Western Washington University
Western CEDAR

WWU Graduate School Collection | WWU Graduate and Undergraduate Scholarship

2011

Comparative genomics of Bistorta vivipara

Daniel F. Bronny
Western Washington University

Follow this and additional works at: https://cedar.wwu.edu/wwuet

 Part of the Biology Commons

Recommended Citation

Bronny, Daniel F., "Comparative genomics of Bistorta vivipara" (2011). *WWU Graduate School Collection*. 178.
https://cedar.wwu.edu/wwuet/178

This Masters Thesis is brought to you for free and open access by the WWU Graduate and Undergraduate Scholarship at Western CEDAR. It has been accepted for inclusion in WWU Graduate School Collection by an authorized administrator of Western CEDAR. For more information, please contact westerncedar@wwu.edu.

COMPARATIVE GENOMICS OF BISTORTA VIVIPARA

By

Daniel F. Bronny

Accepted in Partial Completion
Of the Requirements for the Degree
Master of Science

Moheb A. Ghali, Dean of the Graduate School

ADVISORY COMMITTEE

Chair, Dr. Eric DeChaine

Dr. Dietmar Schwarz

Dr. Jeff Young

MASTER'S THESIS

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Western Washington University, I grant to Western Washington University the non-exclusive royalty-free right to archive, reproduce, distribute, and display the thesis in any and all forms, including electronic format, via any digital library mechanisms maintained by WWU.

I represent and warrant this is my original work, and does not infringe or violate any rights of others. I warrant that I have obtained written permissions from the owner of any third party copyrighted material included in these files.

I acknowledge that I retain ownership rights to the copyright of this work, including but not limited to the right to use all or part of this work in future works, such as articles or books.

Library users are granted permission for individual, research, and non-commercial reproduction of this work for educational purposes only. Any further digital posting of this document requires specific permission from the author.

Any copying or publication of this thesis for commercial purposes, or for financial gain, is not allowed without my written permission.

Daniel F. Bronny
November 14, 2011

COMPARATIVE GENOMICS OF BISTORTA VIVIPARA

A Thesis
Presented to
The Faculty of
Western Washington University

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science

By

Daniel F. Bronny
November 2011

ABSTRACT

High Northern latitudes are predicted to change considerably in forthcoming climate scenarios, and empirical evidence detailing a species' capacity to cope with extreme variability is needed. Tundra plants make for an excellent study because their genetic histories were impacted by the dramatic transitions of historic glacial and interglacial ages. Here, thousands of restriction site-associated DNA (RAD) markers from geographically isolated Alaskan (Arctic) and Coloradan (Alpine) *Bistorta vivipara* (Polygonaceae) populations are compared in an investigation of evolutionary response to rapid climate change.

Non-coding nuclear markers were analyzed in a coalescent framework to estimate an effective ancestral population size (N_a) and divergence date (*t*) of the two populations of ~23 000 individuals and ~140 000 years before present. Nucleotide substitutions per synonymous site (dS) and nonsynonymous site (dN) were calculated for putative orthologous protein-coding sequences to determine the form of selection acting on the subsampled genome in the context of *t*. Most sequences were either 100% conserved or exhibited dS>dN, suggesting purifying selection. The few sequences suggesting positive selection (dS<dN) were identified as retroelements, which are expected to escape purifying selection. There were two exceptions: a putative protein phosphatase and a kinase involved with steroid signaling. The results suggest genetic adaptation is not a readily apparent option for *B. vivipara*'s response to climate change. This, and other organisms whose habitats will shift quickly or disappear, may depend on demographic and plastic responses as alternatives to extinction.

ACKNOWLEDGEMENTS

I am grateful to all who contributed to the development of this work, including:

Eric DeChaine, for leading me to the marvels of evolution;

Dietmar Schwarz and Jeff Young, for offering their time and energy to my growth;

Charles Davis, for connecting me to a talented, supportive, and diverse academic community;

Tallen Xi, who helped with the molecular methods;

Matt Fujita, Frank Reindt, Ingrid Soltero, and Maude Baldwin, for being remarkable teachers and friends;

Kurt Galbreath, who weathered my questions graciously;

Jody Hey, who developed an analytical tool central to this investigation and helped troubleshoot its use;

Joann Otto, whose perspective was a refuge during tempests of uncertainty;

Bruce Clinton and Arlan Norman, without whom this research would never have found me;

the faculty and staff of the WWU Biology department, for their dedication to the graduate student body;

the Hodgson family, for actively supporting plant research;

the Bronny family, for encouraging me as long as I can remember;

and,

Katie Wall, through whom I see my world more profoundly, which gave me the courage I needed every single day.

TABLE OF CONTENTS

ABSTRACT……iv

ACKNOWLEDGEMENTS……...……v

LIST OF FIGURES…….vii

LIST OF TABLES…….x

INTRODUCTION…….1

METHODS…….17

RESULTS…….50

DISCUSSION……..59

LITERATURE CITED…….79

APPENDIX……..89

LIST OF FIGURES

Figure 1. Past Milankovitch cycles and temperature proxies. Q^{-day} is the daily-averaged insolation at the top of the atmosphere on the day of the summer solstice at 65°N latitude calculated from orbital parameters. Sediment and glacial ice show two proxies for past global temperature, from ocean sediment and Antarctic ice respectively. A reduction of d18O indicates warmer conditions. The vertical gray line is current conditions, 2011 C.E. Adapted from Incredio (2009) with permission.……2

Figure 2. Pollen maps of undifferentiated *Picea* (Spruce) at 4 000 calibrated (cal) year (yr) intervals between 17 000 and 5 000 cal yr before present demonstrating habitat tracking. White represents regions with no data, and light blue represents ice. Pollen data comes from Williams *et al*. 2004, images generated by Pollen Viewer 3.2 (Leduc 2003)……3

Figure 3. *Bistorta vivipara* (L.) S. F. Gray scaled to half actual size (from Polunin 1959). Individuals can be highly variable, but typically display leaves of lustrous green above, grayish below, a glabrous stem, and white or pink flowers with bulbils replacing the lower flowers.……...9

Figure 4. Range map of *B. vivipara* compiled from numerous sources (Hultén 1971) with sampling sites highlighted. (Note: the synonym *P. viviparum* has since been segregated from *Polygonum* and placed in the genus *Bistorta*; it is treated as such by the Flora of North America Editorial Committee [2005] and accepted by the International Code of Botanical Nomenclature). Arctic and Alpine populations flow together fairly well in Europe and Asia via numerous intermediate localities, whereas North American Alpine populations, restricted to the high peaks of the Alaska Range, northern Coastal Range, Cascades, and Rocky Mountains, are comparatively discontiguous from North American Arctic populations. One *B. vivipara* was sampled from each numbered site in August 2008: 1) Red Mountain Pass, Colorado (37°53'54"N, 107°42'43"W), elevation 3383 m; and, 2) Noatak River, Alaska (67°58'3"N, 161°51'48"W), elevation 100 m.……..11

Figure 5. Targeted fragment architecture. PEs are Illumina paired-end adapter sequences which bound the fragments to the flowcell during bridge amplification and served as primers for sequencing-by-synthesis; MIDs are molecular identification tags that enabled bioinformatic separation of Arctic and Alpine samples; RSs are remnants of the *PsiI* restriction site (5'-TAA); parenthetic numbers are the segment lengths in bp; the arrows below the fragment indicate the direction, start, and stop locations of the two 101-bp reads; and Xs are portions of the fragment that were not sequenced. The DNA used in comparative analyses begins 7 bp into each read, after the MID and RS. For each 400-500 bp fragment, the last bp sequenced in one read was 120-220 bp away from the last bp sequenced in its opposite. The center of most fragments remained unsequenced.……...21

Figure 6. Workflow for assembling Illumina reads into RAD markers, pairing homologs across samples, and selecting sequences to compare. A) Millions of q.c.'d Illumina reads (dashes) were collapsed into thousands of RAD markers (shapes) with corresponding

coverage information. B) Homologous Arctic and Alpine markers were paired together via a megablast across samples. Zero, one, or several sequences generated hits. Groups of similar sequences from a single sample (cluster families) have presumed origins in gene or genome duplications. C) The longest (or only) Arctic sequence in a cluster family is paired with the Alpine homolog with the highest percent identity (%ID) for final comparative analysis. Grey bars are sequences from the Arctic and Alpine individuals; vertical lines are polymorphic sites with respect to the Arctic sequence. %ID is calculated as identical columns / sequence length. Note the effect of incomplete data on pairing orthologs: 2 is orthologous and 1 and 3 are paralogous, but 2, truncated during q.c., is erroneously passed over for 1.……...26

Figure 7. A census of sequences per cluster family. A histogram counting cluster families with up to 10 members and a boxplot (insert) showing the distribution of cluster families with more than 10 sequences serves to gauge the completeness of my genomic survey. Despite polyploidy, the majority of contigs are the sole member of a cluster.…..……..31

Figure 8. The IMa2 model depicting 6 possible demographic parameters (Hey and Nielsen 2004). N_1, N_2, and N_a are constant effective population sizes, m1 and m2 are gene flow rates, and t is the time of population splitting. Parameters evaluated in this investigation are circled; others were either removed from the model or provided meaningless values….....40

Figure 9. Combined probability curves for time since splitting, *t*, and ancestral effective population size, N_a, from 50+ Markov Chain Monte Carlo (MCMC)-based IMa2 runs each for paired-end (PE) data sets 1 and 2, generated with alternate priors. For the curves in A, priors were set to ~700 000 years and ~80 000 individuals; for B, priors were set to ~1.4 million years and 200 000 individuals, with the exception of (*), which used a prior of 150 000 individuals. For all runs, the mutation rate range priors were estimated from the literature and encompassed 3 orders of magnitude. Each simulation began with 200 pairs of non-coding, unlinked, neutral nuclear markers with an average size of 94 bp. After a burn-in of 230 000 steps, the simulations ran for 2.1 million iterations, saving the parameter values and genealogies every 100 steps, for a total of 21 000 saved parameter value sets per simulation. Summary histograms were made for each batch of 200 loci (not shown) in which the x-axis was the target parameter from 0 to the prior divided into 1000 bins, and the y-axis was based on the likelihood of each parameter value occurring in the numerous saved geneologies. The 50+ summary curves from either PE1 or 2 were combined by plotting the sum of the probabilities at each bin against the average bin value from all the runs. The upper (a) and lower (c) bounds of the highest posterior density 95% (HPD95%) interval (black) span a distance on the x-axis that has a 0.95 probability of covering the actual value of *t* or N_a, and may be interpreted as confidence intervals. The peaks of the curves (b, red) correspond to the most likely value of each parameter, given all PE1 or 2 data. Here, the best estimate of *t* is ~140 000 years before present, and N_a is ~23 000 individuals…….....52

Figure 10. Relative proportion of P pairs under purifying selection (green) and positive selection (gold). Absolute counts are given in parentheses…..…….……..……..……...54

Figure 11. Intensity of selection on CG pairs depicted as the distribution of dS/dN values.

The median for pairs under positive selection (dS/dN <1, n=25) was 0.675, the median for pairs under purifying selection (dS/dN > 1, n=95) was 3.975. The y-axis is logarithmic; circles are outliers…...…….54

Figure 12. A) A correlation of chronostratigraphical subdivisions showing, from left to right, formal time divisions of the Middle and Late Pleistocene subseries, North American Stages, and Marine Isotope Stages (MISs), redrawn from Cohen and Gibbard (2011). Solid horizontal lines indicate observed boundaries, the red dashed line indicates the divergence date estimate (*t*), and the grey horizontal bar marks the 95% confidence intervals thereof. B) MIS 7 at a higher resolution. Various proxy climate records report multiple interglacial peaks during MIS 7, a milder interglaciation than the Sangamonian. Redrawn from Lang and Wolff (2011)……...60

LIST OF TABLES

Table 1. A summary of the primary results delivered by this investigation. The variance around the mean value from both paired-end (PE) sets is given where applicable.……55

Table 2. Bioinformatics results for paired-end (PE) data sets.……...56

Table 3. Alignment statistics for mitochondrial (M), chloroplast (C), protein (P), and nuclear non-coding (NC) markers. Total columns refers to all columns aligned in each category for PE1 and 2 combined, not counting gaps. The percent of polymorphic sites for the nuclear genome (P and NC combined, not shown) is 2.10%.....57

Table 4. Adaptive analysis outcomes for paired-end (PE) 1 and 2 data sets…….….58

INTRODUCTION

Do evolutionary changes occur in response to rapid environmental change?

The Anthropocene marks the ongoing global impact of our society, yet our assessment of how the planet's species might respond to forthcoming climate scenarios lacks a thoroughly genetic understanding—this, despite the connections between evolutionary response and the persistence of populations given future change (Geber and Dawson 1993, Travis and Futuyama 1993). In a post-genomic era where predictions warn of environmental change at scales that eclipse the extremes of bygone periodic ice ages (Jansen *et al*. 2007), we can sequence large portions of non-model genomes, infer evolutionary histories with respect to historic environmental change, and register, to some degree, their population demographic and adaptive response potentials.

Consider an environment as a multidimensional space whose *n*-dimensions are defined by myriad factors, many of which are climatological (Hutchinson 1957). Strong climate trends, like rapid warming or cooling, alter the 'shape' of this conceptual environmental space and challenge organisms whose fundamental niches—sets of environmental factors that permit their survival and reproduction—are compromised or no longer contained therein. Some responses to this challenge leave genetic signatures that can be recovered using high-throughput DNA sequencing. In this investigation, I observed evolutionary changes in genomes that persisted through historic climate cycles. The results are presented in terms of a species' response to environmental variability of the past in order to understand how it might respond to variability of the future.

Historical context

The Quaternary (the last 2.6 million years) is characterized by its alternating glacial (100,000 years) and interglacial (20,000 years) ages (durations are approximate). During glacial ages, cool summers failed to offset the glacial growth of winter and the northern portions of North America, Europe, and Asia were buried in ice. The movements of the ice sheets literally left their marks in stone, and are further evident in fossil and pollen records showing the presence and absence of species with predictable environmental tolerance ranges over time (Pielou 1991).

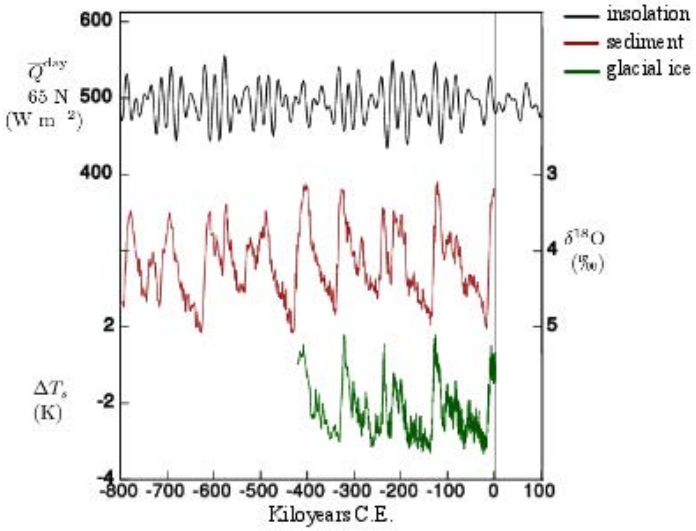

Figure 1. Past Milankovitch cycles and temperature proxies. Q^{-day} is the daily-averaged insolation at the top of the atmosphere on the day of the summer solstice at 65°N latitude calculated from orbital parameters. Sediment and glacial ice show two proxies for past global temperature, from ocean sediment and Antarctic ice respectively. A reduction of d18O indicates warmer conditions. The vertical gray line is current conditions, 2011 C.E. Adapted from Incredio (2009) with permission.

The mechanism driving the glaciations was revealed in Milankovitch's discovery of orbital forcing. He described three gradual astronomical cycles that influence the amount of sunlight reaching Earth: the precession of the axis, the obliquity of the axis, and the eccentricity of the orbit. By modeling a combination of these periods, he deduced historic global insolation, and thus temperature (Hays *et al*. 1976). Since then, empirical evidence of these climate cycles has been recovered from marine sediment and glacial ice cores using temperature-indicative isotope ratios in diatom shells or atmospheric gases (Bradley 1985). Generally, the paleoecological records and Milankovitch cycles agree: there were periodic

trends of warming and cooling (Fig. 1).

Move, adapt, or die

How species responded to the oscillating glacial cycles can be broadly categorized as demographic, adaptive, or neither, due to extinction (Jackson and Overpeck 2000).

Demographic responses affect the size, structure, and distribution of a population or species; possible outcomes might include the isolation of one population from another or the redefinition of a species' range. There are numerous examples of species migrating in step with their preferred habitats during the Quaternary, including biome reconstructions showing latitudinal shifts over time (Shuman *et al*. 2002) (Fig. 2). This kind of habitat tracking is a classic demographic response.

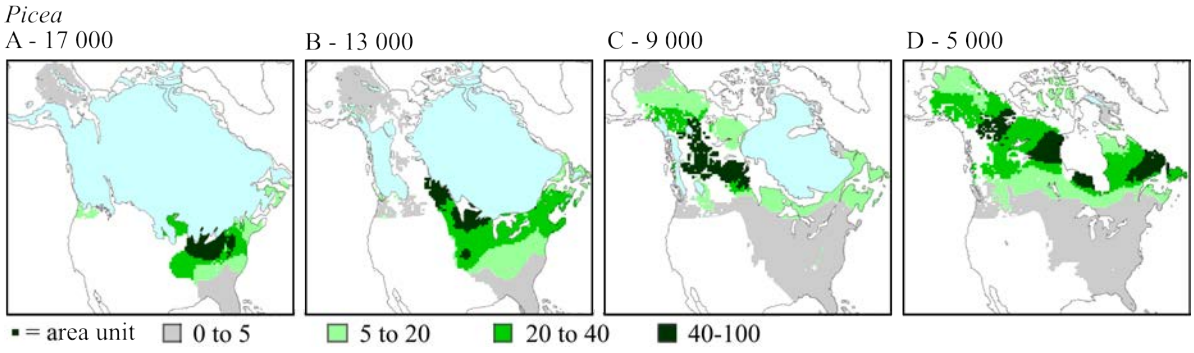

Figure 2. Pollen maps of undifferentiated *Picea* (Spruce) at 4 000 calibrated (cal) year (yr) intervals between 17 000 and 5 000 cal yr before present demonstrating habitat tracking. White represents regions with no data, and light blue represents ice. Pollen data comes from Williams *et al.* 2004, images generated by Pollen Viewer 3.2 (Leduc 2003).

Adaptive responses occur when a population alters its fundamental niche to fit within the constraints of a new environmental space. Population level adaptive changes (e.g. population fitness) and individual responses (e.g. phenotypic plasticity) are facilitated by genetic mutation and natural selection and can be investigated at the level of nucleotide

sequence data. This is because fixed nucleotide changes in population level data can represent a phenotype, and thus fundamental niche, better suited to a particular environmental space, and instances of phenotypic plasticity, where a broad fundamental niche exceeds the extremes of the current environment, are underpinned by a genetic capacity for that niche.

Finally, if a population can not move to suitable habitat, adapt, or otherwise maintain a viable effective population size (N_e) within the constraints of a new environment, then it experiences local and/or global extinction.

Filling the gap

The importance of being able to respond via adaptation increases when the rate or magnitude of environmental changes outpace a species' ability to disperse to suitable habitat. It is even more critical when suitable habitat disappears entirely. Simply put, adaptation can prevent local extinction. Despite this, the literature is biased against genetic adaptation as a response to rapid environmental change, with examples of migration (Coope 1995, West 1980), phenotypic plasticity (Charmantier *et al*. 2008, Moyes *et al*. 2011) and extinction (reviewed in Barnosky *et al*. 2004, Jackson and Weng 1999) being more common. Far less work has been done to understand Darwinian evolution—i.e., selection of heritable traits—in populations that remain in place when the climate changes. Perhaps the bias is an issue of tractability. Until now, it has been difficult to observe genetic adaptation in organisms with long generation times, especially in the context of environmental variability (although, see Bennington and McGraw 1993 and Davis and Shaw 2001), and clear-cut evidence indicating a significant role for evolutionary adaptation to ongoing climate warming is conspicuously

scarce (Gienapp *et al*. 2008).

Comparative genomics provides a new perspective on the past by enabling the mass investigation of genomic changes in populations over time. The field relies on state-of-the-art high-throughput DNA sequencing technology to boost one of biology's oldest and most successful methods (Haubold and Wiehe 2004)—that of comparing closely related organisms to infer function and evolutionary forces (Wiener 1994). Recently, researchers combined a classic molecular method—a restriction digest to subsample a genome at enzymatic recognition sites—with Illumina technology to jump start the field of comparative population genomics (Baird *et al*. 2008). Their innovative restriction-site associated DNA (RAD) markers extended the platform's high-throughput capabilities from sequencing individual genomes to generating thousands of genome-wide markers from multiple individuals.

Using RAD marker methods as a foundation, I developed an application of the modern comparative genomics toolkit for a population that has had the potential for both a demographic and adaptive response to historic climate change. In the process, I observed genetic structures influenced by the redistribution, mixing, isolation, and adaptive pressures of North American glaciers with resolution that has only recently become possible due to advances in nucleotide sequencing technology (next-generation sequencing).

