillas & Barbadilla, 2017). Although the use of SNPs is becoming increasingly common, microsatellites

remain a popular choice (Garrido-Cardenas, Mesa-Valle, & Manzano-Agugliaro, 2018; Puckett, 2017). As recently as 2018, the number of published studies investigating genetic structure using SNPs genotyped by reduced representation sequencing were only 10% of the number that used microsatellites (Sunde, Yıldırım, Tibblin, & Forsman, 2020). Some reasons that microsatellites remain popular include their ease of use, affordability, and comparability with previously published studies that used the same microsatellite markers (Hodel et al., 2016). Microsatellites have also become easier to develop, as potential microsatellite markers can be identified in silico from the whole genome sequences of one or more individuals (Hodel et al., 2016).

Reduced representation sequencing approaches such as RADseq are used because whole genome sequencing of many individuals is costly (Fuentes-Pardo & Ruzzante, 2017). As the cost of whole genome sequencing continues to fall and available computing power increases, reduced representation sequencing may become unnecessary and be replaced by whole genome resequencing. However, both reduced representation and especially whole genome sequencing have stricter minimum DNA quantity and quality requirements than other markers, including microsatellites and other SNP genotyping techniques (e.g., MassARRAY) (Campbell & Narum, 2009; Jordon-Thaden et al., 2020). For example, the wing clips taken from *P. smintheus* individuals at Jumpingpound Ridge yield sufficient DNA for microsatellite and MassARRAY SNP analysis, but not enough for RADseq. Whole genome amplification is a potential solution to low starting DNA quantities; however, it is both costly and can result in biases in amplification, the introduction of artifacts, the amplification of off-target (e.g., contaminating DNA) sequences (Sabina & Leamon, 2015), and lower numbers of loci sequenced than non-amplified libraries (Medeiros & Farrell, 2018). As part of Chapter 5 I attempted whole genome amplification of wing clip DNA followed by RADseq, and was unsuccessful – the resulting RADseq libraries were dominated by a small number of highly replicated sequences. Although the availability of a whole genome sequence for comparison would alleviate some of the errors introduced by whole genome amplification (e.g., by screening out primer artifacts and off-target sequences), the process remains costly and, in my experience with *P. smintheus* wing clips, prone to errors. Given the current state of both whole genome sequencing and whole genome amplification, for studies with limited

starting DNA both microsatellites and targeted SNP panels will remain the viable options for molecular markers for the near future.

6.3 Contributions to the *Parnassius smintheus* model system

Parnassius smintheus populations in western Alberta have now been extensively studied for the past 26 years. Previously developed tools for their study have included mark-recapture protocols (Roland, Keyghobadi, & Fownes, 2000), microsatellite markers (Keyghobadi, Roland, & Strobeck, 1999), SNP markers within the phosphoglucose isomerase gene (Jangjoo, 2018), an adult transcriptome (Jangjoo, 2018), and recently, a shotgun sequenced genome (Allio et al., 2020). Some of these, including mark-recapture and collection of wing clips for genetic analysis, are used as part of the annual survey of the Jumpingpound populations. Others, such as the transcriptome, have been used to investigate specific questions, such as whether expression levels differ between recently dispersed and non-dispersed individuals (Jangjoo, 2018), but have continued to have important and unanticipated secondary uses in subsequent projects. For example, I used the transcriptome as a tool to differentiate expressed and non-expressed SNP loci when choosing which SNPs to include in the panel I used in Chapter 5. This abundance of tools and datasets, in combination with extensive genetic and dispersal data collected annually over many years, make the Jumpingpound Ridge *P. smintheus* populations an emerging model system in empirical landscape and population genetics.

