Phylogeography of the plants disjunct between Western North America and the Great Lakes Region: patterns, case studies, and implications for future comparative research

By

Chloe P. Drummond

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Botany)

at the

UNIVERSITY OF WISCONSIN-MADISON

Date of final oral examination: 07/09/2018

The dissertation is approved by the following members of the Final Oral Committee:

Kenneth J. Sytsma, Professor, Botany
David A. Baum, Professor, Botany
Thomas J. Givnish, Professor, Botany
Sara C. Hotchkiss, Professor, Botany
Cécile Ané, Professor, Botany and Statistics

ACKNOWLEDGEMENTS

I want to thank my PhD advisor, Kenneth J. Sytsma, for his guidance and mentorship. My scholarship, teaching, and enthusiasm for botany have been greatly inspired and strengthened by my experience conducting research in his lab and teaching in his classroom. Dr. Sytsma's thoughtfulness and attention to my academic success helped me find my path through graduate school and foster my own sense of meaningful participation in the botanical evolution community. For his support and patience, I am immensely grateful.

I would like to thank the members of my PhD committee: David A. Baum, Thomas J. Givnish, Sara C. Hotchkiss, and Cécile Ané. By working with them at each stage of my doctoral thesis, I was challenged to shape questions and choose analytical approaches with consideration for multiple disciplines, which broadened my botanical lens. I am thankful for their willingness and eagerness to provide me with advice and support during my doctoral training, and I look forward to continuing this journey with them as colleagues.

Embarking on a career in plant systematics would not have happened without the mentorship of my undergraduate research advisor, Michael J. Moore, at Oberlin College. Dr. Moore first inspired me with his passion for gypsum endemism, and I conducted research in his lab only to find that I, too, was eager to investigate evolutionary histories to make sense of botanical patterns. He has been cheering me on throughout my graduate career, and his excitement for my progress and successes has been uplifting.

Last, but certainly not least, I thank my family and friends for their unending support. My dad, who came on multiple, week-long field excursions to the outskirts of North America and the United States, barely bat an eyelid when asked to trek through the spiny Devil's-club or hike through bear country to collect samples. My mom and Kathy, who read through many grant

proposals, continued to remind me to take a step back and wrestle with the big picture and the larger meaning of my work. My *unni*, Jocelyn, who has been the best older sister and a compass-guide throughout my life, helped me navigate the ups and downs of graduate school life. My grandparents who are no longer here, instilled in me a core value of education, and my pursuit of a PhD was borne out of that family value. My friends in Madison, Chicago, Boston, and San Diego, who were there for major life events and life transitions, kept me grounded and focused.

I did not attain this degree on my own, and I am grateful for the mentors, friends, and colleagues who have supported me along the way.

TABLE OF CONTENTS

ABSTRACT

More than 100 vascular plant taxa have a discontinuous geographic distribution between western North America and the Great Lakes Region. Each of these species is distributed in a pattern characterized by occurrence from Alaska south through the western Rocky Mountains, with reappearance in the Great Lakes Region, where they are often locally endangered. Some of these species have additional occurrences in the Black Hills of South Dakota. In this thesis I provide an updated review of this disjunct biogeographic pattern and presents three in-depth case studies of this geographic distribution: *Rubus parviflorus* Nutt. (Thimbleberry, Rosaceae), *Oplopanax horridus* (Sm.) Miq. (Devil's-club, Araliaceae), and *Aconitum columbianum* Nutt. (Columbian monkshood, Ranunculaceae). For each of these case studies I present 1) how old these species are and how their age supports or refutes different historical migration hypotheses that are centered around Pleistocene geological events, 2) the historical geographic range of these species and how they came to occupy their current distribution, and 3) population dynamics, such as bottlenecks and gene flow, that contribute to the genetic diversity of the Great Lakes Region populations. These analyses leverage geo-referenced specimen data and next-generation sequencing. They uncover different underlying phylogeographic histories between the case studies, indicating that this shared biogeographic pattern is a more complex, pseudoparallel pattern. The results have implications for future research on the western North America-Great Lakes Region disjunct plants as well as future conservation efforts in the Great Lakes Region.

CHAPTER 1: Review of the phylogeography of disjunct taxa with emphasis on the Western North American-Great Lakes Region disjunct distribution: questions, methods, and setting the stage for future research on this system.