Study system

The Arctic provides an excellent study system for climate change research. It is relatively unimpeded by direct anthropogenic influence and comes with a well-documented climate history. Life is spread thin and low to the ground in an ecosystem of herbaceous, dwarf shrub, or lichen vegetation where summers are too cold to allow tree growth (Billings

1974). These conditions are echoed, with some variation, in lower latitudes at high altitudes as Alpine tundra. Together, Arctic and Alpine vegetation cover approximately 8% of the planet's terrestrial surface (Körner 1995) with 1629 Arctic and over 10,000 Alpine pteridophytes, gymnosperms, monocots, and dicots (Löve and Löve 1975, Walker 1995). These organisms are influenced largely by abiotic factors (e.g., wind, radiation, freeze/thaw cycles), rather than biotic interactions (e.g., grazing, modified microclimate from other plants) (Billings 1974), and a good body of research on their natural history and adaptations to extreme conditions exists (Borgen and Bengt 1997, Ives and Barry 1974).

Furthermore, the Arctic is experiencing the fastest rate of temperature increase on the globe today (Comiso 2002), and, perhaps rightfully, is gathering cultural and scientific attention. The importance of Arctic history has been clearly stated in the context of making reliable predictions about its future. Previous research intending to assess the potential responsiveness and resilience of arctic ecosystems draws on environmental manipulation experiments, eco-physiological and plant demographic studies, and a consideration of paleoecological and pedogenic processes in the high Arctic (Arft *et al*. 1999, Parsons *et al*. 1995, Robinson *et al*. 1998, Walker *et al*. 2005 Wookey *et al*. 1995). My project will increase our understanding of historic processes in the Arctic/Alpine ecosystem as we progress from paleoecological data, through genetic research, and into next-generation sequencing, bioinformatics, and population genomics.

Arctic and Alpine plants have already persisted through extremes, with population fragmentation and redistribution playing a key role in their recent histories. Biologists as early as Darwin (1859) postulated that high-latitude residents were pushed southward by advancing glaciers. Additionally, as massive volumes of water were locked up in glacial ice,

ocean levels dropped and redefined continents' coasts. This revealed the landmass, from the Lena river in northeast Asia to the Mackenzie river in the Yukon territory, now partially submerged by the Bering Strait. It was coined 'Beringia,' and its ecological significance was championed early by the prominent cryophytologist, Eric Hultén (1937). Evidently ice-free through the ages (Hamilton and Thorson 1983, Hopkins 1967, Tarasov 2000), Beringia played an important role as a glacial-age refuge for plants and animals (Abbott *et al*. 2000, Brubaker *et al*. 2005). So, not only were high-latitude residents pushed southward by advancing ice, but they may also have moved into refugial pockets like Beringia. Other proposed havens include coastal refugia, where coastal ranges prevented glacial ice from reaching the sea, and mountaintop islands in the ice, called nunatuks, where hardier plants may have survived (Pielou 1994).

The particulars of advancing and retreating ice during glacial-interglacial transitions also affected the distributions and population dynamics of ice age organisms. For example, the Laurentide and Cordilleran ice sheets initiated and moved in from the East and West, respectively, leaving an ice-free corridor between the Cascades and Rocky Mountains until the two continental sheets fused (Fig. 2B). Whether the glaciological corridor was an ecologically inviting one is debated. Irrespective of population richness, this area of bogs, marshes, and icy rivers and lakes connected the refugia north and south of the glaciers several times during the Quaternary (Pielou 1994).

The impact of these scenarios on the Arctic/Alpine ecosystem was undoubtedly large, but the general biotic response was likely different on a species by species basis (Stewart *et al*. 2010, Taberlet *et al*. 1998). High-latitude vegetation history has been extensively studied (Abbott and Brochmann 2003, Comes and Kadereit 1998, Hopkins *et al*. 1982, Jackson *et al*.

1997, Simakov 2002) and the emerging picture is complex. Yet, as more histories are investigated, a more complete understanding of this rapidly changing landscape is revealed.

For taxa whose gene flow is restricted due to geographic isolation and limited dispersal capabilities, genetic structure is dependent on historical events (Schaal *et al*. 1998). I chose one such arctic resident and looked to its DNA to infer what happened in the past to result in the current state of its genome.

Target species

Bistorta vivipara (Polygonaceae), commonly called Viviparous Knotweed or Alpine Bistort, is one of the most ubiquitous and characteristic of Arctic plants. The plant is a long term Arctic resident. Fossil evidence of this species persisting in the Yukon Territories through the last interglacial age (approximately 120k years ago) and glacial maximum (approximately 18k years ago) is confirmed (Zazula *et al*. in press). It was named for its habit of generating self-germinated bulbils attached to the parent plant, although this name is somewhat misleading. Vivipary implies seeds or fruit that sprout before they fall from the parent. Although this occurs in *B. vivipara*, it is not exclusively the case, giving rise to what would more correctly be described as false vivipary.

The plant has an unbranched stem, 4-30 cm high, bearing a terminal spike-like raceme of white or pinkish flowers. The inflorescence is 4-8 mm in diameter, and the flowers are small, 2.5-4.0 mm long, with seed-like fruits (bulbils) replacing the lower flowers. The leaves, 1-10 cm long, are linear-lanceolate to subrotund, normally tapering at the base, with upper leaves being reduced and sessile (Polunin 1959) (Fig. 3).

Bistorta vivipara is similar to other plants thriving at low temperatures. These plants, called cryophytes, evolved in a treeless landscape and are metabolically inhibited above 25°C (Billings, 1974). They face low temperatures, short growing seasons, sudden changes in growing season length, limited nutrients, and spatial heterogeneity, or 'oligotrophic' (Henry *et al*. 1986) environments. Generally, these obstacles are all overcome. Cryophytes are low-lying and small, staying safe from wind and generating a protective microphytoclimate (Billings 1974). Their size reduces the need for overall primary production, and most of them are perennial, reducing the need to make a new photosynthetic apparatus each year. Polyploidy is common in high-latitude plants, and many are capable of facultative vivipary, so that a new seed set is not required for reproduction. This, and other mechanisms to reproduce without relying on exchanging gametes with another individual, overcome the instability of factors like pollinator interaction and are widely employed. In fact, 'selfing' is so common in the tundra that there is not a single heterostylous, or self-incompatible, arctic population (Baker 1959). *Bistorta vivipara* displays all of these cold-adapted characteristics.

Figure 3. *Bistorta vivipara* (L.) S. F. Gray scaled to half actual size (from Polunin 1959). Individuals can be highly variable, but typically display leaves of lustrous green above, grayish below, a glabrous stem, and white or pink flowers with bulbils replacing the lower flowers.

Cytotaxonomical investigations of Alpine Bistort report complete and incomplete somatic chromosome counts from 80-132 (Löve and Löve 1975), suggesting a 6- to 11- ploid organism. The basic chromosome number for the *Bistorta* genus is 12. Individuals collected

in Northwest Alaska near Ogoturuk creek (Johnson and Packer 1968) and Barrow (Packer and McPherson 1974) contained >100 chromosomes, a Canadian Arctic individual harbored >110 (Mosquin and Haley 1966), and exact counts of 120 are reported from Central Northern Canada (Löve and Ritchie 1966) and Boulder, Colorado (Löve *et al*. 1971). Additional instances of decaploid (2n=120) individuals are recorded in various parts of its Arctic-Alpine distribution, and the species was considered, at one point, authoritatively decaploid (1971).

The current distribution of the target species is nearly circumpolar (Fig. 4), with disjunct Arctic and Alpine populations. This wide geographic separation between them may have occurred during a glacial age, after which, as the climate warmed, the cryophytes occupying more southern lowlands retreated northward with the ice but also found refuge in temperate mountains (Löve and Löve, 1974). While we can only suspect the cause of their current distribution, we can directly investigate its effects by examining the current population genomic structure.

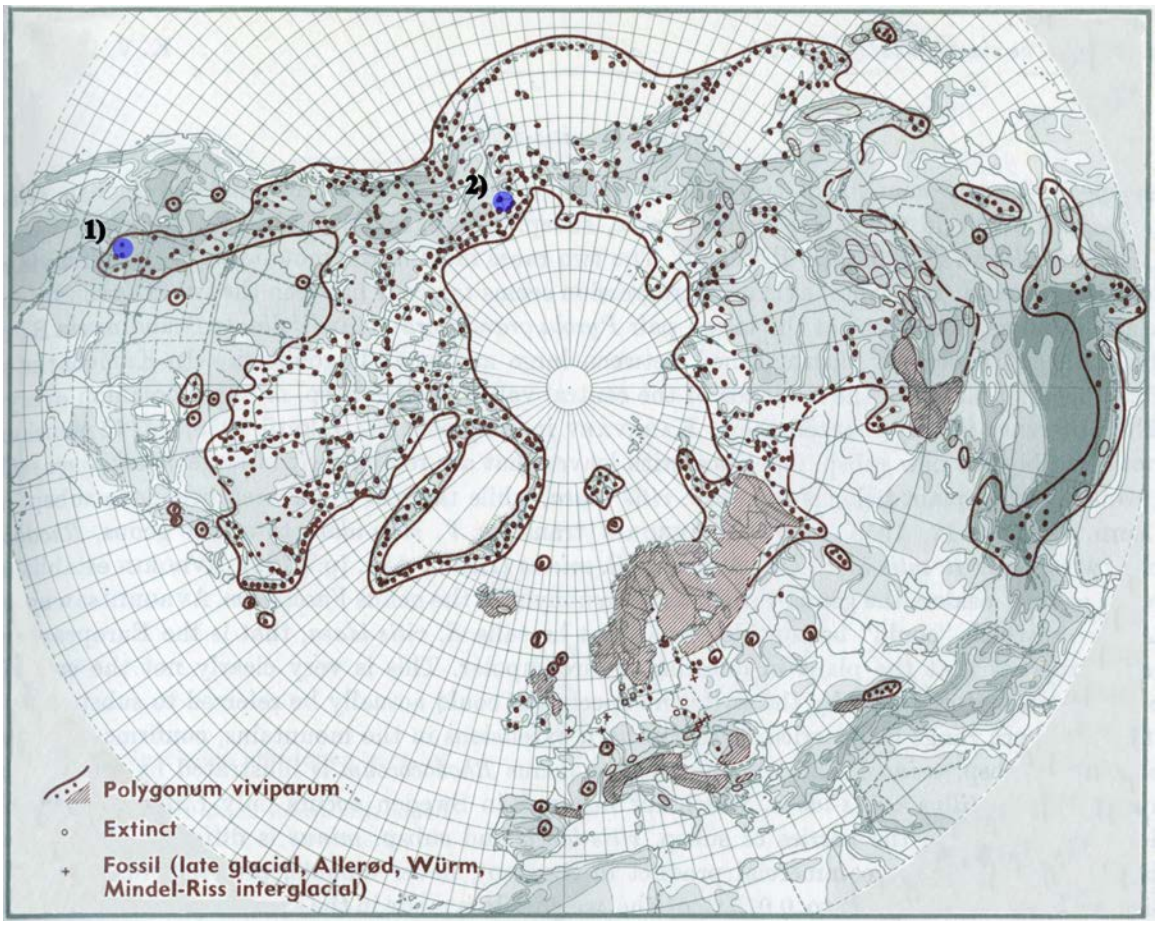

Figure 4. Range map of *B. vivipara* compiled from numerous sources (Hultén 1971) with sampling sites highlighted. (Note: the synonym *P. viviparum* has since been segregated from *Polygonum* and placed in the genus *Bistorta*; it is treated as such by the Flora of North America Editorial Committee [2005] and accepted by the International Code of Botanical Nomenclature). Arctic and Alpine populations flow together fairly well in Europe and Asia via numerous intermediate localities, whereas North American Alpine populations, restricted to the high peaks of the Alaska Range, northern Coastal Range, Cascades, and Rocky Mountains, are comparatively discontiguous from North American Arctic populations. One *B. vivipara* was sampled from each numbered site in August 2008: 1) Red Mountain Pass, Colorado (37°53'54"N, 107°42'43"W), elevation 3383 m; and, 2) Noatak River, Alaska (67°58'3"N, 161°51'48"W), elevation 100 m.

Objective

My goal was to observe genomic changes in a species that has weathered rapid environmental changes. To do this, I generated a large library of DNA markers and observed nucleotide polymorphism between an Arctic and Alpine population, including potentially adaptive substitutions, in a plant affected by Quaternary climate cycles. The potential to respond adaptively stems from isolation in the slightly different environments of Arctic and Alpine tundra; the potential to respond demographically stems from the population shuffling associated with repeated glaciations. So, I addressed the populations' susceptibilities to demographic limitations (e.g. range shifts and fragmentation) due to climate while simultaneously exploring their potential for genetic adaptation; thus acknowledging both of Jackson and Overpeck's (2000) alternatives to extinction in my assessment of a species' capacity to respond to change. With this in mind, I generated the following hypotheses and tested them using the tools and techniques of comparative genomics.

Hypotheses and predictions

Demographic response

- $H1_{alt1}$: *Arctic and Alpine populations of* B. vivipara *are panmictic and exhibit signals of gene flow.*
- $H1_{alt2}$: *Arctic and Alpine populations of* B. vivipara *are genetically isolated and have been split for a number of years,* t.

My first set of hypotheses addressed whether or not Arctic and Alpine populations of *B. vivipara* seem to be isolated from one another. I tested this by generating a robust estimate of *t*, or time since splitting, based on accumulated mutations in non-coding DNA

from the two populations. Support for $H1_{alt1}$ would have looked like a very shallow divergence date, suggesting that the dispersal capabilities of this species connect the Arctic and Alpine gene pools sufficiently to mask any signal of isolation. Support for $H1_{alt2}$ was a deeper estimate of *t*, suggesting that gene flow between these populations was significantly restricted at some point, either by inhospitably warm geographic space during interglacial ages, or by massive ice sheets during glacial ages. I expected to find support for $H1_{alt2}$ in the form of divergent non-coding DNA showing that the populations are or were, in fact, split. I further predicted that the last critical event to severely restrict gene flow and isolate these populations was the climax of the last glacial age, approximately 18 000 years ago, and that this event would correspond with my estimates of *t*.

Support for $H1_{alt2}$ permitted further investigation based on the following hypotheses and provided a temporal context for any potential adaptive response.

Adaptive response

These hypotheses address whether or not adaptive genetic changes have occurred in the genomes of the two populations since splitting. Although similar in many ways, the Arctic and Alpine tundra have some fundamental differences: the photoperiod in the Arctic is three months of continuous sunlight versus the 24-hour day/night cycles experienced at lower latitudes; the diversity and density of pollinators, competitor plants, and other animals is greater in the Alpine ecosystem; and, low-elevation Arctic cryophytes experience less solar radiation than their mountaintop counterparts. Different selective pressures across geographic space may have led to nucleotide polymorphisms that can be observed in the putative protein-coding sequences.

- *$H2_{alt1}$: The Arctic and Alpine populations did not acquire variation in protein-coding sequences since* t, *time since splitting.*
- *$H2_{alt2}$: The Arctic and Alpine populations did acquire variation in protein-coding sequences since* t, *time since splitting.*

The following sub-hypotheses address the form and magnitude of natural selection affecting the markers in the *B. vivipara* genomes:

- *$H3_{alt1}$: The majority of protein-coding sequences in Arctic and Alpine populations are under purifying selection; the minority exhibit signals of positive selection since splitting.*
- *$H3_{alt2}$: The majority of protein-coding sequences in Arctic and Alpine populations are under positive selection; the minority exhibit signals of purifying selection since splitting.*

To test my adaptive response hypotheses, I observed synonymous and nonsynonymous nucleotide substitutions in homologous protein sequences from the two populations. Synonymous nucleotide substitutions occur at codon positions that do not alter the primary sequence of amino acids; nonsynonymous substitutions result in an exchange of amino acids or the insertion of a stop codon. The signature of purifying selection was defined as instances where synonymous substitutions per synonymous site (dS) outnumbered nonsynonymous substitutions per nonsynonymous site (dN); positive selection was defined as the inverse; and neutrality was defined as instances where dS = dN. Support for $H2_{alt2}$ provided a means of calculating dS and dN. Support for $H3_{alt1}$ would have looked like highly conserved protein-coding sequences whose differences did not lead to changes in the protein's primary structure. Support for $H3_{alt2}$ would have looked like a dS/dN ratio below

unity for the majority of the sequences. (Note: dS/dN is used, rather than its inverse, to be consistent with the program that applied the tests for selection [see Methods]).

Given that genetic adaptation to climate over the time frame of Quaternary climate cycles has not yet been shown in plants, I did not expect many markers to exhibit positive selection since splitting. The populations do face some fundamentally different environmental pressures, which may have resulted in some sequences responding via positive selection, but I predicted they would be a minority. Thus, I expected to find support for $H3_{alt1}$.

Research strategy

A library of DNA sequences was generated from two individuals: one *B. vivipara* from the Arctic tundra and one from an Alpine tundra ecosystem. Each genome was subsampled using a restriction digest to generate thousands of homologous markers between the individuals. The pangenomic markers from both individuals were sequenced simultaneously using Illumina's (Solexa©) Genome Analyzer II (GAII) platform. The markers included functional and non-functional DNA, which required annotation and organization prior to analysis. Protein-coding markers were identified and separated from anonymous, non-coding, nuclear data using similarity search algorithms and bioinformatics databases. The numerous remaining markers were analyzed in a coalescent framework to generate an estimate of *t*, the time since splitting. Plastid sequences were identified and treated as linked genetic markers and the frequency of variable sites in the three genomes accessed (nuclear, chloroplast, mitochondrial) was compared. Finally, dS and dN was calculated for all protein-coding sequences to determine the form of selection at play in the

subsampled genome in the context of *t*.

Significance

Multi-locus approaches to generating demographic parameters are relatively new in the literature (Galbreath *et al*. 2011, Peters *et al*. 2008, Xiang *et al*. 2008), and have yet to be effectively employed using hundreds—let alone thousands—of unlinked nuclear loci from two individuals, which places this investigation in relatively uncharted territory. In light of the theory that more loci are more informative than more individuals in coalescent modeling (Felsenstein 2005), my approach has the potential to resolve historic demographic parameters with relatively narrow confidence intervals—limited, perhaps, only by the computational demands of the numerous required simulated genealogies. My application of tests for selection, based on classic work on the expected frequency of nucleotide substitutions at non-synonymous and synonymous codon positions (Nei and Gojobori 1986), to a species with a long generation time in the context of historic climate cycles is novel.

By using a range of genomic techniques on a single target species, I contributed to our emerging understanding of the history of the Arctic ecosystem and helped sharpen the tools at the junction between ecology, global change studies, and Quaternary paleobiology. I did this in the general framework of assessing a species' capacity to respond to environmental change in the hopes that it leads to a better understanding of future biotic responses, brings awareness to the changing Arctic/Alpine tundra, and contributes to our understanding of fundamental evolutionary processes—i.e. genomic change over time.

METHODS

Overview

Generally, comparative genomics relies on a stepwise process of:

1) Generating sequence data;
2) Reconstructing homologous collinearity (i.e. selecting sequences to compare);
3) Aligning multiple sequences; and,
4) Identifying evolutionarily constrained DNA.

(Margulies and Birney 2008)

These steps were my guide as I generated numerous short DNA sequences from the genomes of two geographically isolated individuals, selected and classified the putative homologs, and employed coalescent- and selection-based analyses on these comparative genomic pairs.

Generating sequence data

My molecular methods were based on the Cresko lab's development of restriction-site associated DNA (RAD) markers (Baird *et al*. 2008). RAD methods are similar to analyses using restriction fragment length polymorphisms (RFLPs) and amplified fragment length polymorphism (AFLPs), in that they reduce the complexity of the genome by subsampling only at specific sites defined by restriction enzymes (Davey and Blaxter 2011)—but surpass these methods in cost and time in that RAD libraries are compatible with next-generation sequencing.

To generate sequence data, I: 1) extracted and amplified genomic DNA from two individual *B. vivipara*; 2) digested the DNA with the endonuclease *PsiI*; 3) prepared the

fragments for Paired-End (PE) Illumina Sequencing; 4) enriched a size-selected portion of the fragments; and, 5) sequenced the resulting DNA library in a single Illumina GAII flowcell channel.

Collecting samples

Bistorta vivipara was collected from two populations separated by approximately 4 600 km: Noatak River, Alaska (Arctic) and Red Mountain Pass, Colorado (Alpine) (Fig. 4). Plants were collected by hand in summer 2008 (DeChaine and Walla), kept *in silica*, and transferred to a 4°F Western Washington University Herbarium (WWB) freezer. Voucher specimens were deposited in WWB (to be accessioned). Genomic DNA (gDNA) was purified and amplified from leaf tissue of one Arctic and one Alpine individual using DNeasy and REPLI-g kits (Qiagen) in 2010. REPLI-g kits provide uniform DNA amplification across the entire genome with minimal sequence bias (Hosono *et al*. 2003). An alkaline denaturation reduces damage to the template DNA from heat denaturation, and an overnight isothermal reaction with Phi 29 DNA polymerase provides high fidelity amplification based on Multiple Displacement Amplification (MDA) technology rather than PCR, for highly representative unbiased amplicons.

Subsampling the genome

I digested the DNA product from each plant in a three-hour reaction with fresh *PsiI* (New England Biolabs) following the manufacturers' protocols. A long digest was used to increase the likelihood that many or all of the same recognition sites were cleaved in both samples. *PsiI*'s suitability for extended digests (up to 8 hours) made it appropriate for this

application, and its 6-base recognition site (5'...TTA^TAA...3') was carefully considered. The variety of commercially available enzymes differ in their nucleotide recognition sequences, and each results in a different collection of various-sized fragments in a digest of the same genome. My aim was a subset of gDNA fragments of similar size, so I needed an enzyme that would make a large library of fragments in my target size range, yet not so large as to risk low coverage per locus in the sequencing reaction. The targeted size range, 400-500 base pairs (bp), was the largest suitable insert size for the GAII sequencer and was used to maximize the distance between paired-end reads.