I developed two novel genomic tools for *P. smintheus* that I then used to generate two different datasets. First, in Chapters 2, 3, and 4 I used a ddRADseq approach to genotype *P. smintheus* individuals from populations across western Alberta at thousands of SNP loci. In doing so, I both developed a protocol (i.e., identified restriction enzymes, appropriate fragment size selection, and optimal PCR conditions) for ddRADseq for whole body *P. smintheus* samples, as well as created a dataset that may be used to address additional landscape and population genetic hypothesis. I expect that the ddRADseq dataset may be used in the future to continue to address questions of how landscape affects genetic structure in *P. smintheus*. For example, I did not examine how

the distribution and density of the larval host plant *Sedum lanceolatum* could affect genetic diversity, separately from the size of each meadow patch. Furthermore, future studies could make use of the ddRADseq data I generated to investigate loci that may be underlying local adaptation among populations of *P. smintheus* in the foothills and front ranges of the Rocky Mountains.

Second, in Chapter 5 I developed a small SNP panel of a few hundred loci, including both expressed and putatively non-expressed loci, that I used to genotype individuals from the Jumpingpound Ridge populations across four years (1995, 2005, 2008, and 2013). This SNP panel is currently being used to examine the genetic effects of an experimental extinction on Jumpingpound Ridge, as well the effects of the demographic bottlenecks on inbreeding and fitness. This novel tool that I developed may continue to be used in the future in conjunction with, or as a replacement for, microsatellite genotyping of the Jumpingpound Ridge populations, which has been extensively used in past studies (Caplins et al., 2014; Jangjoo, Matter, Roland, & Keyghobadi, 2016, 2020; Keyghobadi, Roland, & Strobeck, 1999).

6.4 Weather variability and genetic diversity

Both weather and climate (the long-term average of weather) affect the abundance and persistence of populations, which in turn affects their genetic diversity. Populations at the edge of their species' climatic envelope are predicted to have lower rates of reproduction and survival than populations at more optimal conditions, resulting in lower and more variable population sizes and therefore decreased genetic diversity over time. This is seen in the lizard *Ameivula ocellifera*, where populations in regions with more favourable climate conditions (e.g., higher temperature seasonality) have greater genetic diversity than populations occupying less favourable conditions (Oliveira et al., 2018). In the trout *Oncorhynchus mykiss*, higher average precipitation during the winter is associated with greater genetic differentiation, potentially as a result of greater spring runoff driving higher embryo mortality (Hand et al., 2016). Another mechanism by which weather and climate affect genetic diversity is through local adaptation. Patterns of isolation by environment, where genetic distances between populations are larger when differences in

their local environments are greater, emerge when there is local adaptation to environmental conditions which reduces the successful survival and reproduction of dispersers (Wang & Bradburd, 2014). For example, differences in alkalinity and chemical compounds between lakes is a predictor of genetic differentiation in the brine shrimp *Artemia franciscana*, likely due to reduced hatching and survival outside of their natal conditions (Frisch et al., 2021).

In the above examples, and indeed in most empirical population and landscape genetic studies, static climate variables (e.g., long-term average temperature or precipitation) are often used when assessing the impact of climate on genetic variation. Furthermore, where variability in weather and climate is considered, it is often as average intra-annual variability; for example, temperature seasonality in the study of the lizard *A. ocellifera* refers to the long-term average of the difference between maximum and minimum temperatures within each year (Oliveira et al., 2018). However, inter-annual variability in both weather (e.g., droughts, extreme temperatures) and climate (e.g., El Niño-Southern Oscillation, Pacific Decadal Oscillation) are also important drivers of population size and genetic diversity. Uncommon but extreme weather, such as extreme high or low precipitation or severe storms, can have long-lasting impacts on population genetics. When weather events cause bottlenecks, genetic diversity may be lost both genome-wide, as alleles are lost both during the initial decline in population size and through stronger genetic drift for as long as the population size remains low (Clark, Marchand, Clifford, Stechert, & Stephens, 2011), and at any loci experiencing selection during the weather event, such as loci that increase drought resistance (Whitney et al., 2019 but see Dillon et al., 2015). As climate change is associated not only with shifting average conditions, but also with increasing weather variability and more frequent extreme weather events, incorporating weather variability when modelling biological patterns and processes, including predictors of genetic diversity, is increasingly important. Studies of both current distributions of genetic diversity, such as in the bettong *Bettongia gaimardi* (Proft et al., 2021), and models of genetic diversity under various climate change scenarios, such as for the woodpecker *Dendrocopos medius* (Cobben et al., 2011) confirm that increased weather variability is associated with lower genetic diversity.