ABSTRACT

Disjunct geographic distributions of plants are of interest to botanists because they pose floristic, biogeographic, and evolutionary questions. These questions can be answered with methods that incorporate phylogenetics, timing, paleoecology, geology, population genetics, demographic modeling, and niche modeling. The recurring North American disjunct distribution between western North America (WNA) and the Great Lakes Region (GLR) was last reviewed in 1981 (Marquis and Voss 1981). We revisit it here to discuss new analytical approaches and lines of evidence that can be used to illuminate the underlying mechanisms of this distribution. To consider specific hypotheses that have been put forward, we generated an updated list of the taxa showing this distribution using regional species lists and relying on multiple sources of distribution maps to bin taxa into six broadly-defined disjunct categories. We tested phylogeographic congruence in this system by projecting suitable niche space back to the Last Glacial Maximum (LGM) for 24 of the species in our list. We used these niche models to test a null hypothesis that all species persisted in a Midwestern glacial refugium during the LGM. Our final list of taxa contained 114 entries across generic, species and subspecific ranks. These projected historical distributions indicated that taxa within the same disjunct categories had different ranges of projected suitable niche space at the LGM, providing support for differing, or pseudo-parallel biogeographic histories within this system. In addition, all but two models rejected the hypothesis that the species were able to persist in a Midwestern glacial refugium. We

hope that this review, updated taxon list, and discussion of pseudo-parallel distributions in the WNA-GLR disjunct system will be used as a starting point for future research on taxa of this distribution.

INTRODUCTION

Biogeographic distributions are critical for understanding past, present, and future biodiversity. They work in tandem with evolutionary, ecological, and geological processes, resulting in the variation we observe. Two perspectives on distributions come to the forefront in any discussion of plant biogeography: the where, how, and when of a singular taxon's distribution and what it interacts with, and the who, how and when of the taxa of a given location, and how they interact. The former, it could be argued, has become a common approach in recent biogeographic and phylogenetic research. Researchers are able to claim expertise on a particular taxonomic group of interest and are concerned with the holistic context of that group. The latter requires a broader, pattern-based approach and has previously existed in the realm of the naturalist who conducts floristic studies, describing and cataloging regional observations. In the case of a recurring biogeographic pattern, however, both of these perspectives are brought together to ask fundamental questions of biodiversity: what are the biogeographical histories of the taxa, how and when did the taxa get there, what determines the maintenance of the distribution, does the distribution help explain variation in the taxa, what parallels are there (if any) among the taxa, and can ecologically relevant traits or relatedness help explain which taxa have this distribution? Biologists have been interested in these biogeographic questions (Wen et al. 2013, Whitaker et al. 2013) because they attempt to uncover the forces that drive patterns of

biodiversity, they set the stage for examining evolution in context, and simply, they satisfy a curiosity about the plants we see day-to-day in our geographic contexts.

Because the field of biogeography is broad, we will focus our review on the phylogeographic investigation of recurring disjunct patterns and discuss the questions that can be asked, and the analytical tools required to answer them. Discontinuous distributions are of particular interest because they provide insight into the larger forces that impact biodiversity, such as glacial cycles, continental drift, and climate change (e.g. Hewitt 1996, Friis et al. 2016), they reveal biological phenomena, such as long-distance dispersal (e.g. Higgins and Richardson 1999, Cain et al. 2000, Givnish et al. 2004), and they serve as models of genetic divergence and speciation (e.g. Hewitt 1996, Pettengill and Moeller 2012). There are classic, striking examples of intra- and intercontinental disjunct distributions that are shared across multiple taxa, and some have been studied recently in the frameworks of historical biogeography (e.g. Givnish and Renner 2004, Wen et al. 2009, Samonds et al. 2012) and phylogeography (e.g. Soltis et al. 1997, Taberlet et al. 1998, Brunsfeld et al. 2001, Soltis et al. 2006, Schmitt 2007, Stubbs et al. 2018). These studies have focused on individual taxa, have looked at patterns across taxa, or have combined these approaches. Recent papers have argued in favor of a combined approach, which integrates comparison of co-occurring taxa with analysis of the traits of each taxon to better predict how each taxon fits the shared geographic pattern (Papadopoulou and Knowles 2016). This approach allows the ecological and evolutionary questions and inferences of each taxon to complement that of another, while still recognizing the differences among taxa that contribute to their individual biogeographic histories. This approach, however, is not an easy task because it requires investment in both narrow and broad scopes. However, it is crucial for studies of a taxon with a shared, exceptional distribution to also consider a regional, place-based perspective. This

is not only helpful for fostering discussion of the distribution among leading experts on particular taxa, but it also cultivates floristic knowledge, which is important for any future work on biodiversity.

Our review focuses on phylogeography, which is the study of geographic distributions at the interface of systematics/phylogenetics and population genetics (Avise et al. 1987, Edwards et al. 2016). This field unites a powerful suite of tools to answer questions at the population level, such as population history, timing, evolution, structure, and genetic diversity, that are often brought together to uncover the mechanisms that underlie geographic distributions and impact biodiversity. Recent advances in genomic sequencing technology and demographic modeling have created a surge in phylogeographic research (Soltis et al. 2006, Hickerson et al. 2010, Bowen et al. 2014, Edwards et al. 2016). With genomic sequencing that produces hundreds of loci or thousands of single nucleotide polymorphisms (SNPs) across the genome, new methods are being tested to finally tease apart long-standing phylogeographic problems, such as parallel distributions (Soltis et al. 2006), rampant hybridization (Hipp et al. 2018) and the threat of climate change on small, isolated populations (Devlin-Durante and Baums 2017). Here we use the vascular plants that are disjunct between western North America (WNA) and the Great Lakes Region (GLR) to illustrate the potential of phylogeographic research and review the phylogeographic methods involved in teasing apart disjunct distributions. We summarize some of the broader patterns found in the WNA-GLR disjunct taxa to aid future research on this system.