No previous information existed on the size or content of the *B. vivipara* genome, so I used *in silico* digests of the *Arabidopsis thaliana* genome (ADB 2009-02-02) (Huala *et al.* 2001) and the ratio of its C-value to an estimated C-value for *B. vivipara* to select an effective enzyme. *Polygonum aviculare* (Polygonaceae) has a reported C-value of .855, approximately 5 times that of *A. thaliana* (Marie and Brown 1993). This served as my C-value estimate for *B. vivipara*. The *Arabidopsis* genome gives 4 934 fragments in the 400-500 bp range when cut at *PsiI*'s 6-base recognition site. I predicted 5 times that many (24 670) in my target species. Each fragment in the library had 202 bases called; two samples required 9 966 680 bases to be called to sequence this possible library once. Given that a single GAII lane may return as many as 3.5 billion base calls, I expected ample (~350X) oversequencing. Theoretically, choosing a 5-base cutter would have traded excess coverage for more unique loci because we expect more instances of 400-500 bp fragments between 5-base recognition sites than 6-base recognition sites in a randomly generated genome. However, given the margin for error in my C-value estimate and the number of bases to be called, the 6-base cutter was a more conservative choice.

Sequencing the RAD library

Fragments from the restriction digest were prepared for high-throughput sequencing to enable a comparison of restriction-site associated DNA (RAD) from the Arctic and Alpine individuals. The purified products of the restriction digests were processed following Illumina PE Library protocols (Illumina, Inc. 2008) for polishing fragment ends, adding adenosine overhangs, and ligating adapters.

The sequencing platform requires proprietary adapters (Solexa©) to flank the DNA of interest and bind the fragments to the GAII flowcell. It is possible to append a unique barcode, or molecular identification tag (MID), to the Illumina adapters that allows DNA from multiple samples to be sequenced simultaneously and separated later bioinformatically. I used custom oligos as adapters that were Solexa© sequences lengthened to include a 4-bp MID. The sequences were: 5' ACACTCTTTCCCTACACGACGCTCTTCCGATC-xxxx and 5' -p-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-xxxx, where x=[MID]. The MIDs were: Arctic (ACGT) and Alpine (AGCT). Adapter-ligated fragments from both samples were pooled together and size selected in the 400-500 bp range using a QIAquick Gel Extraction Kit (Qiagen). These size-selected fragments were enriched with limited opportunity for bias using the cycling conditions described in the Illumina protocol. The enrichment was validated via nanodrop, to confirm its purity; gel electrophoresis, to confirm the fragment size; and cloning and Sanger sequencing, to confirm the construct of the fragments; then sequenced in a single lane of a GAII flowcell.

Target fragment architecture

Each fragment in the RAD libraries was designed to have the following composition:

5'- (PE1)(MID)(RS)(plant dna)(RS)(MID)(PE2), where RS is the remnant of the restriction enzyme recognition site, and PE1 and 2 are the Solexa© adapter sequences. The adapters, MIDs, and RSs comprised 78 bp, leaving ~320-420 bp of Bistort gDNA at the center of each 400-500 bp fragment. The GAII called bases starting with the MID/RS on through 94 bp of plant DNA for each of the PE reads. In a 400-500 bp fragment with two 101-base PE reads, the last base called in PE1 is ~150 bp distant from the last base called in PE2 (Fig. 5).

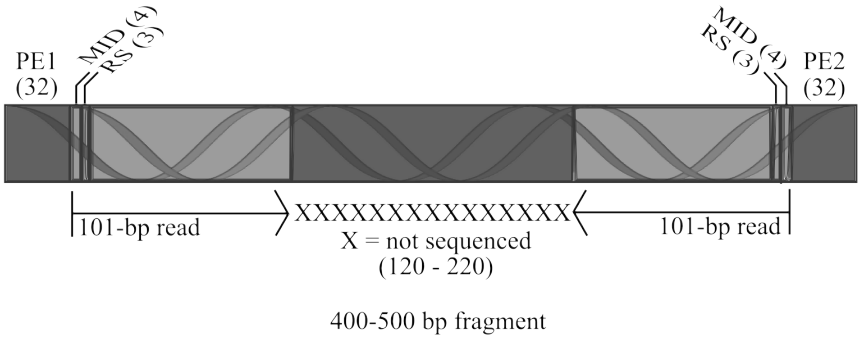

Figure 5. Targeted fragment architecture. PEs are Illumina paired-end adapter sequences which bound the fragments to the flowcell during bridge amplification and served as primers for sequencing-by-synthesis; MIDs are molecular identification tags that enabled bioinformatic separation of Arctic and Alpine samples; RSs are remnants of the *PsiI* restriction site (5'-TAA); parenthetic numbers are the segment lengths in bp; the arrows below the fragment indicate the direction, start, and stop locations of the two 101-bp reads; and Xs are portions of the fragment that were not sequenced. The DNA used in comparative analyses begins 7 bp into each read, after the MID and RS. For each 400-500 bp fragment, the last bp sequenced in one read was 120-220 bp away from the last bp sequenced in its opposite. The center of most fragments remained unsequenced.

The sequencing reaction was not intended to call bases for the entire length of the fragments. Rather, the output was a pair of reads called from the ends of the fragment towards the center. They describe nearly the same location in the genome, offset by a few hundred bp. The space by which they are offset may be large enough to allow for recombination in plants. If it is, the two sets, PE1 and 2, can be argued to double the number

of independent records, or samples of data. If it is not, and they are linked, the second set of reads allows independent verification of the analyses performed at each locus. In this investigation, the two sets remained separate in the analysis pipeline and the results were compared.

Flowcell results

Approximately 6.7 million 101-bp reads were split into their respective samples, Arctic and Alpine, via exact matches at their start to a MID. There is no *a priori* bias for one population or another in terms of how the molecules will respond to DNA library processing. Considering the stochastic elements in that process, one expects the exact number of Arctic and Alpine reads to differ, yet hopes they are roughly the same. There were ~3.2 million Alpine reads (48%) and ~3.5 million Arctic reads (52%).

The reads contained a total of ~1.3 billion base calls—37% of what was advertised as possible. It is uncertain why there were relatively few reads, although it is likely related to the internal quality filter of the GAII. The lower-than-expected output of called bases highlights the importance of choosing a restriction enzyme conservatively.

Quality control pipeline

Reads of poor overall quality were removed from the data prior to analysis. Initial quality control (q.c.) of the raw data was carried out with the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), a suite of software written for processing short reads associated with next-generation sequencing. Each base called by the GAII is assigned a quality score similar to the widely accepted PHRED scores used in Sanger sequencing.

PHRED scores are linked to the probability of calling a base incorrectly (Ewing *et al.* 1998). The Illumina format is simply a PHRED score offset by an integer, incorporated this way to circumvent ASCII-encoding problems in the development of another widely accepted data format, FASTQ (Cock *et al*. 2010). The Illumina quality scores can be reasonably interpreted using this rule of thumb: a score of 30 has a 1/1000 probability of error; a score of 20 has a 1/100 probability of error. To prevent carrying miscalled bases into the final analyses, I used a quality score threshold of 30 at several stages in my q.c. pipeline.

Bases with a score less than 30 were trimmed from the ends of all reads to address declining scores near the limit of the read length. Adapter sequences, which were likely in the library due to excessive ligation, were clipped out of the reads. Reads containing one or more unknown base (N) were discarded. They were further filtered by quality score: unless 90% of a read's bases scored 30 or higher, it was discarded. Then, to facilitate assembly (see below), reads shorter than 38 bases were discarded.

Quality control removed an average of 2.25 million reads per set. The remaining reads were converted to FASTA format for further analyses.

Assembling reads into RAD markers

A set of algorithms called Velvet, v. 1.0.12 (Zerbino and Birney 2008), was used to collapse the millions of q.c.'d PE1 and 2 reads into two lists of RAD markers with corresponding coverage information (Fig. 6A). Velvet is a sophisticated program written to assemble long contiguous sequences (contigs) from many overlapping short reads; I co-opted the program to process my RAD markers. A simplified explanation of its function follows. The program searches for words of length *k* (*k*-mers) in aligned reads and creates a 'node'

wherever the *k*-mer is represented in the reads a number of times specified by the user. It then attempts to extend nodes in either direction if overlapping *k*-mers (also with required coverage) are present. Nodes that reach 2*k*-1 in length are reported as contigs. (Note: I use 'contig' to describe the output of Velvet; 'locus' the physical location in the genome represented by a contig). Each contig from this assembly process represented either a unique locus, or a version thereof. Since both Arctic and Alpine contigs are needed for comparative analyses, Arctic reads were assembled separately from Alpine reads. Each population had two sets of reads, PE1 and 2, which made for a total of 4 assemblies.

To initiate this process, I removed the 7 leading bases from each read and input them in Velvet with a *k*-mer value of 31 and a coverage cutoff of 2. The first 7 bases correspond to the MID, which is artificial, and the RS. Although the RS is native, the RAD methods are based on the absence of variation in the RS. The sites confound the assembly algorithm by misrepresenting reads as contiguous (due to overlap), which they are not.

The *k*-mer value ($k = 31$) was chosen to maximize the program's sensitivity to small differences in reads in the library. Longer *k*-mers work to this end because the complete *k*-mer must be intact in the reads multiple times to count as coverage rather than a unique contig. The program detects allelic variation in the same individual as single nucleotide polymorphisms (SNPs) in the 31-base word and initiates a new node precisely because of that variation. Longer words are more sensitive to single differences, so versions of the same locus with SNPs should generate separate contigs.

This sensitivity came at the expense of excluding data from reads shorter than *k*. These reads were discarded for the simple reason that no *k*-mers could be detected therein, and amounted to ~1.3 million reads (almost 20% of the total reads) per set in this

investigation. The trade-off was viewed favorably in light of targeted analyses that relied on naturally occurring polymorphic sites to infer evolutionary histories. $K = 31$ is the maximum value permitted in the 64-bit computing environment available to me, and larger values are computationally exhaustive. Even without this limitation, a *k*-mer length any higher than 49 would not be suitable for this data, because the minimum length required to output a contig would be longer than the raw input reads.

The coverage cutoff of 2 was chosen to differentiate between sequencing errors and alleles in the DNA library without discarding potentially informative data. SNPs that are carried through the molecular processing from genomes to flowcell should appear multiple times in the reads. Sequencing errors, on the other hand, are not likely to occur at the same position, and will not be represented frequently. Thus, an effective coverage cutoff abandons reads whose unique sequence is likely the result of a miscalled base, yet is not set so high as to miss alleles that were present in the DNA library (even in low numbers). Since Velvet interprets coverage as the number of times a *k*-mer appears in the reads, a coverage cutoff of 2 (when $k = 31$) is equivalent to nucleotide coverage of 2.9. The coverage at each locus was essentially doubled by the existence of a highly similar locus in the other sample. Thus, for inclusion in my comparative analyses, a locus had at least ~6X nucleotide coverage over a string of 61 or more exactly matching nucleotides. The average nucleotide coverage for sequences used in comparative analyses was ~10X.

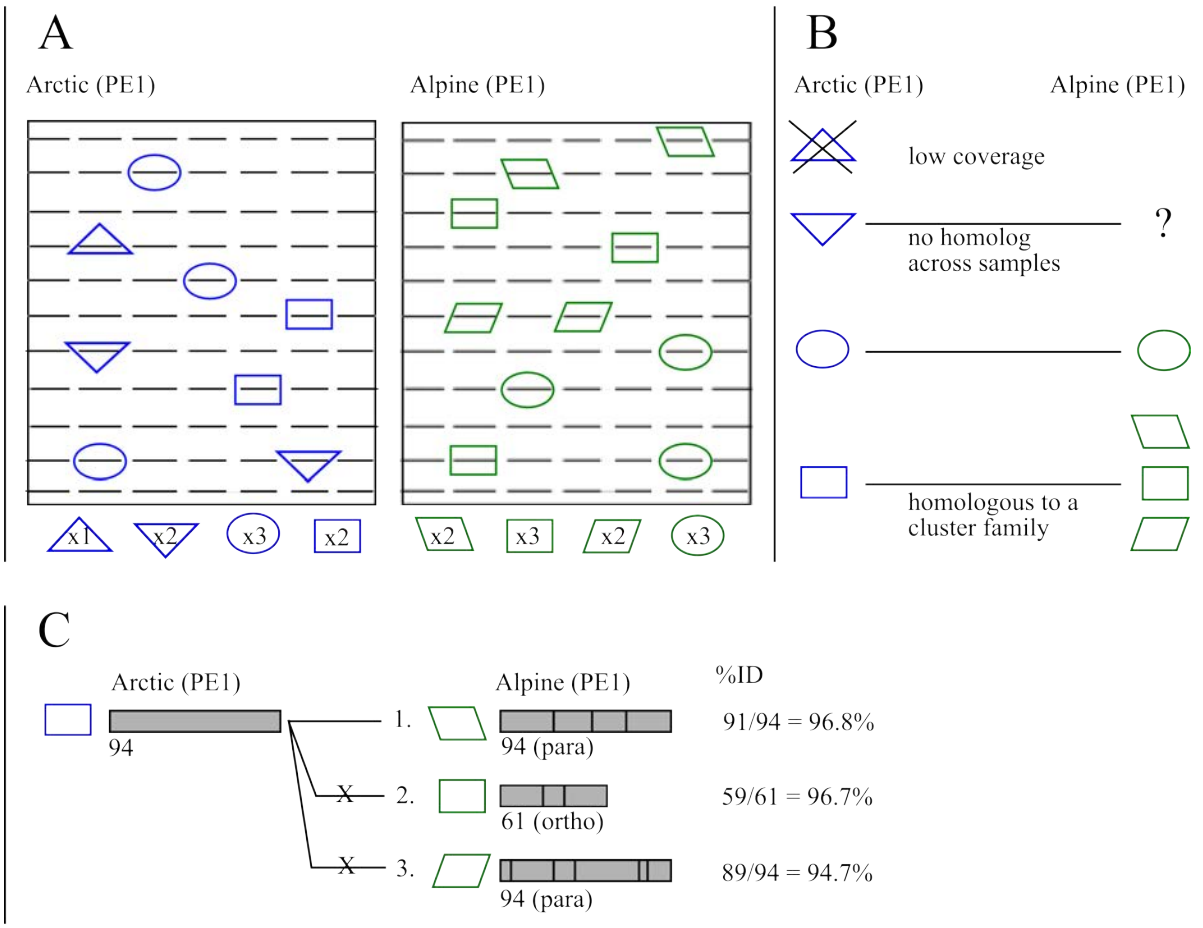

Figure 6. Workflow for assembling Illumina reads into RAD markers, pairing homologs across samples, and selecting sequences to compare. A) Millions of q.c.'d Illumina reads (dashes) were collapsed into thousands of RAD markers (shapes) with corresponding coverage information. B) Homologous Arctic and Alpine markers were paired together via a megablast across samples. Zero, one, or several sequences generated hits. Groups of similar sequences from a single sample (cluster families) have presumed origins in gene or genome duplications. C) The longest (or only) Arctic sequence in a cluster family is paired with the Alpine homolog with the highest percent identity (%ID) for final comparative analysis. Grey bars are sequences from the Arctic and Alpine individuals; vertical lines are polymorphic sites with respect to the Arctic sequence. %ID is calculated as identical columns / sequence length. Note the effect of incomplete data on pairing orthologs: 2 is orthologous and 1 and 3 are paralogous, but 2, truncated during q.c., is erroneously passed over for 1.

Assembly results

The contig lengths (in bases) confirmed effective use of the program, given the data. The average in each assembly, 94, was the longest read the molecular methods were intended to create (101 minus the MID and RS leaves 94). This was optimal. The minimum contig length for each assembly, 61, was expected from Velvet when using my chosen parameters: a read must have 2X *k*-mer coverage for overlapping *k*-mers of 31 to generate a node, limiting the minimum node length to $2k - 1$, or 61.

The maximum node lengths in each assembly were 1228, 645, 561, and 461. Given reads from a RAD library, one does not expect the program to either: 1) bridge the unknown distances between RAD markers; or, 2) extend nodes beyond the read length of the sequencer. While there is no suggestion that the former occurred, the latter certainly did. The algorithm presumably incorporated information from overlapping fragments whose origins involve imperfect cleavage, breakage, or ligation events. This is not unrealistic, especially when the bulk genomic processing permitted imperfect fragments. Fragments with unintended architecture could meet q.c. and assembly requirements, overlap with other reads, and contribute to extending the contigs. The different max lengths in each assembly supports the idea that long contigs are the result of stochastic elements of the DNA library preparation.

The number of contigs assembled is this investigation's best approximation of how many 400-500 bp fragments are actually generated by *PsiI* in *B. vivipara*. The assemblies generated an average of 44 741 contigs, with a standard deviation of < 5%. The small deviation might suggest that the number of contigs reasonably approximates the number of fragments—the exact number of which could only be generated *in silica* if the entire Bistort

genome was known. There are other possibilities for the small deviation. The number of contigs may be converging as a result of the restriction digest: if the enzymes found the same number of same-spaced restriction sites in the same amount of time, but the sites were not the same sites in each sample, the total number of fragments would remain unknown (though the number of fragments generated would be the same). The DNA was saturated with enzymes for an extended period of time to reduce the likelihood of this scenario.

Taken as an approximation of the actual number of fragments generated by *PsiI* in a completely digested genome, the contig count nearly doubles the number of fragments predicted *a priori*. This further highlights the importance of a conservative choice for a restriction enzyme when developing RAD markers for a non-model organism.

Reconstructing homologous collinearity

After the sequences were in hand, my goal was to select those whose comparison would be evolutionarily informative.

Defining homology

Homologous sequences may arise from either population/species splitting events or gene duplication events, and these circumstances must be carefully considered before analysis. Homologs are related to each other by descent from a common ancestral DNA sequence. They may be orthologous—arising from a speciation or splitting event, or paralogous—arising from a gene or genome duplication event. Testing my hypotheses relied on a comparison of orthologs—sequences that were once the same and have been accumulating changes since the splitting of the population. They contain the information

needed to recover estimates of *t* and rates of change in protein-coding sequences. Comparing paralogs for these estimates is problematic. For example, dating divergence using paralogs could grossly misrepresent *t* if variation in the sites has been accumulating since an ancient gene duplication. Unfortunately, the data do not allow a sure way to differentiate paralogs and orthologs, and the paralogs, logically, exist within the data because *B. vivipara* is polyploid. Therefore, it was necessary to adopt several assumptions and precautions in evaluating homologs as orthologs for comparative analysis.

Pairing homologs

I used megablast, the BLAST search task tuned for finding matches among closely related sequences (Altschul *et al*. 1990), several times in this investigation as a tool to inspect and curate the data. This similarity search algorithm aligns sequences, determines their percent identity (%ID), and assigns each pair a similarity score. The score correlates to %ID, which the program's default settings define as the number of columns in an alignment with an exact match divided by the total number of columns, including gaps. Unless otherwise noted, all megablast searches were executed using default parameters except for an e-value threshold of 1e-6, which corresponds, roughly, to a million to one chance that the similarity between pairs is a matter of random chance.

To identify homologs, I used a megablast search across samples (Fig. 6B). The output from the search was parsed using a Perl program (Appendix 2). The script recorded each instance of a query (an Arctic contig) that had a match (an Alpine contig), while an additional filter excluded matches if the alignment that identified them as similar was less than 60 columns. This eliminated instances of contigs being classified as homologs when only short

motif-like portions of DNA embedded in the contig generated a megablast match. The threshold of 60 was chosen based on the minimum contig length from the assemblies. The results were the first confirmation that the RAD-based methods generated homologous sequences from multiple samples. Approximately 25% of the contigs from one sample had a homolog in the other as detected using this method.

Cluster families

Alleles from the same individual were clustered together via a megablast of one list of contigs to itself. Most (~90%) of the contigs had no significant similarity to others from that sample. These are interpreted as different RAD markers from across the genome—the welcome outcome of the RAD methodology. The remaining ~10% were highly similar versions of contigs from the same sample. These alternates were there because the assembly algorithm explicitly included allelic differences in the output. So, the contigs from one sample, while mostly unique sequences, were predisposed to grouping into clusters, or sequences that share a certain %ID. The %ID threshold for belonging to a cluster family was set based on information from the megablast of each population's data to itself: the mean, minimum, and maximum %ID of sequences recognized as homologs within a sample was 94, 77, and 99. I used the minimum value, 77, as the %ID threshold for cluster families to be sure all homologs from a single sample were grouped together.

Cluster family members have an implied evolutionary relationship: they are different versions of the same locus from different gene copies in one individual. They are probably paralogs that arose from gene or genome duplication events in the history of the species, though I cannot be certain. They might also be explained as discontiguous loci showing

variability in a repetitive genome-wide motif that includes *PsiI*'s recognition site. Although the data did not permit the distinction, I assumed that cluster families were groups of paralogs with origins in duplication events.

Since ploidy events are an opportunity for gene copies to arise and acquire variation (Doyle *et al*. 2008), and *B. vivipara* is a known polyploid, I expected instances of clusters when assembling the data. So why are the majority of contigs the sole member of a cluster? Several possible explanations exist, including: 1) variants existed in the individual, but were not in the flowcell due to stochastic DNA library preparation; 2) variants did not exist in the individual, (i.e. all gene copies were homozygous); and, 3) the q.c. and assembly requirements discarded evidence that alternate gene versions were sequenced. Again, the data did not easily confirm or refute these possibilities.

As a proxy measure of completeness, the frequency of various-sized cluster families suggests my survey includes DNA from both duplicated genomes and gene families, but only covers a portion of the organism's genetic makeup (Fig. 7). There are relatively few instances where alleles from all 10 genome copies could comprise a complete cluster family.