In Chapter 4, I looked at whether average and extreme weather conditions were predictors of genetic diversity and differentiation in *P. smintheus* populations at a broad (tens to hundreds of kilometers) spatial scale. In Chapter 5, I examined the genetic consequences of repeated bottlenecks, which are likely driven by early winter weather, on the Jumpingpound Ridge populations. In Chapter 4, I found that both average (mean daily low temperatures) and extreme weather (the lowest November snow depth over a 10 year period) predicted genetic diversity and differentiation. These results were congruent with the immediate effects of the bottlenecks examined in Chapter 5; specifically, at the larger scale populations with lower snow cover, which is expected to result in more frequent and severe demographic collapses (Roland & Matter, 2016), were more genetically differentiated from other populations and had lower expected heterozygosity. These results at the larger scale are consistent with how the Jumpingpound populations lost genetic diversity after each bottleneck. Both sets of results support the importance of weather variability (in this case, variability in snow cover specifically) on genetic diversity.

Interestingly, at Jumpingpound Ridge there is no observed long-term decline in genetic diversity despite the observed bottlenecks; while genetic diversity is lost in each population in the short term, it is recovered within several years to pre-bottleneck levels as a result of gene flow (Caplins et al., 2014; Jangjoo et al., 2016, 2020). This apparent difference between the potential effect of years with low snow cover on genetic diversity at Jumpingpound Ridge (i.e., a temporary decline and recovery, but no long-term change) versus among the more widely distributed *P. smintheus* populations (i.e., a predictor of differences in genetic diversity among populations) likely reflects the different spatial and temporal scales covered in Chapters 4 and 5. While the Jumpingpound populations may individually be losing and then recovering genetic diversity through each bottleneck, the average genetic diversity across the Jumpingpound metapopulation (like other populations in the East Kananaskis region) is lower than other *P. smintheus* populations that experience different weather and landscape conditions (e.g., greater meadow connectivity and snow cover). Here, using data from two spatial and temporal scales I have provided complementary evidence for the way that genetic diversity can be reduced through reduced population sizes (as seen at a smaller spatial but longer temporal scale;

Chapter 5, and what landscape and weather variables are associated with those losses (at a larger spatial scale, but observed at a single point in time; Chapter 4).

6.5 Literature cited

Allio, R., Scornavacca, C., Nabholz, B., Clamens, A.-L., Sperling, F. A., & Condamine, F. L. (2020). Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Systematic Biology*, *69*(1), 38–60.

Campbell, N. R., & Narum, S. R. (2009). Quantitative PCR assessment of microsatellite and SNP genotyping with variable quality DNA extracts. *Conservation Genetics*, *10*(3), 779–784.

Caplins, S. A., Gilbert, K. J., Ciotir, C., Roland, J., Matter, S. F., & Keyghobadi, N. (2014). Landscape structure and the genetic effects of a population collapse. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1796), 20141798.

Casillas, S., & Barbadilla, A. (2017). Molecular population genetics. *Genetics*, *205*(3), 1003–1035.

Clark, R. W., Marchand, M. N., Clifford, B. J., Stechert, R., & Stephens, S. (2011). Decline of an isolated timber rattlesnake (*Crotalus horridus*) population: Interactions between climate change, disease, and loss of genetic diversity. *Biological Conservation*, *144*(2), 886–891.