REVIEW

Overview of the WNA-GLR disjuncts

The taxa disjunct between western North America (WNA) and the Great Lakes Region (GLR) pose floristic, biogeographical, and evolutionary problems. A simplified description of their range is from Alaska south through the western Rocky Mountains, reappearing in the GLR where they often occur in smaller, isolated populations (Marquis and Voss 1981). The floristic problem is one focused on the GLR. Are there, or were there characteristics of this region that made it suitable to accumulate multiple, predominantly western species? Will these disjuncts be a long-term contribution to the flora of the GLR in the face of environmental change? The biogeographical problem is one focused on how and when these taxa came to occupy their current distribution. For example, the absence of west-east migration patterns in animals means that animal dispersal is less likely to explain this distribution, and the presence of glacial Lake Agassiz that covered parts of Ontario, Manitoba, Saskatchewan, Minnesota, and the Dakotas ~13,000 years ago may have hindered, or aided, dispersal along the receding glacier during the Wisconsin glaciation (Marquis and Voss 1981, Dyke and Prest 1987). So, how and when did these taxa occupy the GLR? Finally, the evolutionary problem is one of divergence, adaptation, and introgression. Is isolation reducing gene flow or causing local adaptation? Are species boundaries maintained when sister species have overlapping ranges in the GLR? These three problems fall within the realm of phylogeography. They are not mutually exclusive, but we will discuss them separately.

Regional influences on the GLR flora: The GLR contains geological and ecological features that might explain its suitability for western species outside of their typical range. These features are most prominent in the "Driftless Area" of Minnesota, Wisconsin, Iowa and Illinois. This region was not covered with ice during the glacial cycles of the Pleistocene, and regions on either side were glaciated, but not at the same time (Alden 1904, Chamberlin and Salisbury 1906,

Holliday et al. 2002). The Driftless Area is thought to have been covered with tundra-like habitat during the Last Glacial Maximum (LGM) (Baker et al. 1980, Delcourt and Delcourt 1981). However, *Quercus* pollen dating to the LGM has also been found in trace amounts in the Driftless Area, which has suggested the possibility of deciduous forest stands in this region during the LGM (Jackson and Overpeck 2000). In addition, the Driftless Area features algific talus slopes, which are home to a few rare and endangered arctic and periglacial, relict disjunct species (Smith 1949, Pusateri 1993). These slopes consist of fractured limestone in which ice freezes during the winter and cool air passes through, creating a cold microclimate year-round. Given the absence of glacial cover, the Driftless Area has been suggested as a glacial refugium for plants and animals (Fernald 1925, Rowe et al. 2004, Lee-Yaw et al. 2008, Li et al. 2013) and might have been the only remaining cool, mesic habitat for periglacial species towards the end of the Pleistocene as ice was retreating.

Another regional influence on the GLR flora is climate change. Recent studies on the flora of the Midwest, and Wisconsin in particular, have shown that projected climate change will move suitable niche space northward for many plants (Ash et al. 2017, Spalink et al. in press). Whether plants with highly restricted ranges in the GLR will be able to persist in the future is a conservation concern, and species distribution models (SDMs) have been used to provide insight into future occupiable space (Vitt et al. 2010). The threat of extirpation in the GLR can also be increased with low levels of genetic diversity, and thus lower adaptive potential or higher rates of inbreeding, which can be caused by isolation and lack of gene flow, and small populations.

How and when did these taxa occupy the GLR: The primary competing hypotheses for the cause of geographic disjunction are vicariance and dispersal. The WNA-GLR disjunct taxa were last reviewed by Marquis and Voss (1981), where the authors discussed hypotheses of how

and when western North American floristic elements arrived in the GLR. Their hypotheses were centered around the Pleistocene glacial cycles because the extent of the ice sheets in North America reached lower than 40° N at their maximum (Ehlers and Gibbard 2007) and would have disrupted plant distributions. Marquis and Voss (1981) proposed: 1) the taxa were western-restricted pre-Pleistocene and advanced eastward along a receding ice-sheet, followed by intermediate local extirpation due to incoming competition and/or the warmer, dryer post-glacial conditions that restricted cool, moist conditions to the northern GLR and the Black Hills of South Dakota, 2) the taxa were widespread pre-Pleistocene and only survived the glacial cycles in the East in a Driftless Area refugium of Wisconsin and Iowa, followed by recolonization of newly unglaciated areas, and 3) wind-dispersed plants may have arrived via long-distance dispersal. They argue, based on the diverse dispersal mechanisms of these plants, that the species arrived in the GLR at different times and via different routes, not *en masse*. Pleistocene history and refugia hypotheses have been tested extensively (Taberlet et al. 1998, Eidesen 2007, Marr et al. 2008, Provan and Bennet 2008, Rebernig et al. 2010, Shafer et al. 2010, Schneeweiss and Schonswetter 2010, Poncet et al. 2013), and the hypotheses put forward by Marquis and Voss (1981) can also be tested using phylogeographic methods.