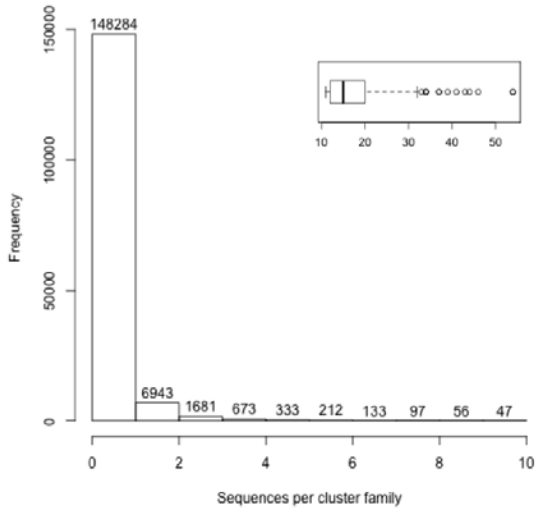

Figure 7. A census of sequences per cluster family. A histogram counting cluster families with up to 10 members and a boxplot (insert) showing the distribution of cluster families with more than 10 sequences serves to gauge the completeness of my genomic survey. Despite polyploidy, the majority of contigs are the sole member of a cluster.

Plant genomes harbor large gene families (Wall *et al*. 2008), so cluster families might be a mixture of alleles from duplicated genomes and paralogs from gene families. This must be the case with cluster families larger than 10, which existed in the data, but were the vast minority. The consequences of partial genome sampling are considered in the interpretation of results (see Discussion).

Pairing orthologs

The cluster families confounded a simple comparison of Arctic and Alpine markers. There are several permutations of homologs in a megablast across samples: each query and its match may or may not belong to a cluster. To make a comparison at a locus, then, was to arbitrarily choose which sequences from each sample to compare (Fig. 6C).

I faced two primary obstacles to confidently pairing orthologs: incomplete data and the ambiguity of paralogs and orthologs. The final decision to pair orthologs was influenced by both. I paired the longest sequence from the larger database (Arctic), to take advantage of the available data, with its highest scoring hit in the Alpine sample, to increase the chance that the comparison truly is orthologous, based the assumption described below. If a sequence was homologous to one without cluster family members, that pair was taken as the ortholog. Pairs made according to these criteria were carried into comparative analysis, and are herein referred to as comparative genomic (CG) pairs.

Due to stochastic molecular processing and/or q.c. methods, the true ortholog may not have been represented in the opposite sample. If it was, and it was shorter than other cluster members, it may have been passed over as the best hit because of a higher similarity score to a paralog with a longer length (see example in Fig. 6C). Because all the alleles of a given

locus were not necessarily present in the data, and the pairing process needed an anchor, I started with the longest cluster family member from the sample with the most data.

The critical assumption made by taking the highest scoring Alpine hit as its ortholog is that an ortholog is always more similar than a paralog. If all gene and genome duplication events pre-date the population split and the molecular clock is fairly accurate over the time scales of the Quaternary, then paralogs are older than orthologs, have more time to accumulate mutations, and are less similar. I predicted a shallow population divergence date (the last glacial maximum) relative to the origin of the species, so the idea that ploidy and gene duplication events pre-date the split seemed plausible. On the other hand, ploidy events can happen virtually overnight (Pikaard 2001), and mutation rates of different gene copies and different loci may vary. Genetic simulations to test my assumption that paralogs are more divergent than orthologs might be possible, but have not yet been done and are outside the scope of this investigation. Other investigations have faced the "perils of paralogy" (Martin and Burg 2002), and more sophisticated approaches to sorting orthologs from paralogs are desirable.

Current approaches that infer protein orthologs from massive post-genomic era alignments were only marginally helpful for choosing *Bistorta vivipara* sequences to compare. Current strategies for parsing orthologs from paralogs vary depending on their basis of inference: graph-based methods rely on BLAST searches and tree-based methods rely on phylogenetic analysis. The former, like those in OrthoMCL (Li *et al*. 2003) and Multi-Paranoid (Alexeyenko *et al*. 2006), provide orthologs from hundreds of complete genomes given protein data. The latter, like those in Orthostrapper (Storm and Sonnhammer 2002), SDI tree reconciliation (Zmasek and Eddy 2002), and LOFT (van der Heijden *et al.*

2007) take user-provided multi-species genome data and build phylogenetic trees with the target species. If either group of tools could show that Arctic and Alpine homologs share a higher %ID or closer relation with orthologs from another species than they do with each other, I could infer the pair did not diverge at the population split, but rather earlier, and should not be compared as orthologs. There were several roadblocks to using these tools to this end. The first was the relative scarcity of completely sequenced plant genomes. Of the two BLAST-based programs above, the first references just 3 genomes from Plantae and the second, none. Even with completed genomes, the tree-based programs are not built for automated comparisons with numerous partially sequenced genes. Rather, they sort and classify homologs from long alignments with multiple species. Some programs (e.g. Ensembl Compara [Hubbard *et al*. 2007], TreeFam [Li *et al*. 2006], and OrthologID [Chiu *et al*. 2006]) combine graph- and tree- based approaches, but again, the hitch is automation and relevant source data. Only OrthologID draws on a database of completed Plant genomes and places a query sequence in a phylogeny, but there is no support for including both halves of a homologous pair to assess their relative positions. I manually entered 10 putative protein-coding pairs into OrthologID, and only 3 produced orthologs from their database of 4 complete (and one partial) plant genomes. Each pair was incorporated into two trees to visually confirm that the Arctic/Alpine homologs held the same position in the phylogeny. The confirmation added support that the select pairs were more likely orthologs than paralogs, but it would be impractical to mine and parse ortholog databases in this way without automation. A better approach would be to link OrthologID's strategy of clustering and automatic tree building to a comprehensive plant gene family database, like GreenPhylDB (Rouard *et al*. 2010). However, more advanced programming is needed to

generate and interpret trees with multiple samples' queries and their interspecific orthologs automatically. Finally, none of these contributions to the complex task of classifying homologs are applicable to non-coding DNA, which comprised the bulk of the sequences I sampled. So, the algorithm that paired the longest Arctic sequence with the most similar Alpine sequence was used to make CG pairs.

Alignment of multiple sequences

As the orthologs were paired, they were aligned to properly assess the form, number, and location of nucleotide changes as observed in the two samples. I incorporated MUSCLE v3.8.31 (Edgar 2004), an alignment program based on the now-classic algorithms developed by Needleman and Wunsch (1970), Smith and Waterman (1981), and Altshcul *et al.* (1997), to power the alignments in a script that (if needed) reverse complimented one or the other sequence in a CG pair, aligned them, and placed them in a small .fasta file (Appendix 3).

Identification of evolutionarily constrained sequences

Comparative genomic pairs were destined for one of four analyses depending on their putative content: chloroplast (CP), mitochondria (M), nuclear protein-coding (P), or nuclear non-coding (NC). Sequence content was determined by querying annotated databases with each pair in BLAST or BLAST-like searches. Pairs whose content was ambiguous were excluded from downstream analyses.

Identifying plastid sequences

All contigs were searched (megablast) against a collection of complete plastid

genomes downloaded from NCBI (Appendix 1). Ten chloroplast genomes and 12 mitochondrial genomes from diverse taxa across Plantae comprised the custom database. For each contig with a match, the most similar sequence in the database was either CP or M. If the two most similar database sequences were either both CP, or both M, the contig was assigned to the appropriate category. If the two highest scoring matches did not correspond, the contig was flagged as an ambiguous plastid sequence.

The relatively small plastid database allowed me to use the complete list of contigs, not just those selected as orthologs, to serve as the query. This way, every contig had a chance to be recognized as a plastid sequence and carry its cluster family members into a category, regardless if it was used as one half of a CG pair. The buckwheat chloroplast genome was the best hit for many contigs, which was expected, given that it was the closest relative to *B. vivipara* in the custom plastid database.

The combined database was important for unambiguous categorization. Plant plastid genomes may be exchanging genetic materials with a relatively high frequency (Goremykin *et al*. 2008). This became evident from the data when the contigs were blasted to separate CP or M databases: only a few sequences found a match in one, but not the other plastid database. It was these few unambiguous plastid markers I carried into downstream analyses. They were double-checked by blasting them against NCBI's complete nucleotide database (nt). The best hit for each was a known CP or M sequence, and several highest scoring matches were to *B. vivipara* sequences already in the nt database.

Identifying protein-coding sequences

Gene-finding algorithms based on statistical techniques, empirical evidence, or a

combination of both, have been thoroughly developed for genomic data. Some, like GLIMMER (Salzberg *et al*. 1998) and GeneMark (Lomsadze *et al*. 2005), use interpolated or hidden Markov models (I/HMMs) to analyze codon content and recognize protein-coding patterns. Others, like NCBI's ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) pick out the potential genes in a large gDNA sequence by automatically parsing BLAST results. Popular web portals (e.g. Oak Ridge National Laboratory [ORNL]) integrate multiple established programs in a pipeline to benefit from both approaches. Pattern-searching tools are most effectively employed on microbial genomes due to their low fraction of intron-containing genes. IMMs trained on eukaryotic genomes (human, mouse, and yeast) exist, and are incorporated into ORNL's pipeline. Like most genome annotation pipelines, however, it is intended for use on small numbers of vast gDNA sequences. The GrailEXP (Hyatt *et al*. 2000) component can run pattern-finding analyses in large batches (required for numerous markers), but the required binary files are no longer distributed at ORNL. Both GrailEXP and GeneMark provide the option to train IMMs on user-supplied databases prior to detecting ORFs and could potentially find protein-coding sequences in the *Bistorta* data. Still, these techniques were originally intended to aid the annotation of new, nearly complete genomes. Given the short, unlinked sequences of RAD markers, the complex developments in eukaryotic pattern-finding algorithms would likely go unused. This, and the need to process hundreds of small sequences iteratively, prompted the use of a gene- (or partial gene) finding strategy based solely on BLAST-like searches.

I used USEARCH's blastx-like similarity search task (Edgar 2010) and Perl scripts to search the largest publicly available protein database—NCBI's "non-redundant" (nr)—with a representative list of contigs to identify protein-coding sequences. Each open reading frame

(ORF) for each sequence was translated into protein data using the standard genetic code. ORFs were allowed to start at the beginning of a sequence even if it was not a start codon, end at the end of a sequence without a stop codon, and start immediately following a stop codon, all of which was appropriate for sequences that may only partially cover a gene. Unlike a default BLAST search, which reports all similar sequences in the database, USEARCH moved on to the next query once a match was found. This reduced run time without compromising purpose: a query was flagged as a putative protein-coding sequence at the first instance of similarity to a sequence in nr.

The massive nr database required that I use a reduced number of Bistort queries. To do otherwise would have been a computationally exhaustive task with a long run time. To make the search tractable, I combined the Arctic and Alpine contigs and sorted them into clusters of 77 %ID using USEARCH's default definition of similarity, which uses the length of the shorter sequence as the denominator, excluding gaps. This had the effect of lowering the BLAST definition of identity from 77% and ensured that all homologs clustered together. I used the longest member of a cluster to represent the homologs. If it found a match in nr, then its cluster family members were assumed to be protein-coding.

To execute the search, I downloaded the nr database in binary via ftp, converted the database to FASTA format using the BLAST task fastacmd, replaced the protein identities with integers, split up the database into 333 smaller databases, and searched each portion serially with the representative sequences described above using a Perl script (Appendix 4). Replacing the protein names with integers reduced the amount of memory required per search (nr protein names are long and information rich), and the identities were not needed to flag the contigs as protein-coding. Each portion of the nr database was searched using the

default USEARCH parameters, except: 1) the maximum sequence length was 35 000 (longer than the longest in nr, so that all were included); 2) the maximum number of rejects was zero, so that a search whose query failed to find a match after several attempts was not terminated until all sequences in that portion of the database were examined; and, 3) the maximum number of accepts was 3, to increase the likelihood that the reported match is the best match in the database, which would eventually help me get the correct frame information for each P sequence. As with megablast searches, the expect-value threshold was set to 1e-6.

Each representative sequence with a match initiated the final process of categorization via a custom program (Appendix 5). First, the match and its cluster family members were screened to see if the locus was already flagged as a plastid in the plastid database search. This was liable to happen, given that nr contains CP and M proteins and contigs flagged as plastid were not removed from the nr search query. If a contig had not been previously categorized, the representative sequence and all its cluster members were committed to the P category.

Having populated master lists for each category with the names of all contigs that could be argued to be CP, M, or P sequences, the script screened each CG pair by these lists. If either name was on the CP, M, or P list, it was designated accordingly. If either name appeared on more than one master list, its association with a category was ambiguous, and it was designated an 'error' pair. Comparative genomic pairs not associated with CP, M, P, or error categories were classified as NC.

Analysis of non-coding pairs

IMa2

I used IMa2 (Hey 2010), a program that can recover historic demographic parameters from multi-locus data sets, to estimate two parameters: 1) time since splitting, *t*; and, 2) ancestral population size (N_a); using thousands of NC CG pairs. It runs Markov chain Monte Carlo (MCMC) simulations based on a population model that includes up to 6 different demographic parameters (Fig. 8). Parameters not relevant to data with a population sample size of 1 were either excluded from the model or provided meaningless computations that were ignored.

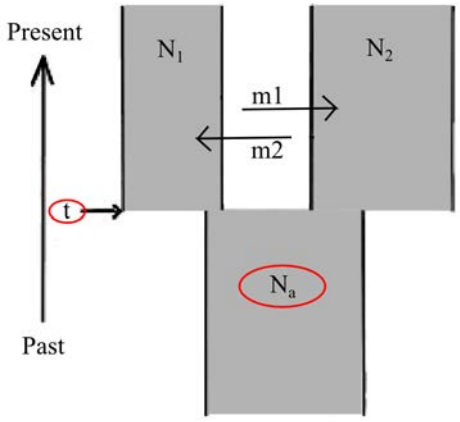

Figure 8. The IMa2 model depicting 6 possible demographic parameters (Hey and Nielsen 2004). N_1, N_2, and N_a are constant effective population sizes, m1 and m2 are gene flow rates, and t is the time of population splitting. Parameters evaluated in this investigation are circled; others were either removed from the model or provided meaningless values.

I ran more than 50 independent MCMC simulations of 200 NC loci each for PE1 and 2 data sets. Two hundred loci is the program's default maximum. It can be recompiled to handle more, but this leads to problems: analyzing more loci requires weeks or months of run time, huge memory demands, and program instability. It is desirable to analyze thousands of loci in the same run, but is apparently beyond the current theoretical and technical means (at least with a sample size of 1, see Discussion). I used the best available computing cluster, Odyssey, supported by the FAS Science Division Research Computing Group at Harvard University (HU), which was running IMa2 at a maximum of 200 loci and failed at attempts of analyzing more.

The activity of the MCMC simulations is based on coalescent theory. Due to the stochastic nature of genealogical processes, we cannot recreate the true genealogical history

(gene tree) that gave rise to the contemporary homologs; but we can use the probability that two sequences coalesce (merge together, in the direction of the past) in the previous generation to our advantage. The equation of that probability, Kingman's coalescent (1982), incorporates various demographic parameters. In IMa2, Choi, Hey, and Nielsen implemented the coalescent so that it does not include population size and migration rates, but rather the splitting time parameter and mutation parameters. This fit my data—those parameters could be estimated with a sample size of one individual per population given numerous loci.

At each locus, a genealogy was generated with random values for the various parameters and then updated with new values. The new values were either accepted and saved, for comparison to the next, or rejected and the old values were retained, depending on which values resulted in the more likely genealogy, given the data, according to Metropolis-Hastings criteria (Hey and Nielsen 2007). Each round of updating and evaluating the simulation state counted as one step, and I ran the simulations for 2.1 million steps. Every 100 steps, the current genealogies and parameter values were saved, giving me 21 000 genealogies from which I recovered the posterior likelihood of the targeted parameter values. The 100 steps between saving states gave the chains sufficient time to explore the full state space before saving, making each saved state effectively independent. Exploring the state space properly via mixed and heated chains was vitally important and these elements are further described below. To estimate the posterior probability of t and N_a, their values from zero to a user-defined maximum (the 'prior') were split into 1000 bins and a histogram was generated plotting the frequency of each value among the saved genealogies. The peak of this curve, converted to the appropriate demographic units, provided my final demographic

parameter estimates and the 95% highest posterior density (HPD) intervals thereof.

The target demographic units for *t* and N_a were time of population splitting at *t* generations in the past (years), and effective ancestral population size (individuals), respectively. Generally, values fitted in the IMa2 model are scaled by the neutral mutation rate, u, and/or the inheritance scalar (i.e. 1 for autosomal diploid loci), so conversion from parameter values to target units was required. In the model, time is estimated as *t*(u), so these values were divided by the geometric mean of the mutation rates included in the input file to get *t* in years. N_a is estimated as N_a(u). To convert to individuals, the formula $(4N_au)/(20uG)$, where G = years per generation, was employed, because the parameter value is proportional to the inverse of the coalescent rate per generation (Hey 2010b). The inheritance scalar of 5 accounted for decaploid individuals.

The posterior probability curves from each run within the paired-end data sets were combined to generate my final estimates of *t* and N_a. Each of the 1000 bins, from zero to the prior, were nearly the same in each run. They differ because the scalar variable for each parameter in the model is not the constants themselves, but the constants scaled by the mutation rate priors (in mutations/locus/year). Most of the sequences are 94 bp, but there is some variation in length. Variation in length led to a variation in mutation rate priors, and thus variation in the flat values at each bin. I averaged the x-axis units (years or individuals) from all the runs to generate the bin values for the final posterior probability curves, and summed the probability for each of the bins from all the runs of PE1 or 2 to generate the final y-values. The result was one curve each for *t* and N_a for both PE sets that represent the combined posterior likelihood results from the numerous simulations of 200 loci each.

I followed several precautions to assess the correct use of the program and the

convergence and mixing of the Markov chains. Some checks were automated and applied to every run. Others required repeating runs with different command line inputs and were too long to repeat with all 100+ runs. I ran these tests on 3 random sets of 200 loci and the results were assumed to be applicable to all runs. In this manner, I determined how long the program needed to run to get reliable results, what priors would be useful, and how changing priors and random number seeds affected repeatability.

I used Perl-driven automated checks to parse the IMa2 output files for update rates, swap rates, autocorrelation, and effective sample size (ESS) values for all runs. Update rates, or the percent of new values accepted versus rejected, were well above zero and swap rates between chains were high; these checks indicate the chains are doing a fair job of exploring the state space and evaluating the full range of parameters. I confirmed near-zero autocorrelation of output parameters and high ESS values for all runs. Non-zero autocorrelation values indicate long term trends, meaning a stationary distribution has not been reached (Hey 2010b). ESS values should be high (theoretically, as high as the number of iterations completed), but in practice (due to the properties of the chains), can be lower. ESS values for all chains in all runs were greater than 30 (average 153 604).

As an alternative method of evaluating stationarity and convergence, I ran 3 representative sets three times, identically, except for the random number seed (which generates the random starting parameters) and observed that the output values converge to the same parameter estimates at the end of the run.

Burn-in

The initial phase of the MCMC simulation (the 'burn-in') allows it to run long enough

that the current values are independent of their starting point. It is unlikely that the randomly generated starting values are correct, so the burn-in gives the chains time to explore alternate genealogies and parameter values before those values are used to generate posterior probabilities.

I found 230 000 steps to be an effective burn-in for my runs. I used the automated checks described above to confirm stationarity for all runs at this point, and I observed trend plots of the values over time from 3 representative sets. Update rates, swap rates, and ESS values were non-trivial, and the trend plots, which chart the attempted parameter values at each step, did not indicate long term trends. After 230 000 steps, the chains were mixing well and the genealogies began to be saved.

Markov-chains

I used 40 Metropolis-coupled chains with the suggested heating figures from the IMa2 documentation for a medium-sized data set with a geometric heating model for all runs. Metropolis-coupled chains allow multiple chains to run simultaneously and are strongly suggested for data sets containing multiple loci (Hey 2010b).

I confirmed the chains were effectively mixing for the duration by running my automated and visual checks after 2.1 million steps. As after the burn-in, update rates, swap rates and ESS values were high, and the trendplots indicated stationarity. Autocorrelations were zero, or near-zero for all runs. Again, if the charts of the updated values over the course of the simulation showed a trend, the posterior probabilities were probably not from the true stationary posterior distribution, and the program was not run long enough.

Priors

I provided prior ranges for mutation rates, N_e, and *t* that were larger than *a priori* expectations, but not excessively higher than needed.

The mutation rate range prior covered three orders of magnitude and encompassed what I estimated from the literature to be a realistic rate for nuclear non-coding plant DNA. The geometric mean of the mutation rate was used to convert the parameter value *t* into years, thus it was critical that the true mutation rate fell within the prior range if estimates of *t* were to be reliable. Whole genome studies in humans (Roach *et al*. 2010), and extensive mutation studies in mice (summarized in Russell and Russell 1996, Drake *et al*. 1998) honed in on a rate of ~1.1 X 10^{-8} mutations per bp per generation; Bayesian multilocus analyses from the genus Cornaceae (Cornales) arrived at substitution rates from ~1.5 X 10^{-8} to 8 X 10^{-8} in nuclear genes (Xiang *et al*. 2008). The former studies are sophisticated investigations on mutation rates in organisms with completely sequenced genomes. The latter used MCMC based analyses to estimate posterior likelihoods of the mutation rate in a species monophyletic with rosids, asterids, and *Bistorta vivipara*. I multiplied an upper (1.1 X 10^{-7}) and lower (1.1 X 10^{-9}) bound, in mutations per bp per generation, by the length of each locus, and used the ratio of these limits as limits on the ratios of the mutation rate parameters in the IMa2 runs.