Cobben, M. M. P., Verboom, J., Opdam, P. F. M., Hoekstra, R. F., Jochem, R., Arens, P., & Smulders, M. J. M. (2011). Projected climate change causes loss and redistribution of genetic diversity in a model metapopulation of a medium-good disperser. *Ecography*, *34*(6), 920–932.

Dillon, S., McEvoy, R., Baldwin, D. S., Southerton, S., Campbell, C., Parsons, Y., & Rees, G. N. (2015). Genetic diversity of Eucalyptus camaldulensis Dehnh. Following population decline in response to drought and altered hydrological regime. *Austral Ecology*, *40*(5), 558–572.

Frisch, D., Lejeusne, C., Hayashi, M., Bidwell, M. T., Sánchez-Fontenla, J., & Green, A. J. (2021). Brine chemistry matters: Isolation by environment and by distance explain population genetic structure of *Artemia franciscana* in saline lakes. *Freshwater Biology*, *66*(8), 1546–1559.

Fuentes-Pardo, A. P., & Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology*, *26*(20), 5369–5406.

Garrido-Cardenas, J. A., Mesa-Valle, C., & Manzano-Agugliaro, F. (2018). Trends in plant research using molecular markers. *Planta*, *247*(3), 543–557.

Hand, B. K., Muhlfeld, C. C., Wade, A. A., Kovach, R. P., Whited, D. C., Narum, S. R., … Luikart, G. (2016). Climate variables explain neutral and adaptive variation

within salmonid metapopulations: The importance of replication in landscape genetics. *Molecular Ecology*, *25*(3), 689–705.

Hodel, R. G. J., Segovia-Salcedo, M. C., Landis, J. B., Crowl, A. A., Sun, M., Liu, X., … Soltis, P. S. (2016). The report of my death was an exaggeration: A review for researchers using microsatellites in the 21st century. *Applications in Plant Sciences*, *4*(6), 1600025.

Jangjoo, M. (2018). Spatial and temporal patterns of neutral and adaptive genetic variation in the alpine butterfly, *Parnassius smintheus*. *Electronic Thesis and Dissertation Repository*.

Jangjoo, M., Matter, S. F., Roland, J., & Keyghobadi, N. (2016). Connectivity rescues genetic diversity after a demographic bottleneck in a butterfly population network. *Proceedings of the National Academy of Sciences*, *113*(39), 10914–10919.

Jangjoo, M., Matter, S. F., Roland, J., & Keyghobadi, N. (2020). Demographic fluctuations lead to rapid and cyclic shifts in genetic structure among populations of an alpine butterfly, *Parnassius smintheus*. *Journal of Evolutionary Biology*, *33*(5), 668–681.

Jordon-Thaden, I. E., Beck, J. B., Rushworth, C. A., Windham, M. D., Diaz, N., Cantley, J. T., … Rothfels, C. J. (2020). A basic ddRADseq two-enzyme protocol performs well with herbarium and silica-dried tissues across four genera. *Applications in Plant Sciences*, *8*(4), e11344.

Keyghobadi, N., Roland, J., & Strobeck, C. (1999). Influence of landscape on the population genetic structure of the alpine butterfly *Parnassius smintheus* (Papilionidae). *Molecular Ecology*, *8*(9), 1481–1495.

Medeiros, B. A. S. de, & Farrell, B. D. (2018). Whole-genome amplification in double-digest RADseq results in adequate libraries but fewer sequenced loci. *PeerJ*, *6*, e5089.

Oliveira, E. F., Martinez, P. A., São-Pedro, V. A., Gehara, M., Burbrink, F. T., Mesquita, D. O., … Costa, G. C. (2018). Climatic suitability, isolation by distance and river resistance explain genetic variation in a Brazilian whiptail lizard. *Heredity*, *120*(3), 251–265.

Proft, K. M., Bateman, B. L., Johnson, C. N., Jones, M. E., Pauza, M., & Burridge, C. P. (2021). The effects of weather variability on patterns of genetic diversity in Tasmanian bettongs. *Molecular Ecology*, *30*(8), 1777–1790.