The role of divergence and introgression: The western-GLR disjuncts are a good study system for investigating speciation and introgression. Geographically isolated populations are often described as the most common model of speciation because of the inherent reduction in gene flow (e.g., Wiens et al. 2004). Without genetic experiments, the signature of divergence might also be found in the taxonomic ranks of the disjunct taxa. With enough divergence, we might expect to find some genera that have disjunct sister species that occupy western North America and the GLR. To avoid circular reasoning, interpretation of this as divergence would

require that the species rank was not applied based on geography. The isolation of the GLR populations could also be fostering local adaptation, and this can be tested with reciprocal transplant studies and fine-scale measurements of niche overlap. Finally, species boundaries for some of the taxa that are disjunct in the GLR are blurred where they have range overlap with their sister species (Iltis 1965). This makes the western-GLR disjuncts a potential system for investigating the genetics of hybrid incompatibility (i.e. Orr 1995), using models of reinforcement that have been put forward to describe the maintenance of species boundaries (Servedio and Noor 2003). We would not expect a universal pattern of either introgression or reinforcement for these disjunct taxa, and since both might be occurring in this system, each taxon should be investigated individually.

Methods for uncovering the biogeographical history of the WNA-GLR taxa

Marquis and Voss (1981) posed three competing hypotheses to explain the history of the WNA-GLR distribution: 1) peri-glacial dispersal from a western Pleistocene refugium, 2) vicariance of a widespread distribution and persistence in a Driftless Area refugium, and 3) long-distance dispersal to the GLR. Many studies in the past decade have looked at inter- or intraspecific disjunct distributions specifically tied to glacial refugia hypotheses (Bermingham and Moritz 1998, Catling 2009, Wang et al. 2009, Beatty and Provan 2010, 2014, Gugger et al. 2011, Escobar Garcia et al. 2012, DeChaine et al. 2013, Li and Wen 2013, Gavin et al. 2014). These studies have used a combination of phylogenetics, timing of population divergence, population genetic diversity, and historical suitable niche space to uncover the underlying mechanisms of disjunct distributions. We discuss the evidence from each type of analysis that would support each of the three hypotheses proposed by Marquis and Voss.

Phylogenetics and demographic modeling are two common methods for inferring migration. Ancestral areas can be estimated along a phylogeny using parsimony, Maximum Likelihood (ML), or Bayesian optimization (Joy et al. 2016). A phylogenetic analysis will reveal the ancestral range of the populations and will help determine the ancestral range of the GLR populations in particular. The ancestral range of all of the populations will not be able to refute any of the hypotheses, but it can provide insight into the demographic history of the species. For example, if Alaska served as an ancestral range, this would suggest a refugium bottleneck in the history of the species. Major areas in Alaska remained ice-free during the last glacial cycles (Brigham-Grette and Gualtieri 2003), and Beringia was proposed as a refugium for arctic plants by Eric Hultén in his vast study of arctic flora (1937). His theory has been supported by more recent evidence (Nimis et al. 1998), and the refugium bottleneck hypothesis could be tested with further with demographic modeling. The ancestral range of the GLR populations, however, might provide evidence for one of the migration hypotheses. For example, if the ancestral range of the GLR populations was in a previously glaciated region, such as British Columbia, this might require post-glacial dispersal because the British Columbia populations would not have existed. Based on timing, this evidence would refute the hypothesis of vicariance and restriction in a Driftless Area refugium. However, some phylogeographic and geological evidence has raised the possibility of ice-free refugia in parts of Alberta and British Columbia, as well as an ice-free corridor, where plant populations might have persisted during glacial periods (Shafer et al. 2010). This scenario would have to be ruled out with demographic modeling that tests for evidence of genetic bottlenecks in the British Columbia populations. In addition, an ancestral range for the GLR populations in Alaska or south of the glacial line would not be able to distinguish between the competing migration hypotheses, and alternative lines of evidence would

be required. Migration history can also be determined with demographic modeling using approaches such as Approximate Bayesian Computation (Beaumont 2010) or Extended Bayesian Skyline models (Ho and Shapiro 2011). These methods estimate gene flow and directly test scenarios of migration and refugial isolation (Rovito and Schoville 2017). Results of this analysis will show the ancestral range of the GLR populations, which can be interpreted the same way as the phylogenetic results. In addition, if this analysis shows evidence of bottleneck events in the GLR, this would support the vicariance hypothesis with restriction in a Driftless Area refugium or the hypothesis of long-distance dispersal.