I ran the program several times on 3 test sets of loci with successively smaller N_e priors to find that as long as they were greater than the equivalent of 40 000 individuals, the complete posterior marginal distributions were visible. I set the mutation rate-scaled prior for the final runs to 1.75, the equivalent of ~80 000 individuals.

By a similar approach, I found runs with *t* value priors greater than 0.5 contained the

upper bounds of the posterior distribution and generally converged on the same posterior probabilities. The scalar 0.75—the equivalent of ~700 000 years—was used as the prior for the final runs.

I explicitly tested the influence/bias of the priors and found none. There was little variation in the posterior distribution of parameter values in runs with priors greater than 1.75 and 0.75 for N_e and *t* (Fig. 8B).

Substitution model

I used the Infinite Sites (IS) model (Kimura 1969) as the nucleotide substitution model for the simulations. Under this model, all mutations in the history of a sequence occur at a different site. The IS model is best used for relatively recent splitting—one does not expect multiple mutations per base pair along a lineage unless the branch is very many generations long—so, it suited the NC CG pairs.

IMa2 model assumptions

IMa2 fit my data to a model that made several important assumptions: 1) there was no migration between populations; 2) the loci were under selective neutrality; 3) there was no recombination within loci; and, 4) there was free recombination between loci. I argue that my data fit these assumptions and were suitable for use in this framework.

By excluding migration, I assumed that Arctic and Alpine populations were at some point true island populations, i.e., there was not another population exchanging genes with the sampled populations that was more closely related than the two. The IMa2 model does include a migration parameter that would estimate the degree of isolation between the

populations, but this requires a sample size > 1. Regardless of its tractability given the data, the effect of *Bistorta* migration between Arctic Alaska to the southernmost Rocky Mountain population is assumed to be nearly nil. Ornithophily is probably the Alpine Bistort's best means of long-distance dispersal. The bulbils share none of the characteristics of wind dispersed fruits; plus, they are consumed by and pass through ptarmigan digestive tracts intact and viable (Clarke and Johnson 2005). However, Alaskan ptarmigan (*Lagopus lagopus*) tend to winter in Alaska (Irving *et al*. 1967), and Coloradan ptarmigan (*Lagopus leucurus*) migrate less than 10 km seasonally (Hoffman and Braun 1975). The remaining effect of gene exchange over the thousands of kilometers separating the *Bistorta* populations (or even the intermittent Alpine populations north of Colorado) by long distance avian migrants that consume *Bistorta* fruits was assumed to be minimal.

The second assumption was that the variation within the data was not affected by directional or balancing selection. The NC CG pairs were categorized by their lack of similarity to any known proteins or non-coding conserved plastid sequences and were assumed to fit this characterization.

The salient implication of the third and fourth assumptions is that the length of the contigs was short enough and the genomic distance between them long enough to restrict free recombination within loci and permit it between them. The CG pairs were, on average, 94 bp long, with a min and max of 61 and a few hundred. Although recombination rates vary widely in the literature, I was comfortable assuming recombination does not occur over such small distances in plant genomes, and certainly not for the majority of NC markers. With regard to free recombination between loci, despite the similarly variable estimated rates of crossing over in plants, even loci with a low rate of crossing over per generation are

effectively unlinked over longer time frames. Rare instances of recombination or linkage between markers may be included in the sheer volume of data, but such cases are likely the minority and their effects were swamped out by unlinked loci. Assuming the distribution of *PsiI*'s recognition site is pangenomic in *B. vivipara*, I was comfortable treating all NC loci as having segregated independently over time, which, if true, gives the multi-locus analysis its power (Felsenstein 2005).

Analysis of plastid pairs

The chloroplast markers within each sample were linked together and aligned as two long loci. The same procedure was followed for mitochondrial markers, and the cytoplasmic genomes were inspected visually using MEGA version 5 (Tamura *et al*. 2011). The lack of variation in both genomes was apparent, and the alignments were processed by a variant of the code in Appendix 6 to generate statistics on the frequency of polymorphic sites.

Analysis of protein-coding pairs

My goal was to estimate the form and magnitude of natural selection on the P CG pairs by calculating dS and dN for each.

I used Perl code (Appendix 6) to segregate identical P pairs from those exhibiting the requisite polymorphism. Pairs were deemed identical if each column in the alignment matched, disregarding gaps. To determine the reading frame of each pair, also requisite for my calculations, I searched NCBI's nr exactly as described for identifying protein-coding sequences, except this time the queries were the already categorized pairs. Sequences that did not find a match in this search were excluded from further analyses. Matchless

sequences were probably cluster family members carried into the P category by their representative sequence and were not sufficiently similar to a database protein to be recovered in the reciprocal search.

I wrote another program (Appendix 7) to execute SNAP (Synonymous Nonsynonymous Analysis Program) (Korber 2000) iteratively on each polymorphic pair. The pairs were put into frame and loaded in SNAP, which provided dS and dN by: 1) counting the number of synonymous and nonsynonymous sites; 2) counting the number of substitutions at synonymous and nonsynonymous sites; 3) dividing the number of substitutions by the number of sites (usually called pS and pN); 4) applying the Jukes-Cantor correction (Jukes and Cantor 1969) for multiple substitutions to transform pS to dS, etc.; and 5) dividing dS by dN. The final ratio is unavailable from sequences displaying one, but not both types of substitutions. Pairs with $dS/dN > 1$ or $dS > dN$ were interpreted as examples of purifying selection; those with $dS/dN < 1$ or $dS < dN$ were interpreted as examples of positive selection. Ratios with greater distance from unity were interpreted as having a greater magnitude of selective pressure.

RESULTS

Summary

Most protein (P) pairs analyzed were governed by purifying selection since the splitting of the Arctic and Alpine populations, ~140 000 years ago. Very few of the P pairs gave a signal of positive selection, and none gave a signal of neutral selection (dS=dN). The size of the ancestral population was ~23 000 (Table 1).

Bioinformatics

Paired-end (PE) 1 and 2 results were compared throughout the investigation (Table 2). Both sets of data follow the same pattern, with the exception of the quality control (q.c.) segment: 11% more reads were discarded from PE2 than PE1. Compared to the PE1 data, the PE2 data lost 246% more reads after trimming low-quality bases from the ends, 80.5% more when reads < 5 bp were discarded after removing adapter sequences, 18% percent more failed to pass the stringent quality filter, and 5% more were lost when reads < 39 bp were discarded. In one instance, more PE1 than PE2 reads were discarded: 123% more PE1 reads contained one or more unknown base (N). When normalized by the number of reads remaining after q.c., PE1 and 2 results vary by less than 0.05% (except max contig length, see above).

Alignment statistics

The proportions of variable sites in the three partially sequenced genomes— mitochondrial, chloroplast, and nuclear— rank from lowest to highest, in that order (Table 3). Polymorphic sites in the mitochondrial and chloroplast genomes were an order of magnitude

less frequent than in the nuclear genome.

All columns containing a dash (-) were ignored when generating alignment statistics. Terminal gaps overwhelmed internal gaps by the thousands due to cases where the Arctic and Alpine homologs were different lengths. They were not informative when estimating the frequency of polymorphic sites in the alignments, difficult to separate bioinformatically from internal gaps, and thus, not counted.

Three CP markers were discarded for the alignment statistics. The match to mismatch ratio in the outliers (98:73, 218:16, and 78:14) accounted for 75% of the variability in 41 CP markers.

Demographic analyses

Time since splitting, *t*, is ~140 000 years ago, and the ancestral population size, N_a is ~23 000 individuals. Fifty-five coalescent simulations using 200 loci and 1 simulation using 152 loci tapped the PE1 NC pairs for results; similarly, 51 plus 1 were run on the PE2 NC pairs. Taken one at a time, the peak of the posterior probability curves for *t* and N_a was each simulation's best estimate of those parameters. The values above HPD95%Lo and below HPD95%Hi contained 95% of the area under each curves. These are typically interpreted as confidence intervals, i.e., the probability of the actual value falling within that range is 0.95. The curves were combined as described in Methods to generate final estimates (Fig. 9).

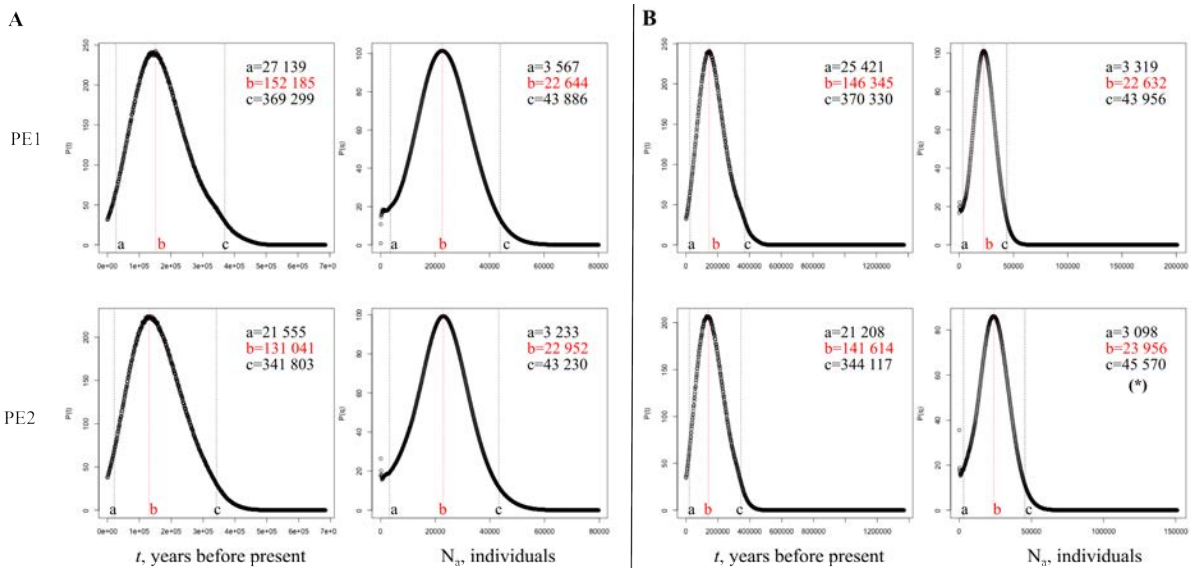

Figure 9. Combined probability curves for time since splitting, *t*, and ancestral effective population size, N_a, from 50+ Markov Chain Monte Carlo (MCMC)-based IMa2 runs each for paired-end (PE) data sets 1 and 2, generated with alternate priors. For the curves in A, priors were set to ~700 000 years and ~80 000 individuals; for B, priors were set to ~1.4 million years and 200 000 individuals, with the exception of (*), which used a prior of 150 000 individuals. For all runs, the mutation rate range priors were estimated from the literature and encompassed 3 orders of magnitude. Each simulation began with 200 pairs of non-coding, unlinked, neutral nuclear markers with an average size of 94 bp. After a burn-in of 230 000 steps, the simulations ran for 2.1 million iterations, saving the parameter values and genealogies every 100 steps, for a total of 21 000 saved parameter value sets per simulation. Summary histograms were made for each batch of 200 loci (not shown) in which the x-axis was the target parameter from 0 to the prior divided into 1000 bins, and the y-axis was based on the liklihood of each parameter value occuring in the numerous saved geneologies. The 50+ summary curves from either PE1 or 2 were combined by plotting the sum of the probabilities at each bin against the average bin value from all the runs. The upper (a) and lower (c) bounds of the highest posterior density 95% (HPD95%) interval (black) span a distance on the x-axis that has a 0.95 probability of covering the actual value of *t* or N_a, and may be interpreted as confidence intervals. The peaks of the curves (b, red) correspond to the most likely value of each parameter, given all PE1 or 2 data. Here, the best estimate of *t* is ~140 000 years before present, and N_a is ~23 000 individuals.

Adaptive analyses

Protein (P) pairs exhibiting positive or purifying selection were determined by either: 1) the absolute number of substitutions per synonymous or nonsynonymous site after being subject to the Jukes-Cantor correction (dS, dN); 2) the ratio of dS/dN; or, 3) a lack of polymorphism (Table 4). Fifty-eight percent of the P pairs were identical (ignoring gaps, see above) and were interpreted as evidence of purifying selection (Fig. 10). There were more substitutions per synonymous site (dS) than dN in 62.5 ± 1.5% of the polymorphic pairs analyzed in SNAP; these were also interpreted as evidence of purifying selection. In pairs from which dS/dN could be calculated, instances of purifying selection outnumbered instances of positive selection approximately 4 to 1. The median intensity of selection on pairs from which dS/dN was available was 3.975 ± 0.275 for those under purifying selection, and 0.675 ± 0.035 for those under positive selection (Fig. 11).

Twenty-five sequences implied positive selection via dS/dN. Based on a blastx search of NCBI's nr, all but 2 were retroelements, transposases, reverse transcriptases, and integrases. The exceptions were a marker sharing 79% amino acid identity with several putative protein phosphatases, and a putative kinase sharing 61%ID with proteins that interact with receptor kinase VHI.

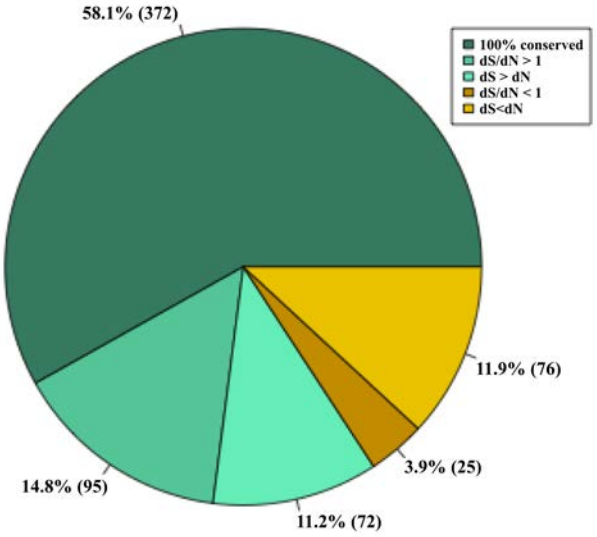

Figure 10. Relative proportion of P pairs under purifying selection (green) and positive selection (gold). Absolute counts are given in parentheses.

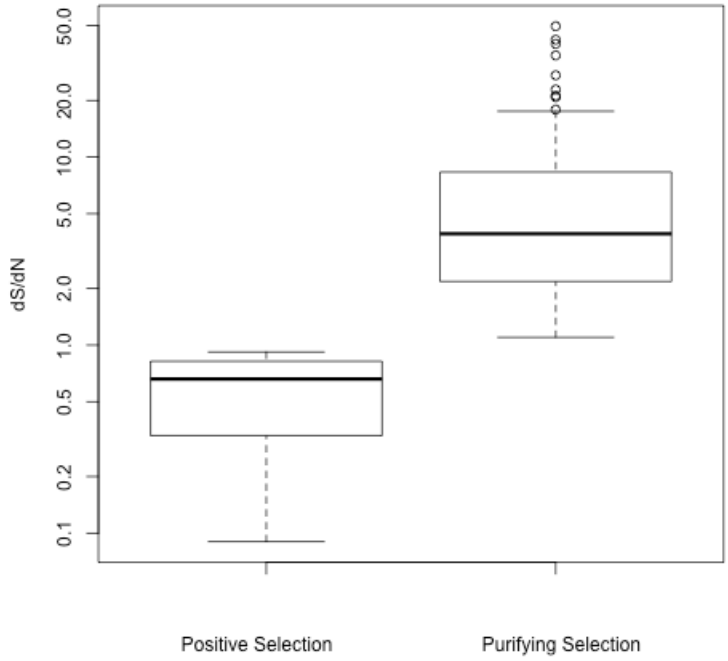

Figure 11. Intensity of selection on CG pairs depicted as the distribution of dS/dN values. The median for pairs under positive selection (dS/dN <1, n=25) was 0.675, the median for pairs under purifying selection (dS/dN > 1, n=95) was 3.975. The y-axis is logarithmic; circles are outliers.

Table 1. A summary of the primary results delivered by this investigation. The variance around the mean value from both paired-end (PE) sets is given where applicable.

Bases sequenced	1 353 797 132
101-bp reads	13 403 932
Reads remaining after q.c.	8 861 830
Contigs	178 926
Average contig length (bp)	94
CG pairs	11 123 ± 416
Nucleotide coverage for CG sequences: min / mean / max	2.8 / 10.4 / 4303.3
Time since splitting, *t* (years)	141 613 ± 10 572
Effective ancestral population size, N_a (individuals)	22 798 ± 154
Protein (P) pairs analyzed	320 ± 25
P pairs lacking polymorphism, purifying	186 ± 13
P pairs with dS>dN, purifying	83.5 ± 5.5
P pairs with dS<dN, positive	50.5 ± 6.5

Table 2. Bioinformatics results for paired-end (PE) data sets.

Flowcell		
	PE1	PE2
Bases sequenced	676 898 566	676 898 566
101-bp reads	6 701 966	6 701 966
Quality control		
Length = 0 after end-trimming	8 392	29 056
Contained unknown base (N)	3 314	1 485
Length < 5 after clipping adapters	24 657	44 495
Removed by quality filter	799 777	943 755
Length < 39 bp	1 312 596	1 374 575
Reads remaining after q.c.	4 553 230	4 308 600
Assembly		
Split by barcode	2 356 673 (51.8%) - Arctic 2 196 557 (48.2%) - Alpine	2 222 421 (51.6%) - Arctic 2 086 179 (48.4%) - Alpine
Contigs	47 564 - Arctic 44 257 - Alpine	44 749 - Arctic 42 356 - Alpine
Contig length (in bases): min / median / max	61 / 94 / 645 - Arctic 61 / 94 / 1228 - Alpine	61 / 94 / 461 - Arctic 61 / 94 / 561 - Alpine
CG pairs		
Total	11 539	10 707
Chloroplast (C)	19	22
Mitochondria (M)	10	8
Protein (P)	358	308
Non-coding (NC)	11 152	10 369

Table 3. Alignment statistics for mitochondrial (M), chloroplast (C), protein (P), and nuclear non-coding (NC) markers. Total columns refers to all columns aligned in each category for PE1 and 2 combined, not counting gaps. The percent of polymorphic sites for the nuclear genome (P and NC combined, not shown) is 2.10%.

Category	M	C	P	NC
Total columns (excluding gaps)	1 783	4 533	65 709	1 981 735
Identical columns	1 780	4 501	64 439	1 939 989
Variable columns	3	32	1 270	41 746
% of polymorphic sites	0.17%	0.71%	1.93%	2.11%

Table 4. Adaptive analysis outcomes for paired-end (PE) 1 and 2 data sets.

	PE1	PE2
P pairs with frame information	345	295
P pairs lacking polymorphism	199	173
P pairs analyzed in SNAP	146	122
dS>dN, purifying	89	78
dS<dN, positive	57	44
dS/dN > 1, purifying	51	44
dS/dN < 1, positive	13	12
dS/dN value, purifying: min / median / max	1.1 / 3.7 / 49.67	1.1 / 4.25 / 39.99
dS/dN value, positive: min / median / max	0.09 / 0.71 / 0.92	0.24 / 0.64 / 0.86

DISCUSSION

Demographic analyses: setting the stage

A Middle- to Late Pleistocene divergence

The probability that the Alaskan and Coloradan populations of *B. vivipara* became isolated from each other within the last 4 glacial ages is 0.95. Out of more than 450 million simulated genealogies, none suggest they split later than the older boundary of the glacial age known as pre-Illinoian A (as demarcated by marine isotope stage [MIS] 12 [Cohen and Gibbard 2011]), ~400 000 years ago. The most likely divergence date coincides with the late Illinoian glacial maximum, ~140 000 years before present (Fig. 12A). Thus, $H1_{alt1}$ was clearly supported: the populations stopped exchanging genes at some point—most likely during the second to last glaciation, and almost certainly during the Pleistocene. The confidence intervals on *t* span cool and warm ages, so I cannot definitively suggest that either glaciers or uninhabitable warm environments prevented these populations from exchanging gametes, but it is clear that the turbulent changes of the Quaternary evoked the well-studied species response: to move.

Habitat tracking

In order for the homologs from the sampled individuals to coalesce, the range must have looked different than it does today. At least once, a founding population contributed generations of offspring that tracked suitable habitat into isolated refugia. The species altered its range to match the environment that met its basic needs for survival and reproduction; thus, in terms of its capacity to respond to historic climate change, it is fair to say habitat tracking may be counted among its potential future responses.