Puckett, E. E. (2017). Variability in total project and per sample genotyping costs under varying study designs including with microsatellites or SNPs to answer conservation genetic questions. *Conservation Genetics Resources*, *9*(2), 289–304.

Roland, J., Keyghobadi, N., & Fownes, S. (2000). Alpine parnassius butterfly dispersal: Effects of landscape and population size. *Ecology*, *81*(6), 1642–1653.

Roland, J., & Matter, S. F. (2016). Pivotal effect of early-winter temperatures and snowfall on population growth of alpine *Parnassius smintheus* butterflies. *Ecological Monographs*, *86*(4), 412–428.

Sabina, J., & Leamon, J. H. (2015). Bias in Whole Genome Amplification: Causes and Considerations. In T. Kroneis (Ed.), *Whole Genome Amplification: Methods and Protocols* (pp. 15–41). New York, NY: Springer.

Schlötterer, C. (2004). The evolution of molecular markers—Just a matter of fashion? *Nature Reviews Genetics*, *5*(1), 63–69.

Sunde, J., Yıldırım, Y., Tibblin, P., & Forsman, A. (2020). Comparing the performance of microsatellites and RADseq in population genetic studies: Analysis of data for pike (*Esox lucius*) and a synthesis of previous studies. *Frontiers in Genetics*, *11*, 218.

Wang, I. J., & Bradburd, G. S. (2014). Isolation by environment. *Molecular Ecology*, *23*(23), 5649–5662.

Whitney, K. D., Mudge, J., Natvig, D. O., Sundararajan, A., Pockman, W. T., Bell, J., … Rudgers, J. A. (2019). Experimental drought reduces genetic diversity in the grassland foundation species *Bouteloua eriopoda*. *Oecologia*, *189*(4), 1107–1120.

Appendix A

Table A1 Barcoded Illumina adapter and PCR primer sequences used during double digest restriction site associated DNA sequencing. Barcodes are bolded for all sequences. P1 adapters ligated to *Nla*III cut sites, and P2 adapters ligated to *Eco*RI cut sites. PCR primers were Illumina i5 (forward) and i7 (reverse), with added barcodes on the reverse primer. Individuals pooled and sequenced in a single Illumina Hiseq lane all received a unique combination of adapter and primer barcodes.

Adapter/primer	Sequence
Adapter P1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CATG**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**TGCA**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**ACGT**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**GTAC**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CTAG**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**AGCT**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**TCGA**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**GATC**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**GGTTG**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**GAGTGG**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**GCCAAT**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CCAACA**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CTGTAA**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**AGTTCCA**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**TATAATG**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**TACAGCA**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CGATGTA**CATG
	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGTAGCACATG
Adapter P2	AATT**CATG**AGATCGGAAGAGCGAGAACAA
	AATT**TGCA**AGATCGGAAGAGCGAGAACAA
	AATT**ACGT**AGATCGGAAGAGCGAGAACAA
	AATT**GTAC**AGATCGGAAGAGCGAGAACAA
	AATT**TCGA**AGATCGGAAGAGCGAGAACAA
Forward Primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG
Reverse Primer 1	CAAGCAGAAGACGGCATACGAGAT**ACATCG**GTGACTGGAGTTCAGACG TGTGC
Reverse Primer 2	CAAGCAGAAGACGGCATACGAGAT**ATTGGC**GTGACTGGAGTTCAGACG TGTGC

Appendix B

Table B2 Trinotate gene ontology functional categories and possible specific functions, assigned to the 39 expressed *Parnassius smintheus* loci genotyped on the Sequenom MassARRAY iPLEX platform.