A key factor that can nuance the migration hypotheses is time. Time is necessary for understanding the environmental context and the plausibility of migration scenarios. The pre-Pleistocene hypotheses are only appropriate if the taxon of interest is older than the Pleistocene. Thus, it is necessary to determine the age of the taxon in order to discuss these hypotheses. In addition, timing of population divergences will provide evidence for either vicariance or periglacial dispersal. Population ages younger than the Last Glacial Maximum (LGM) (~ 22,000-18,000 years ago; Martinson et al. 1987, Yokoyama et al. 2000), would point to a periglacial dispersal route. Older, pre-Pleistocene population ages would point to a vicariance event and restriction in a Driftless Area refugium. In addition, direct fossil evidence can be used in support of the vicariance scenario, as in the case of *Dryas drummondii*, for which there is fossil evidence in eastern North America that indicates a pre-Pleistocene widespread distribution (Miller and Thompson 1979). There are three main methods for incorporating time: fossil-calibrated chronograms, generation time- and mutation rate-calibrated coalescent models, and the fossil record. Fossil-calibrated chronograms estimate node ages in years, and these can be used to estimate the age of the node of interest. Coalescent trees estimate node ages in generations unless

population size and generation time are known. Both of these methods can be used to estimate the age of population divergences and they can leverage next-generation sequencing data, such as targeted sequence capture and reduced representation sequencing (Eaton et al. 2016, Harvey et al. 2016). In addition, there are coalescent models such as SNAPP (Bryant et al. 2012) and SVDQuartets (Chifman and Kubatko 2014) that can use SNP data from reduced representation genomic sequencing (i.e. Genotyping-by-sequencing and RAD-Seq) to summarize individual gene trees and account for incomplete lineage sorting (ILS). The fossil record can provide support for the presence of a taxon at a particular place and time. However, because the record is incomplete, fossils should be interpreted as evidence of the presence and not the absence of a taxon.

Additional evidence to tease apart these hypotheses comes from population genetics. Population genetics can reveal genetic bottlenecks in outcrossing species (Cornuet and Luikart 1996). Evidence of a genetic bottleneck in the GLR populations compared to the western North American populations would indicate either a long-distance dispersal founder event, or restriction in a glacial refugium in the Driftless Area. Microsatellite genetic markers have been used to capture intra-specific variation for population genetics (Balloux and Lugon-Moulin 2002, Selkoe and Toonen 2006). However, reduced-representation genomic sequencing, which produces thousands of anonymous SNPs across the genome, is now being harnessed for population genetic applications (Davey and Blaxter 2010, Davey et al. 2011, Narum et al. 2013). Bottleneck events (i.e. population restriction followed by population expansion) can be detected from the allele frequency spectrum (AFS) or by calculating Tajima's D, which is a uni-dimensional representation of the AFS (Marth et al. 2004, Cornuet and Luikart 1996). The signature is heterozygosity excess, which occurs when alleles are lost faster than heterozygosity

is reduced (Cornuet and Luikart 1996). In addition, bottleneck events will result in loss of heterozygosity, which is difficult to recover from if there is a lack of subsequent gene flow into the population. Evidence of reduced heterozygosity and signatures of a bottleneck in the AFS for the GLR populations would refute the periglacial dispersal hypothesis. Support for long-distance dispersal or restriction in a glacial refugium would come from a young population divergence in the GLR or an older population divergence, respectively.

Finally, historical projections of species' niches can be used to determine whether there was suitable niche space in the GLR at the LGM. Species distribution models (SDMs) have been used to project species distributions into the past to analyze previous areas of habitation, migration history, and potential refugium areas (Nogues-Bravo 2009, Poncet et al. 2013, Gavin et al. 2014). These models use presence/absence or presence-only waypoint data to summarize the current niche of a species and use historical climate models to estimate where the niche was located in the past. Because SDMs estimate potential occupiable space, not realized niche space, they are better interpreted as evidence of the absence of a species rather than the presence of a species. If suitable niche space in the LGM exists in western North America, but does not include the GLR, this refutes hypotheses that include presence of that species in the GLR at the LGM. In contrast, if suitable niche space in the LGM does include the GLR, the species could have been there, but was not necessarily present. Due to the political, economic and social pressures for improving climate research globally, general circulation models (GCMs) for projecting climate through time have become more complex and fine-tuned (IPCC Reports, http://www.ipcc.ch/). In addition, SDM approaches, such as MaxEnt (Phillips et al. 2006) are becoming more statistically rigorous (e.g. Thuiller et al. 2009, Guisan et al. 2011, Calabrese et al. 2013). There are limits to these analyses due to limitations of the data, such as sampling bias of

locality data and resolution and accuracy of the environmental and climate data. In addition, there are limits due to the models, such as the underlying assumptions of niche conservatism and species occupying the full limits of their range (Radosavljevic and Anderson 2014, Martinez-Meyer et al. 2004, Araujo and Peterson 2012, Gavin et al. 2014, Dormann et al. 2012, D'Amen et al. 2013). Given these limitations, SDMs should be interpreted with care.