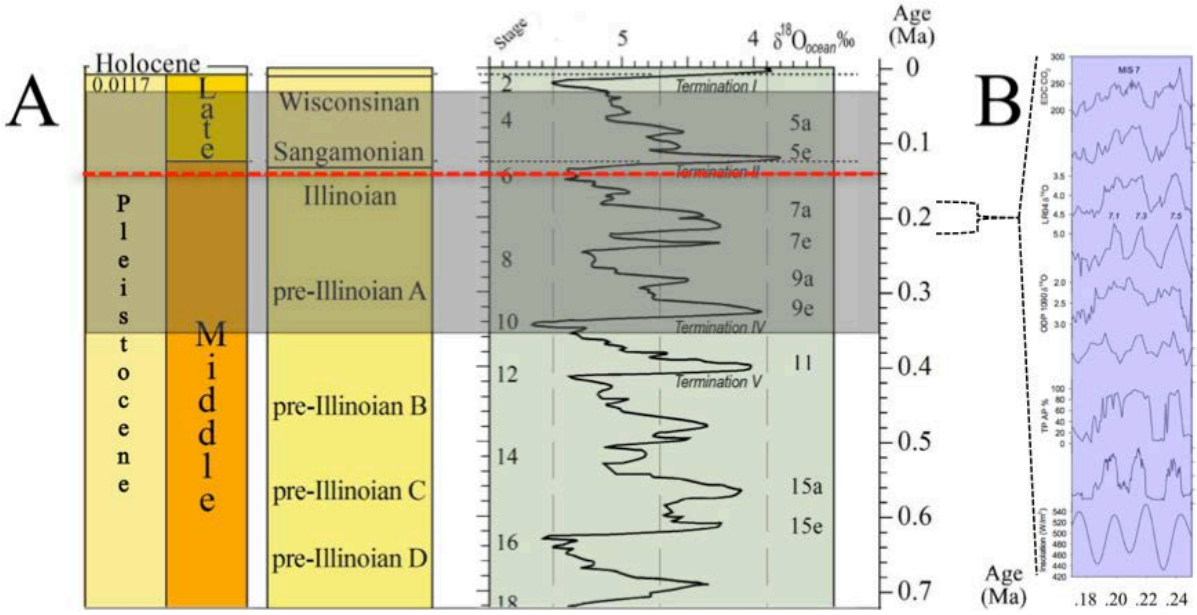

Figure 12. A) A correlation of chronostratigraphical subdivisions showing, from left to right, formal time divisions of the Middle and Late Pleistocene subseries, North American Stages, and Marine Isotope Stages (MISs), redrawn from Cohen and Gibbard (2011). Solid horizontal lines indicate observed boundaries, the red dashed line indicates the divergence date estimate (*t*), and the grey horizontal bar marks the 95% confidence intervals thereof. B) MIS 7 at a higher resolution. Various proxy climate records report multiple interglacial peaks during MIS 7, a milder interglaciation than the Sangamonian. Redrawn from Lang and Wolff (2011).

Fluctuating population size

The range of the ancestral population was not determined by this investigation, but its size is evident from my estimate of ancestral effective population size (N_a). That figure, ~23 000 individuals, could be compared to extant effective population sizes (N_e) to confidently add fluctuating N_e to *B. vivipara*'s potential demographic responses. If its preferred environmental space increased since splitting, local populations may be larger than the founding population. On the other hand, $N_a > N_e$ would suggest that the current state is not only the result of the plant moving to refugia, but also reducing N_e, and, consequently, genetic diversity. The data leave the comparison unanswered because they lack the

information needed to estimate N_e. The RAD markers are a useful starting point, though, for designing primers for loci that could be sequenced from multiple Alaskan and Coloradan individuals, which would then provide means to estimate N_e. Although not directly confirmed, a change in population size is expected to follow a change in range, and it is probably fair to say *B. vivipara* also responded demographically via a change in N_e.

Glacial influence

Some potential distributions of the founding population are offered which aid the interpretation of the demographic results. Initially, I predicted the last glacial age (Wisconsonian) split the populations. I imagined its preceding interglaciation as similar to the current and that the prehistoric distribution echoed the plant's modern range in Western North America. As the climate cooled and the two great continental ice sheets formed, the Arctic populations might have sprawled south and mingled with more southerly populations that had shifted down slope from Alpine refugia into the now cool lowlands. Birds and animals roaming the corridor between the Cordilleran and Laurentide might have assisted the plant's panmixia by transporting indigestible bulbils. Eventually, as the ice sheets fused together, the northern and southern populations would have retreated to their respective glacial age refugia. Alternatively, perhaps an Arctic or an Alpine population alone migrated to found the disjuncts. I further imagined that a version of this scenario was repeating itself throughout the long series of Quaternary cycles. Regardless of where the founding population was, its polymorphisms were carried into both daughter populations and the extant genetic markers would bear the signature of the last mixing event.

If a N/S glacial corridor was involved in the population’s dynamic history, what

characteristics of one glacial-interglacial period could inhibit the Alpine Bistort's panmixia while another supported it? Lang and Wolff (2011) concluded there is no reason to think of a single glacial-interglacial 'type' of cycle, and, indeed, a closer look at the Illinoian and Wisconsonian carries a plausible explanation as to why the divergence date settles at not the last, but the second to last glacial maximum. The key differences between the two are in the climate trends that precede them. The interglaciation before the Wisconsonian (the Sangamonian, demarcated by MIS 5e) appears as a severe 'spike' in marine and ice core isotope records, and serves as one of the few boundaries in Quaternary stratigraphy that researchers can all agree on (Fig. 12A). Whether seen as the official beginning of the Late Pleistocene sub-stage, the Tyrrhenian standard stage in marine records, or the variously named ages from northwest Europe, Russia, or North America, the last interglaciation saw a dramatic transition from the Illinoian glacial maximum to warmer-than-today conditions (Clark *et al*. 1993) and back to the formation of Wisconsonian ice sheets in little more than 20 000 years (Cohen and Gibbard 2010). On the other hand, the Illinoian is preceded by a series of almost indecisive temperature swings. These events, demarcated as MIS 7, are only hesitantly described as an interglaciation. At a higher resolution, MIS 7 itself shows up to 5 weak interglacial-glacial cycles, depending on the location and type of record (Fig. 12B). At their coolest, the mini-periods behave similarly to records close to fully glacial, yet at times and at some sites, there is an almost continuous period of weak interglaciation. An unbiased ranking of the relative strengths of the last 9 ages and their transitions highlighted MIS 7 as one of the weakest interglaciations and transitions—it is the exception to the rule that the oscillations become more exaggerated as we move closer to the present (Lang and Wolfe 2011). Perhaps its irregular flirtations with glacial conditions and soft transition into full

glaciation held a cool ice-free corridor open long enough for *B. vivipara* to occupy or otherwise traverse the distance from Alaska to Colorado, whereas the severity of the Sangamonian and its transition into the Wisconsonian prevented it.

Effects of polyploidy on demographic estimates

Certain considerations are required in the interpretation of N_a and *t* because I surveyed the genome copies in a decaploid organism more or less equally.

N_a calculations are based on the allelic diversity in previous generations, so the parameter values in the model were scaled by an integer, 5, to reflect the ratio of an individual to its contribution of alleles to the next generation (each decaploid parent contributes half of the alleles in the offspring).

T estimates were affected in that each duplicated genome was a source of potential paralogs. Analyzing paralogs in IMa2 would date the divergence of the sequences instead of the splitting of the populations. I took precautions in selecting orthologs from cluster families, but the incompleteness of the data suggests that some paralogs were included in the IMa2 runs.

How did paralogs in the IMa2 runs bias the results? The posterior likelihood curve for each run was a summary of the individual posterior probability curves for each locus analyzed. Paralogs may have widened and/or shifted summary peaks, and likewise, the peaks of the combined summaries, to higher values of *t*. The individual posterior probability curves for each locus in a run might display a multi-modal distribution: orthologs peaking at the population split and paralogs peaking at one (or several) other values of *t*. To strengthen the demographic estimates, then, I would next analyze an individual batch of 200 loci one

locus at a time with very high priors to observe the distribution of posterior probability peaks on a locus-by-locus basis.

Effects of cloning on demographic estimates

Clonal population genetics trends should also be considered in the interpretation of N_a and *t*. Genetic variation in clonal populations is expected to be lower than sexually reproducing populations, due to the absence of segregation and recombination. In a survey of allozyme markers, populations of *Bistorta vivipara* reflect these expectations: their genotypic diversity and structure was similar to the average for clonal species (Diggle *et al*. 1998). *T* estimates are not affected by cloning because the mode of inheritance (sexual or asexual) is separate from the neutral mutation rate (Nordborg 2000).

However, the partially clonal populations mean that effective population size estimates in this investigation are probably underestimates. N_a is based on allelic diversity. In clonal populations, heterozygosity is maintained, but the total number of alleles does not change much from generation to generation (Balloux *et al*. 2003). In the IMa2 model, the seemingly slow-to-change allelic diversity resulted in lower estimates of N_a because it did not account for generations of clones.

Violation of assumptions

The demographic results are based on assumptions that, if violated, would influence *t* and N_a in different ways. Here, I revisit the assumptions in the IMa2 model briefly. Assumptions for pairing orthologs and the accuracy of the sequence data are dealt with further below.

If the variability between Arctic and Alpine populations is not the result of microevolutionary changes imposed by physical isolation but rather the result of exchange with diverse populations, then the demographic estimates are not informative. I think the argument for their physical isolation is sound and that gene exchange is minimal, but how would a small amount of migration affect the results? If another population more closely related to both sampled populations were exchanging genes from Arctic to Alpine tundra, it would widen the pool of naturally occurring NC diversity—diversity that I counted as evidence of time passing since the split. Thus, small amounts of migration would shift *t* closer to the present.

If the markers I designated as NC were under natural selection, estimates for *t* and N_a could be off in either direction. If there were a degree of purifying selection, then mutations would be occurring less frequently than I assumed they were, and the results would be underestimates. If they were encouraged by natural selection to continually change, then the actual split time would be younger than the results suggest. The categorization process limits this possibility that selection is in play, but the true strength is in numbers in this case. It is unlikely that most, if any, of the NC loci were under selection.

Lastly, if the assumptions for recombination were violated, then the confidence intervals on both estimates would be artificially narrow. The numerous loci must be counted as independent records if the probability estimates are to have statistical merit. Linked loci introduce false calculations: if two loci are linked, the probability of the parameter values at each step must be calculated based on the likelihood of generating the data at all of the sites together. If a portion of a locus is probable, and another is not, the probability of the data would incorrectly increase the posterior probability of that parameter set if the unlikely

portion were considered as a separate record. Cloning as a reproductive strategy does not link the individual loci, as long as the population is not exclusively asexual (Balloux *et al*. 2003).

Next-generation estimates

The thousands of sequences and simulated geneologies incorporated into the demographic analyses make these results noteworthy. Coalescent-based computer algorithms are extremely effective at exploring polymorphism data (Nordborg 2000), and independent records of genealogic processes (i.e., unlinked neutral loci) are a priority of hypothesis-testing coalescent models. Despite some theoretical and practical assumptions in my methodology, I find the combination of next-generation sequencing and coalescent-based approaches highly functional. The uncertainties of sequencing errors, IMa2 model violations, and paralogy are balanced by analyses of millions of character states (nucleotides) from a single data set. The potential accuracy of the demographic estimates rings of the profound impact post-genomic technology is having in various disciplines.

Adaptive analyses

No evidence of a sweeping adaptive response

In ~140 000 years, the protein-coding (P) sequences examined did not change at all (the majority), changed only at synonymous sites (some), or changed at nonsynonymous sites (very few). Those few with strong signatures of positive selection ($dS/dN < 1$) were identified as sequences expected to escape purifying selection. All but two were similar to some class of retroelement. But for their existence, I conclude: an adaptive response was not

in play in this population over the studied time frame.

One of the two exceptions was a sequence similar to a kinase involved with auxin and brassinosteroid signaling (see NCBI Gene ID: 837960). The other matched several putative protein phosphatases. If either contains a region under positive selection through the glacial cycles, clocking the rate of those evolutionary changes is desirable. As of this writing, 'climate change genes' are still dubious. Some of the most hopeful candidates (e.g. flowering-time genes FRI and FLC [Atwell *et al*. 2010]) are still missed in genome-wide SNP surveys in model organisms (Fournier-Level *et al*. 2011). The *Bistorta* sequences provide a point from which we can expand the search for local adaptation in plant genomes. To learn more, primers based on these sequences can be developed to obtain longer portions of the genes. Comparative analysis of longer markers would be an informative starting point for functional analysis.

The scarcity of markers under positive selection suggests the adaptive potential may be at odds with vivipary and small effective population sizes, two things that reduce the raw material for an adaptive response. Genotypic diversity is reduced both as asexual reproduction increases (Balloux *et al*. 2003) and after a population bottleneck (Landergott *et al*. 2001). The combination of reproducing clonally and losing diverse individuals during the retreat to refugia may explain the lack of adaptive genetic changes in all but two of hundreds of sequences. Or, perhaps it reflects a bias of sampling the entire genome: more gDNA is expected to be under purifying than positive selection. Thus, one might expect examples of purifying selection to dominate in a pangenomic survey.

Interpreting the capacity to adapt

Whether on a gene in a genome or a codon in a gene, natural selection acts locally. The question pertinent to this data, then, is: if an adaptive response did occur since splitting, was it on a locus sequenced in this survey? My assessment of *Bistorta vivipara*'s adaptive capacity is limited to the markers sequenced in this study. The data may contain partial genes under selection, but sites with a fitness-increasing polymorphism may not be included. It is also of note that some potentially adaptive genetic responses would not register in these analyses (e.g. insertions, deletions, and polyploidy); nor, was the genetic basis of phenotypic plasticity assessed.

I further consider the utility of these markers for assessing adaptive potential in the light of contingency tables that accompany MEGA 5's (Tamura *et al.* 2011) test for selection. This alternate implementation of a Nei and Gojobori (1986) – based test for selection has benefits over SNAP, but was not easily automated using Perl. The codon based Z-test for selection in MEGA 5 calculates the probabilities of rejecting one of three null hypotheses: positive, purifying, and neutral selection, given two aligned sequences. The Z-test adds statistical significance to the argument that the observed polymorphisms are the result of selective pressure, not random chance, for each marker. I manually analyzed 10 random CG P pairs with MEGA 5's Z-test and found that SNAP and MEGA 5 agree on the form of selection in all 10. However, only 1 in 10 dS/dN-based measure of selection was supported by P values less than 0.05 in MEGA 5. Most of the markers were too short to statistically differentiate the signals of positive or purifying selection from neutral evolution (although, the probability of a rejecting a hypothesis of neutral evolution was at least 40% greater than the probability of rejecting a hypothesis of either positive or purifying selection in all cases). The non-significant results of a quick sampling of CG P pairs in MEGA 5 calls into question

the usefulness of these markers themselves to gauge adaptive potential.

On the other hand, $H2_{alt2}$ was clearly supported by observed changes in P pairs, and the proportions of variable sites should reflect the influence of selective pressures. The sequences could be longer, but that would run the risk of introducing recombination. Here again, a drawback in methodology may be balanced out by the volume of markers and the ease of their retrieval.

Lastly, more than half of the sequences that did not have polymorphisms at both types of substitution site have a single variable site, and might be more effectively used in population genetic diversity work or genome mapping rather than an analysis of adaptive genetic changes.

Patterns of variation in three genomes

The chloroplast (CP), mitochondrial (M), and nuclear (P and NC) markers show patterns of mutations that we generally expect for plant DNA: the mtDNA is slowest to change, and cpDNA and nuclear DNA is faster.

Similarities between chloroplast and mitochondrial genomes emerged as I was categorizing plastid sequences, which lends support to the idea that horizontal gene transfer is common in angiosperms. Goremykin *et al*. (2008) report that 42.4% of the *Vinis vitifera* chloroplast genome has been incorporated into the mitochondrial genome. In *B. vivipara*, every sequence identified as either CP or M had multiple analogs in archived chloroplast and mitochondrial genomes. This made unambiguous categorization difficult, but perhaps more importantly, highlights the potential importance of nuclear markers in phylogeography. CP and M markers are not free from recombination and chromosomal rearrangement mediated

by the opposite genome, and they harbor little overall variation, thus their usefulness in phylogenetic reconstruction is challenged.

Recent work highlights how historic inference can be missed without the utility of nuclear data (Galbreath *et al*. 2011). Most phylogenetic work in plants has been based on chloroplast phylogenies, which may work well for distantly related species. Yet, comparisons between closely related species are likely to yield more precise and informative measurements of evolutionary patterns and rates (Brown *et al*. 1982). Unfortunately, even with long stretches of plastid DNA, there is not much variation to be had. The vital sets of organelle genes naturally resist evolutionary change (Zurawski and Clegg 1987), and the problem may be compounded by viviparity. The roadblock is essentially this: low levels of sequence variation restrict conclusive results in phylogeography and nuclear sequences are difficult to recover when designing and implementing species-specific primers from scratch. Comparing *Bistorta* genomes demonstrated nicely that we can get informative nuclear markers without prior knowledge of the target genome. Next-generation sequencing methods like restriction-site associated DNA (RAD) markers circumvent the problem of targeting specific nuclear markers. As was predicted for many other disciplines in the life sciences, next-generation sequencing has the potential to revolutionize plant phylogeography by facilitating access to vast portions of nuclear DNA from numerous individuals simultaneously.

Future biotic response

Given that range contractions, isolation, and loss of genetic diversity have driven plant extinctions of the past (Jackson and Weng 1999), the need to protect the remaining gene

pools of tundra plants is apparent. We know *B. vivipara* tracked its preferred habitat into isolated refugia during historic climate cycles. The founding population contributed the initial population genetic structure to the modern populations following a corresponding change in effective population size. Populations were isolated from gene-exchange across geographic space, while individual genomes were passed on, often asexually, to the next generation. Mutations accumulated in neutrally evolving DNA, but, despite differences in rainfall, radiation, competition, and pollinator interaction, the Arctic and Alpine populations seem to be working off of the same blueprints for protein-coding DNA. The lack of evidence for adaptation does not rule out the capacity to adapt at a genetic level: it may be a single nucleotide difference, not sequenced here, that prevents local extinction. But adaptation does require variation. Variation allows the species to occupy different microhabitats, which broadens the tolerance of the population from that of an individual. On the one hand, multiple genome copies may be a source of variation that increases adaptive potential. And, despite viviparity, over the long run, variation may be maintained by refugia: new allelic combinations arise from plant life diverging in isolation (Gavrilets 2003). Once they arise due to mutation, heterozygosity is maintained at that locus via asexual propagation (Balloux *et al*. 2003). Genotypic variation would be boosted at the species level by the mixing of diverse alleles maintained in refugia, provided they have the chance to come back together at some point. On the other hand, the last retreat to refugia was probably a bottleneck event, which reduces variation. Plus, any adaptive potential must draw on the standing variation, for the need to adapt is now.

Given evidence of genetic adaptation is scarce, more pressure is placed on phenotypic plasticity and demographic responses as alternatives to extinction. It has yet to be seen how

a plastic response may help *B. vivipara* out of a potentially threatened position, but it is historically resilient. It survived the Quaternary period in what could be argued to be the most difficult landscape to weather those changes. Explicit (Hultén 1968) and implicit (Wookey *et al*. 1995) attention to the plant's impressive morphological variation is palpable in the literature. Perhaps the genetic basis for a wide environmental tolerance is already carried in this species, regardless of the effects of clonality and refugia on variation. Genome duplications have been long been suspected to aid the durability of tundra plants simply due to association: the relative proportion of polyploidy is greater in the high-latitudes than anywhere else (Löve and Löve 1974). It is possible that polyploidy facilitates plasticity, and the entire system is effectively propagated clonally.

Improvements and future work

Lessons from this exploratory investigation should inform forthcoming research based on these data or techniques. Observations towards potential improvement are discussed in broad categories below.

Effectively sampling populations

The often clonal nature of the *B. vivipara* life cycle encouraged my commitment to a sample size of one—sampling more would have risked potential resources by sequencing identical genomes. However, data from multiple individuals per population would have offered several advantages. Such data would provide a means to estimate N_e and facilitate the use of a migration parameter in the coalescent simulations. Estimating N_e for each population would have allowed nucleotide diversity to be compared within and across

populations, enhancing the interpretation of why so few adaptive genetic changes occurred.

Data from more individuals would have strengthened my conclusions, but more samples can complicate the bioinformatics processing. Theoretical and computational strategies for managing large volumes of data will only get easier as researchers tailor their efforts to next-generation sequencing. As such, we can expect (and benefit from) multi-locus studies that involve a more complex sampling scheme than this one in future investigations.

Effectively generating RAD markers

A longer restriction digest of unamplified genomes might have recovered more homologs from these two samples. The fact that not all, or even a majority, of markers had a homolog across samples suggests I recovered only a fraction of the 400-500 bp fragments that would be generated *in silica* using a complete *B. vivipara* genome and *PsiI*'s recognition site. If the *in vitro* digest were complete, then all the markers would have had a homolog. What, then, prevented a more complete digest of the extracted genomes? To a single restriction enzyme with a limited turnover rate, a genome is a massive physical structure. Adding time to the digest might have allowed the discovery and cleavage of more sites. Adding more enzymes may also have had a similar effect; or, perhaps not. Activity may be hindered above a certain number of enzymes. Careful experiments with the digest length and concentration of enzymes that focused on increasing the volume of DNA that migrates to the desired size band during electrophoresis would probably pay off with a greater number of markers, and thus, of homologous markers, in the bioinformatic stages.

In this investigation, I digested amplified gDNA. This was initially done to prevent having too little DNA to submit to the sequencing center after the size-selecting step. At the

time, paired-end GAII sequencing of small amounts of DNA was untested by my colleagues at HU, so the amplification step was a precaution against wasting an expensive sequencing reaction. Since then, several groups have had successful runs with smaller amounts of DNA for Illumina sequencing. Is it possible the excessive genetic material *in vitro* actually prevented the enzymes from effectively scanning the full genome? This would be especially true if there was preferential digestion for some parts of the genome: the numerous copies would keep the enzymes 'busy' cutting the easily accessible restriction sites. The result of this scenario would be excessively high coverage for some markers. The highest coverage of one marker (4303.3) compared to the average coverage of all markers (10.4) lends support to this possibility, and leads me to think that removing the whole genome amplification step would not only eliminate possible bias, but also increase the efficiency of the restriction digest, and thus the number of homologs to compare.

Effectively using next-generation sequencing

Paired-end sequencing was created to help researchers who randomly fragment genomes, use the small pieces as Illumina inserts, and reassemble the raw reads back into large contigs. The paired-ends work like addresses to help place reads in the proper order when repetitive regions make placement via overlap ambiguous. My samples occupied 1 of 8 available channels on an Illumina flowcell that was shared with research groups using PE data to sequence whole genomes, which automatically carried me into the PE option. I had hoped the distance between my PE reads was large enough to allow for recombination in my target species, which would have doubled the number of independent markers for my analyses. The almost perfectly parallel results from PE1 and PE2 suggest they were linked.