Locus ID	Category	Possible functions
CLocus_99187	Cell function	Apoptosis; transcription; chaperone protein
CLocus_9025	Cell function	Membrane protein
CLocus_28174	Cell function	Phosphate bond hydrolysis
CLocus_162616	Cell function	Cell cycle regulation
CLocus_40552	Cell function	ATP synthase
CLocus_27673	Cell structure	Cell movement; neuron development
CLocus_15073	Development	Segmentation protein runt
CLocus_70308	Developmental	Insect development (*Drosophila*)
CLocus_63430	DNA repair	DNA repair
CLocus_263637	DNA repair	Histone related
CLocus_124730	Heat shock protein	DNAj Heat Shock Protein
CLocus_689	Histone	Histone
CLocus_45337	Immune	Innate immunity by binding to peptidoglycans
CLocus_92401	Immune response	Putative defense protein, antimicrobial
CLocus_111905	Metabolism	Nicotine response – mitochondrial
CLocus_141537	Metabolism	Protease
CLocus_29804	Metabolism	Cytochrome b; ETC

CLocus_77750	Metabolism	Glycogen/glycan biosynthesis
CLocus_22919	Metabolism	Lipid catabolism
CLocus_91307	Organism-level	Microtubule; involved in movement/hearing
CLocus_4488	Organism-level	Neurotransmission (exocytosis)
CLocus_28873	Organism-level	Signal transduction
CLocus_111588	Organism-level	Cuticle melanization and sclerotization
CLocus_6191	Organism-level	Hearing (cilia); courtship
CLocus_87605	Organism-level	Insect molting hormone
CLocus_46539	Organism-level	Insect molting hormone
CLocus_1032489	Polymerase	RNA directed polymerase
CLocus_986766	Polymerase	Polymerase
CLocus_69474	Protein modification	Modifies cullin proteins
CLocus_348206	Protein modification	Protein glycosylation
CLocus_45989	Transcription	mRNA processing and export
CLocus_2336	Transcription	DNA binding
CLocus_113788	Transcription	Transcription regulation
CLocus_81019	Transcription	Zinc finger protein
CLocus_119355	Transcription	Zinc finger protein
CLocus_130947	Transcription	Mitochondrial transcription regulation
CLocus_1050198	Transport	Microtubule based transport
CLocus_953305	Transport	Transmembrane transport
CLocus_184151	Transport	Potassium channel

Curriculum Vitae

Name:	Mel Lucas
Post-secondary Education and Degrees:	Queen's University Kingston, Ontario, Canada 2010-2014 B.Sc. The University of Western Ontario London, Ontario, Canada 2014-2021 Ph.D. (in progress)
Honours and Awards:	NSERC Canada Graduate Scholarship 2015-2016 Queen Elizabeth II Graduate Scholarship in Science and Technology 2017-2018 Province of Ontario Graduate Scholarship 2018-2019
Related Work Experience	Teaching Assistant The University of Western Ontario 2014-2019 Sessional Instructor (Molecular Ecology) The University of Western Ontario 2021
Conference Presentations	Levels of missing data in SNP datasets affect some population genetic analyses in the alpine butterfly *Parnassius smintheus*. Canadian Society for Ecology and Evolution Meeting, Guelph, Canada (2018) Temporal and spatial patterns of genetic diversity in the alpine butterfly *Parnassius smintheus.* The Lepidopterists' Society, Ottawa, Canada (2018) The effect of landscape on genetic differentiation in western Alberta populations of the alpine butterfly *Parnassius smintheus*. Canadian Society for Ecology and Evolution Meeting, Victoria, Canada (2017)

Landscape composition and configuration affect genetic differentiation in the alpine butterfly *Parnassius smintheus*. Entomological Society of Ontario, Guelph, Canada (2017)

Comparing genetic and genomic approaches to assessing population structure and connectivity in an alpine butterfly. Canadian Society for Ecology and Evolution, St John's, Canada (2016)

Assessing the spatial genetic structure of populations of the alpine butterfly *Parnassius smintheus* using RAD sequencing. Canadian Society of Zoologists, London, Canada (2016)