Comparative methods

In addition to addressing each taxon individually, we can assess whether co-distributed plants have a shared biogeographic history using comparative phylogeography. Recent papers have reviewed comparative phylogeography and associated analytical methods (Arbogast and Kenagy 2001, Avise 2000), and pushed the boundaries of comparative analysis (Papadopoulou and Knowles 2016, Satler and Carstens 2016). For co-distributed species, the most parsimonious assumption is that they have a shared biogeographic history (Avise 2000, Soltis et al. 2006). However, studies have shown that shared distributions can also result from different biogeographic histories, and this process of pseudo-congruence has been discussed in the literature for at least 30 years (e.g. Page 1990, Donoghue et al. 2001, Xiang and Soltis 2001, Donoghue and Moore 2003, Soltis et al. 2006, Stubbs et al. 2018). Congruence between co-distributed taxa is thus an assumption that should be tested, and methods are being developed to directly test parallel histories (Satler and Carstens 2016). Below, we discuss the geographic patterns of the WNA-GLR distribution and use historical SDMs to start a conversation on the congruence among taxa that share this distribution.

METHODS

Patterns of the WNA-GLR disjuncts

Papers reviewing biogeographic patterns have aimed to list recurring patterns, draw phylogeographic generalizations, and assess pseudo-congruence (e.g. Donoghue and Moore 2003, Soltis et al. 2006). We take a similar approach for this discussion of WNA-GLR disjunct taxa. We will 1) review the recurring patterns within this disjunct distribution and provide a list of taxa for each pattern, and 2) reconstruct historical climate models for representatives of each pattern to address hypotheses of vicariance, dispersal, and pseudo-congruence.

List of disjunct vascular plants

We constructed a comprehensive list of vascular plant taxa in the GLR. We downloaded species checklists for Minnesota, Wisconsin, Illinois, Indiana, Indiana Dunes National Lakeshore, Michigan, and Ohio from the Consortium of Midwest Herbaria and an Ontario checklist from the Database of Vascular Plants of Canada (VASCAN, Brouillet et al. 2010+). The Consortium of Northeast Herbaria did not have checklists for Pennsylvania and New York and their specimen lists were inflated with non-native plants, so we downloaded PA and NY checklists from the USDA Plants Database (USDA, NRCS 2018). To capture potential eastern disjuncts that occur along the St. Lawrence Seaway, we downloaded Quebec and New Brunswick checklists from VASCAN and checklists for Vermont, New Hampshire, and Maine from the USDA Plants Database. To narrow this Northeast list to the St. Lawrence region we took a conservative approach and combined lists south of the river and retained only the species that were also found in Quebec.

To account for the different taxonomic treatment in each of the sources, we uploaded each list, including authorities, to the iplantcollaborative Taxonomic Name Resolution Service

(Boyle et al. 2013) to get the best name when compared against five databases (The Plant List, USDA, International Legume Database and Information Service, TROPICOS, and the Global Composite Checklist). This taxonomic service chooses the name that is most commonly listed as the accepted name among the databases. We merged all lists into one species list removing duplicate names.

We visually inspected distribution maps of each species, using the USDA Plants Database, the Biota of North America Program (BONAP) North American Plant Atlas (NAPA) (Kartesz 2015), and the Flora of North America database (Flora of North America Editorial Committee, 1993+). We retained only native taxa and binned them into two categories: disjunct and not disjunct. Taxa were disjunct if they fell into one of the following categories (Fig. 1):

I. Western taxa disjunct in GLR

II. Western taxa disjunct in GLR + disjunct in Northeast

III. Western taxa disjunct in GLR + disjunct in the St. Lawrence Seaway

IV. Western taxa disjunct in GLR + Appalachian affinity

V. Western taxa disjunct in Northeast

VI. GLR taxa disjunct in mid-Canada

Taxa in each of these categories were assigned additional properties if applicable: 1) additional mid-continent, stepping-stone occurrence, 2) extension northward on the east coast, 3) potential disjunct, and 4) disjunct north and south of the glacial line. Potential disjuncts were those that showed a continuous distribution but were marked as rare or historical between the West and the GLR. Category IV (Appalachian affinity) is incomplete because it was difficult to

define this category, and we have only included a few examples here. Category IV is therefore not an exhaustive list. In order to restrict this review to North American taxa, we took a conservative approach and did not include GLR disjuncts with the southwest since these taxa often have continuing distributions into Central and South America. Our final list was reviewed by Ted Cochrane of the UW-Madison Herbarium who provided assistance checking for non-native species and validity of our distributional assignment.