It would have been more effective to reduce my insert size to 200, just short of twice the read length. That way, I would have sequenced the middle of the fragments, had longer contigs, and analyzed a single set of data. Longer markers would have strengthened all my analyses, especially the tests for selection. On the other hand, longer markers might increase the likelihood of recombination, so a recombination step, such as the 4-gamete test, would be recommended for the analysis pipeline.

I took risks of introducing artificial variation with whole genome amplification, enrichment of the size-selected DNA, and massively parallel sequencing, because I relied on the assumption that data carried into the final analyses was a faithful representation of the state of the genome(s) I extracted from both plants. Violations of this assumption have unpredictable effects on the results. I followed protocols devised to reduce bias at both amplification steps, but I suspect the quality control (q.c.) pipeline was my most effective weapon for safeguarding against sequencing errors. Its importance for any next-generation sequencing project cannot be overstated. The q.c. results of this investigation are particularly relevant for researchers relying on PE reads: the PE2 reads had dramatically lower quality scores than PE1, especially near the limit of the read length. Interestingly, there were more unknown bases (reads containing N) in PE1 than PE2.

Effectively handling paralogs and orthologs

At its worst, the ambiguity of paralogs and orthologs caused the erroneous comparison of sequences for any number of my CG pairs. Given the incompleteness of the data, I have no doubt some pairs were, indeed, paralogous. When generating divergence estimates, these pointed to the timing of a duplication event, not the split, and led to an

unwelcome widening of my confidence intervals on *t*. In tests for selection, these gave meaningless counts of synonymous and nonsynonymous polymorphisms, for a duplicated gene could accumulate changes at equal rates for all sites if it was not under any selective pressure (Zhang 2003). My hope is that given the sheer amount of data and the application of my assumptions when pairing markers, orthologous pairs outnumbered paralogous pairs and the results were skewed towards the correct interpretations. It is reassuring that the results are compatible with *a priori* expectations: *t* falls within the Quaternary period, and instances of purifying selection outnumber instances of positive selection.

One possible idea for quantitatively testing whether paralogs accumulate more mutations than orthologs would be a carefully devised simulation incorporating duplication events, population splits, and the requisite substitution rates and models for different genomic regions. This posit of future work, or any forthcoming techniques to untangle orthologs from paralogs would undoubtedly be welcome by the research community. Even with our current understanding, there may be some powerful analyses yet to be done with the Arctic/Alpine *Bistorta* data. For example, if we assume cluster family members are paralogs, the demographic and adaptive analyses I applied across samples could be applied within samples, and, with some thought, yield such desirable outcomes as dating ploidy events or finding evidence of functional selection on gene copies in a polyploid.

Effectively analyzing thousands of loci

There was no precedent in the literature for analyzing thousands of loci for much else but SNPs and genetic mapping when I designed this investigation. The Perl programming language became invaluable for implementing tasks that were well-established for use with a

few sequences on the many. The demographic analyses with IMa2 were particularly challenging because of the computational demands (e.g., random access memory [RAM], CPU time) when analyzing multiple loci in a coalescent framework. Even with HU's world-leading computing facilities, I was unable to process more than 200 loci at a time without breaching RAM limits. To use the thousands of markers I recovered, the runs were performed separately, which was awkward and time consuming. My method of combining the results, summing the probability from all the simulations for each of 1000 bins from 0 to the prior for *t* and N_a, is 'legal' for combining runs performed on the same data set, but is not usually done for the results from different data sets. The only reason it worked for these data is that the batches of 200 loci, although different, all had the same characteristics (i.e. nuclear, unlinked, non-coding DNA of approximately the same length, from the same genomes). The method worked well for x-axis values less than the median and was less accurate closer to the maximum. The reason for this is the difference between the constant of a larger bin in two different runs is potentially greater than that of a two smaller bins because it is a larger number scaled by the mutation rate. For example, the bulge on the right side of the final posterior probability curve for *t* (Fig. 8A) is probably an artifact of my method of combining results, not an increase in probability for those parameter values. Just recently, the author of IMa2 published a manuscript outlining a theoretical framework using thousands of loci simultaneously, hinting at a practical application of high-throughput sequence data for estimating demographic parameters (Wang and Hey 2010). It does require multiple individuals per population, which my data lack, but would be an excellent direction for improvement from here.

Conclusion

Global change studies forecast a reorganization of terrestrial biomes that will give organisms less time to move and/or adapt than transitions in and out of glacial ages (Jansen *et al*. 2007). The response of *Bistorta vivipara*, a hardy tundra species, to transitions past suggests its capacity to respond is primarily demographic. Some demographic responses, like retreat to refugia, may combat extinction, but only if refugia still exist or can be colonized by extant populations. Unfortunately, the severity of the predicted change does little to encourage either adaptation or viable shifts in population structure as sustainable. With genetic adaptation ruled out as a response and potentially no way to keep up with (or find) ideal environmental conditions, the remaining option is irrevocably final—unless phenotypic plasticity can overcome the potential challenge.

Broadly applied to other plant species, the implications are enormous. Some species may fare better than others (Willis *et al*. 2010), but plants that provide us with food, clothing, and shelter will probably struggle to adapt genetically in the time period imposed by anthropogenic climate scenarios. As a result, we may see massive rearrangements in population demographics worldwide that tend toward circumstances—i.e. loss of genetic diversity—previously associated with extinction (Jackson and Weng 1999).

LITERATURE CITED

Abbott RJ, Brochmann C. 2003. History and evolution of the arctic flora: in the footsteps of Eric Hultén. Molec Ecol 12(2): 299-313.

Abbott RJ, Smith LC, Milne RI, Crawford RMM, Wolff K, Balfour J. 2000. Molecular analysis of plant migration and refugia in the arctic. Science 289(5483): 1343-46.

Alexeyenko A, Tames I, Liu G, Sonnhammer ELL. 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22:e9-15.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215: 403-10.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25(17): 3389-402.

Arft AM, Walker MD, Gurevitch J, Alatalo JM, Bret-Harte MS, Dale M, Diemer M, Gugerli F, Henry GHR, Jones MH, *et al*. 1999. Responses of tundra plants to experimental warming: meta- analysis of the international tundra experiment. Eco Mono 69(4): 491-511.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, *et al*. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465: 627-31.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3(10)e3376: 1-7.

Baker HG. 1959. Reproductive methods as factors in speciation in flowering plants. CSH Symp Quant Biol 24: 177-91.

Balloux F, Lehman L, de Meeûs T. 2003. The population genetics of clonal and partially clonal diploids. Gen Soc Am 164: 1635-44.

Barnosky AD, Koch PL, Feranec RS, Wing SL, Shabel AB. 2004. Assessing the causes of late Pleistocene extinctions on the continents. Science 306: 70-75.

Bennington CC, McGraw JB. 1995. Natural selection and ecotypic differentiation in *Impatiens pallida*. Ecological Monographs 65(3): 303-23.

Billings WD. 1974. Arctic and alpine vegetation: plant adaptations to cold summer climates. Arctic and alpine environments. Ives JD, Barry RG, eds. London: Methuen; 999p.

Borgen L, Bengt J, editors. 1997. Variation and evolution in arctic and alpine plants: proceedings of VI international symposium of IOPB (international organization of plant biosystematists). Copenhagen: Council for Nordic Publications in Botany; 239 p.

Bradley RS. 1985. Quaternary paleoclimatology: methods of paleoclimatic reconstruction. Boston: Allen and Unwin; 472 p.

Brown WM, Prager EM, Wang A, Wilson AC. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225-39.

Brubaker LB, Anderson PM, Edwards ME, Lozhkin AV. 2005. Beringia as a glacial refugium for boreal trees and shrubs: new perspectives from mapped pollen data. J Biogeog 32: 833-48.

Charmantier A, McCleery RH, Cole LR, Perrins C, Kruuk LEB, Sheldon BC. 2008. Adaptive phenotypic plasticity in response to climate change in a wild bird population. Science 320: 800-803.

Chiu JC, Lee EK, Egan MG, Sarkar IN, Coruzzi GM, DeSalle R. 2006. OrthologID: automation of genome-scale ortholog identification within a parsimony framework. Bioinformatics 22: 699-707.

Clark PU, Clague JJ, Curry BB, Dreimanis A, Hicock SR, Miller GH, Berger GW, Eyles N, Lamothe M, Miller BB *et al*. 1993. Initiation and development of the Laurentide and Cordilleran ice sheets following the last interglaciation. Quat Sci Rev 12(2): 79-114.

Clarke JA, Johnson RE. 2005. Comparisons and contrasts between the foraging behaviors of two white-tailed ptarmigan (*Lagopus leucurus*) populations, Rocky Mountains, Colorado, and Sierra Nevada, California, U.S.A.. Arct Antarct Alp Res 37(2): 171-6.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucl Acids Res 38(6): 1767-71.

Cohen KM, Gibbard PL. 2011. Global chronostratigraphical correlation table for the last 2.7 million years v. 2011. [online]. Avail from: http://www.quaternary.stratigraphy.org.uk/correlation/POSTERstratchart %20v2011.jpg [2011 Sept 7]

Comes HP, Kadereit JW. 1998. The effect of Quaternary climatic changes on plant distribution and evolution. Trends Plant Sci 3(11): 432-438.

Comiso JC. 2002. A rapidly declining perennial sea ice cover in the Arctic. Geophys Res

Let 29(20): 17-1-4.

Coope GR. 1995. Insect faunas in ice age environments: why so little extinction? Extinction Rates. Lawton JH, May RM, eds. Oxford: Oxford University Press; 233 p.

Darwin C. 1859. On the origin of species by means of natural selection or the preservation of favoured races in the struggle for life. London: John Murray; 502 p.

Davey JW, Blaxter ML. 2011. RADSeq: next-generation population genetics. Brief Funct Genomics 9(5): 416-423.

Davis MB, Shaw RG. 2001. Range shifts and adaptive responses to Quaternary climate change. Science 392(5517): 673-79.

Diggle PK, Lower S, Ranker TA. 1998. Clonal diversity in alpine populations of *Polygonum viviparum* (Polygonaceae). Int J Plant Sci 159(4): 606-15.

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF. 2008. Evolutionary genetics of genome merger and doubling in plants. Annu Rev Gennet 42: 443-61.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics 148: 1667-86.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acid Res 32(5): 1792-1797.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19): 2460-1.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. accuracy assessment. Genome Res 8(3): 175-185.

Felsenstein J. 2005. Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? Mol Biol Evol 23(3): 691-700.

Flora of North America Editorial Committee. 2005. Flora of North America, north of Mexico, volume 5: Magnoliophyta: Caryophyllidae part 2. New York: Oxford University Press; 656 p.

Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM. 2011. A map of local adaptation in *Arabidopsis thaliana*. Science 334: 86-9.

Galbreath KE, Cook JA, Eddingsaas AA, DeChaine ED. 2011. Diversity and demography in

Beringia: multilocus tests of pale distribution models reveals the complex history of arctic ground squirrels. Evolution 65(7): 1879-96.

Gavrilets S. 2003. Perspective: models of speciation: what have we learned in 40 years? Evolution 57: 2197-215.

Gienapp P, Teplitsky C, Alho JS, Mills JA, Merilä J. 2008. Climate change and evolution: disentangling environmental and genetic responses. Mol Ecol 17: 167-78.

Geber MA, Dawson TE. 1993. Evolutionary responses of plants to global change. Biotic Interactions and global change. Kareiva PM, Kingsolver JG, Huey RB eds. Sunderland: Sinauer; 480 p.

Goremykin VV, Salamini F, Velasco R, Viola R. 2008. Mitochondrial DNA of *Vinis vitifera* and the issue of rampant horizontal gene transfer. Mol Biol Evol 26(1) 99-110.

Hamilton TD, Thorson RM. 1983. The Cordilleran ice sheet in Alaska. Late-Quaternary environments of the United States, Vol 1: the late Pleistocene. Porter SC, ed. Minneapolis: University of Minneapolis Press; 38-52.

Haubold B, Wiehe T. 2004. Comparative genomics: methods and applications. Naturwissenschaften 91: 405-421.

Hays JD, Imbrie J, Shackleton NJ. 1976. Variations in the Earth's orbit: pacemaker of the ice ages. Science 194(4270): 1121-32.

Henry GHR, Freedman B, Svoboda J. 1986. Effects of fertilization on three tundra plant communities of a polar desert oasis. Can J Bot 64(11): 2502-7.

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167: 747-60.

Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. PNAS 104(8): 2785-90.

Hey J. 2010. Isolation with migration models for more than two populations. Mol Biol Evol 27(4): 905-920.

Hey J. 2010b. Documentation for IMa2.

Hoffman RW, Braun CE. 1975. Migration of a wintering population of white-tailed ptarmigan in Colorado. J Wildl Manage 39(3): 485-90.

Hopkins DM, editor. 1967. The bering land bridge. Stanford: Stanford University

Press; 495 p.

Hopkins DM, Matthews Jr. JV, Schweger CE, Young SB, editors. 1982. Paleoecology of Beringia. New York: Academic Press; 489 p.

Hosono S, Far qi AF, Dean FB, Du Y, Sun Z, Wu X, Du J, Kingsmore SF, Egholm M, Lasken RS. 2003. Unbiased whole-genome amplification directly from clinical samples. Genome Res 13: 954-64.

Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W, *et al*. 2001. The Arabidopsis information resource (TAIR): a comprehensive database and web-based information retrieval, analysis and visualization system for a model plant. Nucl Acids Res 29(1): 102-5.

Hubbard TJP, Akan BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, *et al*. 2007. Ensembl 2007. Nucl Acids Res 35: D610-61.

Hultén E. 1937. Outline of the history of arctic and boreal biota during the Quaternary period: their evolution during and after the glacial period as indicated by the equiformal progressive areas of present plant species. Stockholm: Thule; 168 p.

Hultén E. 1968. Flora of Alaska and neighboring territories; a manual of the vascular plants. Stanford: Stanford University Press; 1008 p.

Hultén E. 1971. The circumpolar plants, volume 2, dicotyledons. Stockholm: Almqvist & Wiksell; 463 p.

Hutchinson GE. 1957. Concluding remarks, Cold Spring Harbor Symposium. Quant Biol 22: 415-27.

Hyatt D, Snoddy J, Schmoyer D, Chen G, Fischer K, Parang M, Vokler I, Petrov S, Locascio P, Olman V, *et al*. 2000. Improved analysis and annotation tools for whole-genome computational annotation and analysis: GRAIL-EXP genome analysis toolkit and related analysis tools. Genome Sequencing and Biology Meeting.

Irving L, West GC, Peyton LJ, Paneak S. 1967. Migration of willow ptarmigan in arctic Alaska. Arctic 20(2): 77-85.

Ives JD, Barry RG, editors. 1974. Arctic and alpine environments. London: Methuen; 999 p.

Illumina, Inc. 2008. Preparing samples for paired-end sequencing. [online]. Avail from: http://medicine.yale.edu/keck/ycga/sequencing/Illumina/protocols.aspx [2011 Aug 22]

Incredio. 2009. File:MilankovitchCyclesOrbitandCores.png. [online]. Avail from:

http://en.wikipedia.org/wiki/File:MilankovitchCyclesOrbitandCores.png [2011 Sept 7]

Jackson ST, Overpeck JT. 2000. Responses of plant populations and communities to environmental changes of the late Quaternary. Paleobiology 26(4): 194-220.

Jansen E, Overpeck J, Briffa KR, Duplessy J-C, Joos F, Masson-Delmotte V, Olago D, Otto-Bliesner B, Peltier WR, Rahmstorf S *et al*. 2007. Palaeoclimate. Climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change. Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL, eds. Cambridge: Cambridge University Press, 996 p.

Jackson ST, Overpeck JT, Webb III T, Keattch SE, Anderson KH. 1997. Mapped plant-macrofossil and pollen records of late Quaternary vegetation change in eastern North America. Quat Sci Rev 16(1): 1-70.

Jackson ST, Weng C. 1999. Late Quaternary extinction of a tree species in eastern North America. PNAS 96(24): 13847-52.

Johnson AW, Packer JG. 1968. Chromosome numbers in the flora of Ogotoruk Creek, N.W. Alaska. Bot Notiser 121: 403-56.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. Mammalian protein metabolism. Munro HN, ed. New York: Academic Press; 763 p.

Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61: 893-903.

Kingman JFC. 1982. The coalescent. J Appl Probab 19A: 27-43.

Korber B. 2000. HIV signature and sequence variation analysis. Computational analysis of HIV molecular sequences. Rodrigo AG, Learn GH, eds. Dordrecht: Kluwer Academic Publishers; 312 p.

Körner C. 1995. Alpine plant diversity: a global survey and functional interpretations. Arctic and alpine biodiversity: patterns, causes, and ecosystem consequences. Chapin III FS, Körner C, eds. Berlin: Springer, 332 p.

Landergott U, Holderegger R, Kozlowski G, Schneller JJ. 2001. Historical bottlenecks decrease genetic diversity in natural populations of *Dryopteris cristata*. Heredity 87: 344-55.

Lang N, Wolff EW. 2011. Interglacial and glacial variability from the last 800 ka in marine, ice and terrestrial archives. Clim Past 7: 361-80.

Leduc P. 2003. Pollen Viewer. [online]. Avail from: http://www.ncdc.noaa.gov/paleo/pollen/viewer/webviewer.html [2011 Sept 7]

Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, *et al*. 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucl Acids Res 34: D572-D580.

Li L, Stoeckert Jr CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-89.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucl Acids Res 33(20): 6494-506.

Löve A, Löve D. 1974. Origin and evolution of the arctic and alpine floras. Arctic and alpine environments. Ives JD, Barry RG, eds. London: Methuen; 999p.

Löve A, Löve D. 1975. Cytotaxonomical atlas of the arctic flora. Vaduz: J. Cramer; 598 p.

Löve A, Löve D, Kapoor BM. 1971. Cytotaxonomy of a century of Rocky Mountain orophytes. Arct Alp Res 3: 139-65.

Löve A, Ritchie JC. 1966. Chromosome numbers from central northern Canada. Can J Bot 44: 429-39.

Marie D, Brown SC. 1993. A cytometric exercise in plant DNA histograms with 2C values for 70 species. Biol Cell 78: 41-51.

Margulies EH, Birney E. 2008. Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. Nature Rev Gen 9: 303-13.

Martin AP, Burg TM. 2002. Perils of paralogy: using HSP70 genes for inferring organismal phylogenies. Syst Biol 51(4): 570-87.

Mosquin T, Hayley DE. 1966. Chromosome numbers and taxonomy of some Canadian arctic plants. Can J Bot 44: 1209-18.

Moyes K, Nussey DH, Clements MN, Guinness FE, Morris A, Morris S, Pemberton JM, Kruuk LEB, Clutton-Brock TH. 2011. Advancing breeding phenology in response to environmental change in a wild red deer population. Glob Chng Biol 17(7): 2455-69.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3): 443-53.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3(5): 418-26.

Nordborg M. 2001. Coalescent theory. Handbook of statistical genetics. Balding DJ, Bishop MJ, Cannings C, eds. Chichester: John Wiley & Sons, Inc.; 1488 p.

Packer JG, McPherson GD. 1974. Chromosome numbers in some vascular plants from northern Alaska. Can J Bot 52: 1096-99.

Parsons AN, Press MC, Wookey PA, Welker JM, Robinson CH, Callaghan TV, Lee JA. 1995. Growth responses of *Calamagrostis lapponica* to simulated environmental change in the sub-arctic. Oikos 72(1): 61-66.

Peters JL, Zhuravlev YN, Fefelov I, Humphries EM, Omland KE. 2008. Multilocus phylogeography of a holarctic duck: colonization of North America from Eurasia by gadwall (*Anas strepera*). Evolution 62(6): 1469-83.

Pikaard CS. 2001. Genomic change and gene silencing in polyploids. Tren Genet 17(12): 675-77.

Pielou EC. 1991. After the ice age: the return of life to glaciated North America. Chicago: The University of Chicago Press; 366 p.

Polunin N. 1959. Circumpolar arctic flora. Oxford: Clarendon Press; 514 p.

Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, *et al*. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636-9.

Robinson CH, Wookey PA, Lee JA, Callaghan TV, Press MC. 1998. Plant community responses to simulated environmental change at a high arctic polar semi-desert. Ecology 79(3): 856-66.

Rouard M, Guignon V, Aluome C, Laporte M-A, Droc G, Walde C, Zmasek CM, Périn C, Conte, MG. 2010. GreenPhylDB v 2.0: comparative and functional genomics in plants. Nucl Acids Res doi:10.1093/nar/gkq811.

Russell LB, Russell WL. 1996. Spontaneous mutations recovered as mosaics in the mouse specific-locus test. PNAS 93: 13072-77.

Salzberg SL, Delcher AL, Kasif S, White O. 1998. Microbial gene identification using interpolated Markov models. Nucl Acids Res 26(2): 544-48.

Schaal BA, Hayworth DA, Olsen KM, Rauscher JT, Smith WA. 1998. Phylogeographic studies in plants: problems and prospects. Mol Ecol 7: 465-74.

Shuman B, Webb III T, Bartlein P, Williams JW. 2002. The anatomy of a climatic oscillation: vegetation change in eastern North America during the Younger Dryas chronozone. Quat Sci Rev 21: 1777-91.

Simakov KV, editor. 2002. Quaternary paleogeography of Beringia. Magadan: NESC FEB RAS; 451 p.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. J Mol Biol 147(1): 195-7.