Historical distribution modeling

To test a null hypothesis that the taxa were widespread pre-Pleistocene and the GLR disjunct taxa were restricted to a Driftless Area refugium during the glacial cycles (vicariance scenario), we conducted historical climate niche modeling at the LGM for 3-4 taxa at the species rank from each disjunct category (Table 1). There is an ongoing discussion on how to model paleoclimate (Braconnot et al. 2011, $PMIP_3$ https://pmip3.lsce.ipsl.fr/); Cane et al. 2006; Kageyama et al. 2016; Ludwig et al. 2018). A collaboration between The Paleoclimate Modeling Intercomparison Project Phase 3 and the Coupled Model Intercomparison Project Phase 5 (Sueyoshi et al. 2013) resulted in a model that simulated paleoclimate based on the MIROC-ESM general circulation model (GCM). This GCM incorporates atmosphere, ocean, sea ice, land surface, ocean and terrestrial biogeochemistry, and atmospheric chemistry and aerosols (Watanabe et al. 2011). We chose to download 19 LGM bioclimatic variables from these simulations from Worldclim v.1.4 (Fick and Hijmans 2017, www.worldclim.org). The highest resolution available was 2.5 minutes. In order to project current niche space onto historical climate, we downloaded the same 19 current climate variables at 2.5' resolution, using the MIROC-ESM-CHEM model, which was one of the models used in fifth IPCC report

(www.ipcc.ch, 5th IPCC Report, 2013). This model incorporates variation in surface temperature in addition to the factors already in the MIROC-ESM model (Watanabe et al. 2011). In QGIS v. 2.18 (QGIS Development Team 2018) we cropped the bioclim rasters to North America using the bounding GPS coordinates: 47.1ºW, 179.7ºE, 24.2ºN, 76.2ºN (WGS 84). We downloaded locality data for each species from the Global Biodiversity Information Facility (GBIF) using the R package '*dismo*' (Hijmans and Elith 2013) and restricted the locations to an area bounded by the same North American coordinates. *Cryptogramma stelleri*, *Festuca altaica*, *Galium kamtschaticum*, *Galium palustre*, *Huperzia miyoshiana*, *Juncus articulatus*, *Poa secunda*, *Polystichum lonchitis*, and *Rubus parviflorus* are found outside of North America. We did not want to factor in external niche space for projections of North American distributions, so we took a conservative approach and excluded localities outside of North America from the analysis.

For each species, we tested for correlation between the 19 bioclimatic variables using the '*maptools*' (Bivand and Lewin-Koh 2013) and '*raster*' (Hijmans and van Etten 2014) packages in R. We removed variables that had correlation outside of the -0.8 to 0.8 range based on the Pearson correlation statistic. We used only the uncorrelated variables for each species in our MaxEnt models. We ran a MaxEnt projection for each species and crossvalidated across 100 iterations to increase confidence in the model. Because our data were presence-only, which can be biased toward easily collected areas, we conducted random background sampling instead of providing pseudo-absence points. Pseudo-absence points guess where the true absences are, but if presence-samples are biased, this approach can result in misleading models (Phillips et al. 2009). We kept probability of presence at each site at the default 0.5 because of uncertainty in sampling bias, as suggested in the MaxEnt manual, and we removed duplicate records. We assessed the final models by looking at the output graphs of omission and predicted area, as well

as the calculated area under the receiver operating curve (AUC), which summarize the performance of each bootstrap replicate.

RESULTS

List of WNA-GLR disjunct vascular plants

Our initial list contained 266 taxa at the generic, species and subspecific ranks. After removing non-native, non-disjunct, and hybrid entries, our list contained 114 taxa (Table 2). Angiosperms made up the vast majority of our list (103 taxa). Nine taxa were ferns and two were clubmosses. There were 27 taxa in Category I (Disjunct in the GLR), 6 taxa in Category II (Disjunct in the GLR and disjunct in the Northeast), 45 taxa in Category III (Disjunct in the GLR and disjunct in the St. Lawrence Seaway), 8 taxa in Category IV (Disjunct in the GLR + Appalachian affinity), 21 taxa in Category V (Disjunct in the Northeast), and 7 taxa in Category VI (Disjunct in Mid-Canada). Only one species, *Salix calcicola*, was noted as an eastern North American species that was disjunct in the West. Three taxa were noted as potential disjuncts: *Askellia elegans*, *Crataegus douglasii*, and *Carex flava*. Eighteen of the 41 species that Marquis and Voss reviewed in 1981 were also found in our list. The ones from Marquis and Voss 1981 that we did not include were found to be continuous in distribution.

Historical climate niche modeling

All species had suitable niche space south of the glacial line during the LGM (Fig. 2). Only *Aconitum columbianum* (Category II) and *Poa secunda* (Category V) showed some evidence of suitable niche space in the GLR at the LGM. *Oplopanax horridus, Rubus parviflorus*, *Aconitum columbianum*, *Dryas drummondii*, *Polystichum lonchitis*, *Cryptogramma*