Stewart JR, Lister AM, Barnes I, Dalén L. 2010. Refugia revisited: individualistic responses of species in space and time. Proc R Soc B 277: 661-71.

Storm CEV, Sonnhammer ELL. 2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics 18: 92-9.

Taberlet P, Fumagalli L, Wust-Saucy AG, Cosson JF. 1998. Comparative phylogeography and postglacial colonization routes in Europe. Molec Ecol 7: 453-64.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol doi 10: 1093/molbev/msr123.

Tarasov PE, Volkova VS, Webb III T, Guiot J, Andreev AA, Bezuko LG, Bezuko TV, Bykova GV, Dorofeyuk NI, Kvavadze EV, *et al*. 2000. Last glacial maximum biomes reconstructed from pollen and plant macrofossil data from northern Eurasia. J Biogeog 27: 609-20.

Travis J, Futuyama DJ. 1993. Global change: lessons from and for evolutionary biology. Biotic interactions and global change. Kareiva PM, Kingsolver JG, Huey RB eds. Sunderland: Sinauer; 480 p.

van der Heijden RT, Snel B, van Noort V, Huynen MA. 2007. Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinformatics 8: 83.

Walker M. 1995. Patterns and causes of arctic plant community diversity. Arctic and alpine biodiversity: patterns, causes, and ecosystem consequences. Chapin III FS, Körner C, eds. Berlin: Springer, 332 p.

Walker MD, Wahren CH, Hollister RD, Henry GHR, Ahlquist LE, Alatalo JM, Bret-Harte MS, Calef MP, Callaghan TV, Carroll AB, *et al*. 2005. Plant community responses to experimental warming across the tundra biome. PNAS 103(5): 1342-46.

Wall PK, Leebens-Mack J, Müller KF, Field D, Altman NS, dePamphilis CW. 2008.

PlantTribes: a gene and gene family resource for comparative genomics in plants. Nucl Acids Res 36: D970-76.

Wang Y, Hey J. 2010. Estimating divergence parameters with small samples from a large number of loci. Genetics 184: 363-79.

West RG. 1980. Pleistocene forest history in East Anglia. New Phytol 85: 571-622.

Wiener J. 1994. The beak of the finch. New York: Vintage; 352 p.

Williams JT, Shuman BN, Webb III T, Bartlein PJ, Leduc PL. 2004. Late-Quaternary vegetation dynamics in North America: scaling from taxa to biomes. Ecol Mono 74: 309-34.

Willis CG, Ruhfel B, Primack RB, Miller-Rushing AJ, Losos JB, Davis CC. 2010. Favorable climate change response explains non-native species' success in Thoreau's woods. PloS ONE 5: e8878.

Wookey PA, Robinson CH, Parsons AN, Welker JM, Press MC, Callaghan TV, Lee JA. 1995. Environmental constraints on the growth, photosynthesis and reproductive development of Dryas octopetala at a high arctic polar semi-desert, Svalbard. Oecologia 102: 478-89.

Xiang Q-Y(J), Thorne JL, Seo T-K, Zhang W, Thomas DT, Ricklefs RE. 2008. Rates of nucleotide substitution in Cornaceae (Cornales)-- pattern of variation and underlying causal factors. Mol Phyl Evol 49: 327-42.

Zazula GD, Froese DG, Elias SA, Kuzmina S, Matthewes RW. In press. Arctic ground squirrels of the mammoth-steppe; paleoecology of middens from the last glaciation, Yukon Territory, Canada. Quat Sci Rev.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821-829.

Zmasek CM, Eddy SR. 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics 8: 83.

Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol 18(6): 292-8.

Zurawski G, Clegg MT. 1987. Evolution of higher-plant chloroplast DNA-encoded genes: implications for structure-function and phylogenetic studies. Annu Rev Plant Physiol 38: 391-418.

APPENDIX

1. Complete mitochondrial and chloroplast genomes used to identify M and CP sequences

Mitochondrial

Name	NCBI Accession No.	Order
Arabidopsis thaliana	49256807	Brassicales
Brassica napus	37591045	Brassicales
Beta vulgaris subsp. vulgaris	47118321	Caryophyllales
Marchantia polymorpha	786182	Marchantiales
Nicotiana tabacum	56806513	Solanales
Oryza sativa (indica cultivar-group) isolate 93-11	74100068	Poales
Physcomitrella patens	90991378	Funariales
Sorghum bicolor	114309646	Poales
Silene latifolia	301338014	Caryophyllales
Triticum aestivum	78675232	Poales
Tripsacum dactyloides cultivar Pete	114432085	Poales
Zea mays strain NB	40794996	Poales

Chloroplast

Name	NCBI Accession No.	Order
Arabidopsis thaliana	7525012	Brassicales
Buxus microphylla	149390519	Basal tricolpates
Coffea arabica	116617087	Gentianales
Fagopyrum esculentum subsp. ancestrale	166065336	Caryophyllales
Oryza sativa Japonica Group	11466763	Poales
Populus alba	110227059	Malpighiales
Piper cenocladum	115605001	Piperales
Pinus thunbergii	7524593	Pinales
Ranunculus macranthus	122893969	Ranunculales
Vitis vinifera	91983971	Vitales

2. list_querynames_and_compliments.pl

```
#!usr/bin/perl
open(BLASTOUT, "$ARGV[0]");
@blastlines = <BLASTOUT>;
close BLASTOUT;
open( FILE, '>queries_that_hit.txt');
open( FILE2, '>query_compliments_rc.txt');
$blastlines = join("", @blastlines);
@queryblocks = split('Query=', $blastlines);
shift @queryblocks;
foreach $block (@queryblocks) {
     ($qth) = ($block =~ /(^.*?\n)/);
     $qth =~ s/\n//g;
     $qth =~ s/\s//g;
     @hsp = split('>', $block);
     shift @hsp;
     if($hsp[0]) {
          ($qc) = ($hsp[0] =~ /(^.*?\n)/);
          $qc =~ s/\s//g;
          $qc =~ s/\n//g;
          @id = split("/", $hsp[0]);
          @id2 = split(" ", $id[1]);
          $numcol = $id2[0];
          $numcol =~ s/\n//;
          if ($numcol > 58) {
               if ($hsp[0] =~ /Plus\/Plus/) {
                    print FILE ">","$qth\n";
                    print FILE2 ">", "$qc\n";
                    }
               else {
                    print FILE ">","$qth\n";
                    print FILE2 ">", "$qc","_rc","\n";
                    }
               }
          }
     }
exit;
```

3. restore_and_align_CGpairs.pl

```
#! /usr/bin/perl
$i = "ARC99.2_contigs.fa";
open(ARCCNTIGS, $i);
@c = <ARCCNTIGS>;
close ARCCNTIGS;
```

```
$c = join("", @c);
@contigs = split('>', $c);
shift @contigs;
%arc;
foreach $element (@contigs) {
     ($name) = ($element =~ /(^.*?\n)/);
     $element =~ s/$name//;
     $element =~ s/\n//g;
     $name =~ s/ARC99/>ARC99/;
     $name =~ s/\n//;
     $arc{$name} = $element;
     }
$ii = "ALP99.2_contigs.fa";
open(ALPCNTIGS, $ii);
@cc = <ALPCNTIGS>;
close ALPCNTIGS;
$cc = join("", @cc);
@contigz = split('>', $cc);
shift @contigz;
%alp;
foreach $elment (@contigz) {
     ($nam) = ($elment =~ /(^.*?\n)/);
     $elment =~ s/$nam//;
     $elment =~ s/\n//g;
     $nam =~ s/ALP99/>ALP99/;
     $nam =~ s/\n//;
     $alp{$nam} = $elment;
     }
open (FILE, "$ARGV[0]");
chomp $ARGV[0];
$ARGV[0] =~ s/CGpairs//g;
$ARGV[0] =~ s/_//g;
$ARGV[0] =~ s/.txt//g;
$id = $ARGV[0];
`mkdir CG_pairs_$id`;
@input = <FILE>;
close FILE;
$input = join ('', @input);
@echpr = split ('\n', $input);
$count = 1;
foreach $pair (@echpr) {
     @arcalp = split ('>', $pair);
     shift @arcalp;
     $arcalp[0] =~ s/A/>A/g;
     $arcalp[1] =~ s/A/>A/g;
     $alpwith_rc = $arcalp[1];
```

```
     $arcalp[1] =~ s/_rc//g;
     open (FILE, '>unaligned.txt');
     if ($alpwith_rc =~ m/_rc/) {
          $revcom = reverse $alp{$arcalp[1]};
          $revcom =~ tr/ACGTacgt/TGCAtgca/;
          print FILE "$arcalp[0]\n", "$arc{$arcalp[0]}",
          "\n\n", "$alpwith_rc", "\n", "$revcom";
          close FILE;
          `./muscle3.8.31_i86darwin32 -in unaligned.txt
          -out ./CG_pairs_$id/$id\_pair_$count.fasta`;
          ++$count;
          }
     else {
          print FILE "$arcalp[0]\n", "$arc{$arcalp[0]}",
          "\n\n", "$arcalp[1]\n", "$alp{$arcalp[1]}";
          close FILE;
          `./muscle3.8.31_i86darwin32 -in unaligned.txt
          -out ./CG_pairs_$id/$id\_pair_$count.fasta  `;
          ++$count;
          }
     }
exit;
```

4. run_usearch_serially_FINAL.pl

```
#!/usr/bin/perl
foreach my $file (`ls nr*`) {
    chop($file);
    if ( $file =~ /lite/){
     `./usearch4.1.93_i86darwin32 --query
     /Users/danielbronny/rhs/clusters/ALL99.2_contigs_seeds.fa
     --db $file --maxlen 35000 --minlen 5 --userout
     /Users/danielbronny/rhs/output_of_nr/output_of_$file.uc
     --userfields query+target+ql+qs+frame+qrow+trow
     --maxrejects 0 --maxaccepts 2 --evalue 1e-6`;
     }
    }
exit;
```

5. make_CG_pairs.pl

```
#!usr/bin/perl
open (FILE1, "./blastanalysis/queries_that_hit.txt");
open (FILE2, "./blastanalysis/query_compliments_rc.txt");
@temp1 = <FILE1>;
@temp2 = <FILE2>;
```

```
close FILE1;
close FILE2;
$temp1 = join('', @temp1);
$temp2 = join('', @temp2);
@qwhits = split ('\n', $temp1);
@qcomps = split ('\n', $temp2);
@cgpairs{@qwhits}=@qcomps;
open(P, 'P_names.txt');
open(C, 'C_names.txt');
open(M, 'M_names.txt');
open(U, 'C_or_M_names.txt');
open(RPTS, 'repeats.txt');
@p = <P>;
@c = <C>;
@m = <M>;
@u = <U>;
@rpts = <RPTS>;
close P;
close C;
close M;
close U;
close RPTS;
$p = join('', @p);
$c = join('', @c);
$m = join('', @m);
$u = join('', @u);
$rpts = join('', @rpts);
open(ERR, '>CGpairs_error_.txt');
open(PPR, '>CGpairs_P_.txt');
open(CPR, '>CGpairs_C_.txt');
open(MPR, '>CGpairs_M_.txt');
open(UPR, '>CGpairs_U_.txt');
open(NCPR, '>CGpairs_NC_.txt');
while (($key, $value) = each %cgpairs) {
     if($c =~ /$key|$value/) {
          if($m =~ /$key|$value/ | $p =~ /$key|$value/ | $u
          =~ /$key|$value/ | $rpts =~ /$key|$value/) {
               print ERR "$key","$value\n";
               }
          else {
               print CPR "$key","$value\n";
               }
          }
     elsif($m =~ /$key|$value/) {
          if($c =~ /$key|$value/ | $p =~ /$key|$value/ | $u
               =~ /$key|$value/ | $rpts =~ /$key|$value/) {
```

```
                print ERR "$key","$value\n";
                }
           else {
                print MPR "$key","$value\n";
                }
           }
     elsif($u =~ /$key|$value/) {
           if($c =~ /$key|$value/ | $p =~ /$key|$value/ | $m
           =~ /$key|$value/ | $rpts =~ /$key|$value/) {
                print ERR "$key","$value\n";
                }
           else {
                print UPR "$key","$value\n";
                }
           }
     elsif($p =~ /$key|$value/) {
           if($m =~ /$key|$value/ | $c =~ /$key|$value/ | $u
           =~ /$key|$value/ | $rpts =~ /$key|$value/) {
                print ERR "$key","$value\n";
                }
           else {
                print PPR "$key","$value\n";
                }
           }
     unless($m =~ /$key|$value/ | $c =~ /$key|$value/ | $u
           =~ /$key|$value/ | $rpts =~ /$key|$value/ | $p =~ /
           $key|$value/) {
           print NCPR "$key","$value\n";
           }
     }
close ERR;
close PPR;
close CPR;
close MPR;
close UPR;
close NCPR;
exit;
```

6. find_identical_ppairs.pl

```
#!/usr/bin/perl
open (NODIFF, '>no_polymorphisms.txt');
foreach $file (`ls`) {
    chop($file);
    if ( $file =~ /P_pair_[0-9]*\.fasta/) {
          open ( FILE, "$file");
```

```
@contents = <FILE>;
close FILE;
$contents = join ('',@contents);
@eachtig = split ('>',$contents);
shift @eachtig;
$arc = $eachtig[0];
$alp = $eachtig[1];
$arc =~ s/A/>A/;
$namechk = $arc;
$namechk =~ s/\n.*//g;
$alp =~ s/A/>A/;
$eachtig[0] =~ s/(^.*\n)//;
$eachtig[1] =~ s/(^.*\n)//;
$arcstring = $eachtig[0];
$alpstring = $eachtig[1];
$arcstring =~ s/\n//g;
$alpstring =~ s/\n//g;
@posarc = split ('', $arcstring);
@posalp = split ('', $alpstring);
$colcnt = 0;
$match = 0;
$miss = 0;
$gap = 0;
foreach $column (@posarc) {
     if ($column eq "-") {
          ++$gap;
          ++$colcnt;
          }
     elsif ($posalp[$colcnt] eq "-") {
          ++$gap;
          ++$colcnt;
          }
     elsif ($column eq $posalp[$colcnt]) {
          ++$match;
          ++$colcnt;
          }
     else {
          ++$miss;
          ++$colcnt;
          }
          }
     $total = $match + $miss;
if ($total eq $match) {
     ++$nochange;
     print NODIFF "$namechk", "\n";
     }
```

```
     print "\n$file","\n";
     print "The number of columns was $colcnt\n";
     print "The number of gaps was $gap\n";
     print "The number of match columns was $match\n";
     print "The number of mismatch columns was $miss\n";
     print "The number of eligible columns was $total\n";
     }
     }
print "\nThere are $nochange \"identical\" protein pairs\n\n";
exit;
```

7. run_SNAP_on_all.pl

```
#! /usr/local/bin/perl
open(FILE, "ALL_frame_output.uc");
@lines = <FILE>;
close FILE;
foreach $line (@lines) {
     @tabs = split("\t", $line);
     unless ($tabs[0] =~ /query/){
          $framehash{$tabs[0]} = $tabs[4];
          }
     }
while (($key, $value)=each %framehash) {
     push (@keylist,$key);
     }
$framekeys = join('', @keylist);
open(POLYCHK, "no_polymorphisms.txt");
@chk = <POLYCHK>;
$polychk = join ('', @chk);
close POLYCHK;
open (FILE, "P_pair_ALL.fa");
@readta = <FILE>;
close FILE;
$readin = join("",@readta);
@readin = split(">AR", $readin);
shift @readin;
$count = 0;
$tpp = 0;
$nopoly = 0;
open (ERR, '>not_appearing_in_framesearch.txt');
foreach $elm (@readin) {
     open (FILE2, '>onepair.fasta');
     $elm =~ s/C991/>ARC991/;
     ($name) = ($elm =~ /(^.*?\n)/);
```

```
$name =~ s/\s//g;
if ($polychk =~ /$name/) {
     ++$nopoly;
     }
else{
     $name =~ s/>//g;
     if ($framekeys =~ /$name/) {
          $arcid = $name;
          @arcid = split ("_", $arcid);
          $arcidfinal = "R".$arcid[2];
          $elm =~ s/>AR.*\n/$arcidfinal\t/g;
          ($name2) = ($elm =~ /(>.*?\n)/);
          $name2 =~ s/>//g;
          $name2 =~ s/\n//g;
          $alpid = $name2;
          @alpid = split ("_", $alpid);
          $alpidfinal = "L".$alpid[2];
          $elm =~ s/>AL.*\n/$alpidfinal\t/g;
          $elm =~ s/\n//g;
          $elm =~ s/A(L)/A\nL/;
          $elm =~ s/T(L)/T\nL/;
          $elm =~ s/G(L)/G\nL/;
          $elm =~ s/C(L)/C\nL/;
          $elm =~ s/-(L)/-\nL/;
          $elm =~ s/$/\n/;
          $elm2 = $elm;
          $elm2 =~ s/(L|R).*\t//g;
          @thetwolines = split ("\n",$elm2);
          $seq1a = $thetwolines[0];
          $seq1b = $thetwolines[1];
          $frame = $framehash{$name};
          $seq1a_rc = reverse $seq1a;
          $seq1a_rc =~ tr/ACGTacgt/TGCAtgca/;
          $seq1b_rc = reverse $seq1b;
          $seq1b_rc =~ tr/ACGTacgt/TGCAtgca/;
          if ($frame =~ m/(-)/) {
               if ($frame =~ m/1/) {
                    print FILE2
                    "$arcidfinal","\t","$seq1a_rc","\n";
                    print FILE2
                    "$alpidfinal","\t","$seq1b_rc","\n";
                    `perl SNAP.pl onepair.fasta`;
                    }
               if ($frame =~ m/2/) {
                    print FILE2
                    "$arcidfinal","\t","--","$seq1a_rc","
```

```
                    \n";
                    print FILE2
                    "$alpidfinal","\t","--","$seq1b_rc","
                    \n";
                    `perl SNAP.pl onepair.fasta`;
                    }
               if ($frame =~ m/3/) {
                    print FILE2
                    "$arcidfinal","\t","-","$seq1a_rc","\
                    n";
                    print FILE2
                    "$alpidfinal","\t","-","$seq1b_rc","\
                    n";
                    `perl SNAP.pl onepair.fasta`;
                    }
               }
          else {
               if ($frame =~ m/1/) {
                    print FILE2
                    "$arcidfinal","\t","$seq1a","\n";
                    print FILE2
                    "$alpidfinal","\t","$seq1b","\n";
                    `perl SNAP.pl onepair.fasta`;
                    }
               if ($frame =~ m/2/) {
                    print FILE2
                    "$arcidfinal","\t","--","$seq1a","\n"
                    ;
                    print FILE2
                    "$alpidfinal","\t","--","$seq1b","\n"
                    ;
                    `perl SNAP.pl onepair.fasta`;
                    }
               if ($frame =~ m/3/) {
                    print FILE2
                    "$arcidfinal","\t","-","$seq1a","\n";
                    print FILE2
                    "$alpidfinal","\t","-","$seq1b","\n";
                    `perl SNAP.pl onepair.fasta`;
                    }
               }
          ++$count;
          }
     else {
          print ERR ">", $name, "\n";
          ++$tpp;
```

```
            }
        }
    }
close FILE2;
$wpur=0;
$wpos=0;
$spur=0;
$spos=0;
$ntrl=0;
open (FILE3, '>dsdnlist.txt');
open (FILE4, '>SNAP_summary.txt');
foreach my $file (`ls summary.*`) {
    chop($file);
    open (FILE, $file);
    @summary = <FILE>;
    @secondline = split(" ",$summary[1]);
    if ($secondline[8] =~ /0\.00/) {
          print FILE3 $secondline[2],"\t",
          $secondline[3],"\t","W\t","pos\t",
          $secondline[12],"\t",$secondline[4],"\t",
          $secondline[5],"\n";
          ++$wpos;
    }
    elsif ($secondline[9] =~ /0\.00/) {
          print FILE3 $secondline[2],"\t",
          $secondline[3],"\t","W\t","pur\t",
          $secondline[12],"\t",$secondline[4],"\t",
          $secondline[5],"\n";
          ++$wpur;
    }
    elsif ($secondline[12] gt 1) {
          print FILE3 $secondline[2],"\t",
          $secondline[3],"\t","S\t","pur\t",
          $secondline[12],"\n";
          ++$spur;
          }
     elsif ($secondline[12] eq 1) {
          print FILE3 $secondline[2],"\t",
          $secondline[3],"\t","NA\t","ntrl\t",
          $secondline[12],"\n";
          ++$ntrl;
          }
     else{
          print FILE3 $secondline[2],"\t",
          $secondline[3],"\t","S\t","pos\t",
          $secondline[12],"\n";
```

```
          ++$spos;
          }
     }
close FILE3;
print FILE4 "The total number of P pairs in the input file was
", scalar @readin, "\n";
print FILE4 "The total number of P pairs that had
polymorphisms, frame information, and were run in SNAP is
$count.\n";
print FILE4 "$tpp P pairs did not appear in the output of
framesearch.\n";
print FILE4 "$nopoly P pairs were not run because they lacked
polymorphism.\n";
print FILE4 "$wpos were categorized as weak-pos\n";
print FILE4 "$wpur were categorized as weak-pur\n";
print FILE4    "$spos were categorized as strong-pos\n";
print FILE4 "$spur were categorized as strong-pur\n";
`mkdir SNAPrawout`;
`mv summary.* SNAPrawout/`;
`mv dsdist.* SNAPrawout/`;
`mv dndist.* SNAPrawout/`;
`mv codons.* SNAPrawout/`;
`mv background.* SNAPrawout/`;
exit;
```