stelleri, *Festuca altaica*, *Huperzia miyoshiana*, and *Poa secunda* had conspicuous suitable niche space in Alaska. Of these species, *Oplopanax horridus*, *Dryas drummondii*, *Polystichum lonchitis*, *Cryptogramma stelleri*, and *Festuca altaica* were conspicuously disjunct north and south of the glacial line. Only *Piperia unalascensis*, *Adiantum aleuticum*, and *Aspidotis densa* showed a restricted distribution in the West, and other western species had broader ranges within the Rocky Mountain range and further south. The species disjunct in the GLR (Category I) all showed a broad stretch of suitable niche space across the continent with varying degrees of suitability. The species disjunct in the GLR and disjunct in the Northeast (Category II) showed both restricted suitable niche space in the West (*Piperia unalascensis*), broader suitable niche space across the continent, and disjunct suitable niche space in the GLR (*Aconitum columbianum*). The species disjunct in the GLR and disjunct in the St. Lawrence Seaway (Category III) showed western-restricted suitable niche space and mid-continent restriction (*Galium kamtschaticum*). The species disjunct in the GLR + Appalachian affinity (Category IV) showed cross-continent suitable niche space, but *Galium palustre* was restricted mid-continent, and *Cryptogramma stelleri* also had suitable niche in Alaska. The species disjunct in the Northeast (Category V) showed disjunct suitable niche disjunct on both sides of the continent (*Huperzia miyoshiana* and *Poa secunda*), continuous suitable niche across the continent with disjunction in Alaska (*Festuca altaica*), and western-restricted suitable niche space (*Aspidotis densa*). The GLR species that are disjunct in mid-Canada (Category VI) all showed a restricted mid-continent suitable niche space except for *Spergularia canadensis*, whose suitable range was highly predicted on the west coast. The two *Galium* species that were in separate categories also showed this mid-continent restriction.

DISCUSSION

List of WNA-GLR disjunct vascular plants

Our taxon list is meant to be an up-to-date review of the taxa in this distribution based on the wealth of distributional data now available (Table 2). We recognize that taxa might have been missed due to nomenclatural shifts, incomplete state plant lists, or incomplete map data. We expect that this missing information is most likely to occur at the subspecific rank, given that these names tend to shift more frequently, and map data for subspecies are not always available. In addition, the maps we used to filter taxa highlight whole states/regions if there is a single occurrence in that region, sometimes masking true disjuncts. We made strong efforts to include taxa that we were uncertain of to check against additional map resources and GBIF waypoints, but we might have missed disjuncts based on our initial filtering. Finally, creating discrete categories for the distributions of these species was a difficult task, and we maintained broad strokes to avoid visually teasing apart subtle geographic structure. The binning of our list might change in the future based on additional localities that might not have been captured in the maps we checked.

Overall, we have listed over 60 additional taxa beyond those discussed in Marquis and Voss (1981), illustrating the utility of current distributional data resources. We considered the western disjuncts with the Northeast (Category V) even though it does not specify the GLR. This cross-continent disjunction underwent the same historical pressures as the GLR disjuncts, and many of the questions and analytical methods still apply. At first glance it may seem that the Driftless Area refugium hypothesis does not apply to this category. However, some of these taxa also tend to have a circum-arctic affinity (Table 1, Fernald 1925), and it has been suggested that southern migration along montane/alpine peaks occurred because those habitats mimic the arctic

(Fernald 1925). This was cited for *Aconitum columbianum*, which has found suitable arctic-like disjunct habitat in the algific talus slopes of the Driftless Area (Brink 1982). It is possible that these western-Northeast disjuncts share a similar biogeographic history with some of the taxa that are also found in the GLR, without exhibiting disjunct populations in the GLR.

Another category that we do not explore in depth is affinity with the Appalachian region. We did not do a comprehensive survey of this distribution because the fine-scale regional phylogeographic breaks around the Appalachian Mountains in the northeast have been previously discussed (Soltis et al. 2006), and we did not think it was appropriate to apply broad strokes to this region. In addition, it was difficult to discretize this category with taxa that had a coastal affinity or taxa that were generally widespread in the Northeast, so we included only a few example taxa. This category is certainly important given the number of taxa that we found and did not include (results not shown), but it will be an area of future research.

Historical distribution modeling

Evidence of suitable climate space in the GLR at the LGM does not mean that the species occupied that space, and we avoid interpreting it as such. Our null hypothesis was that the GLR populations were restricted in a Driftless Area glacial refugium after vicariance of a formerly widespread pre-Pleistocene distribution. The null hypothesis is rejected for taxa that do not have suitable niche space in the GLR during the LGM. Only *Aconitum columbianum* and *Poa secunda* showed potential niche space in the GLR, indicating a potential LGM refugium. Our historical projections are not consistent across species and are not consistent within categories. This supports the hypothesis that the WNA-GLR disjuncts are incongruent and have pseudo-parallel biogeographic histories.

Hypotheses to explain this distribution have considered west to east modes of migration (Iltis 1965, Marquis and Voss 1981). Our historical distribution modeling supports this pattern by showing more instances of western North American restriction than eastern North American restriction (Fig. 2). Eastern North American refugia have been suggested by previous phylogeographic studies, however, (Jaramillo-Correa et al. 2004, Godbout et al. 2005, Soltis et al. 2006, Lee-Yaw et al. 2008, Beatty and Provan 2011) and might serve to explain the restricted mid-continent suitable niche space of the mid-Canada disjuncts and the two *Galium* species (Figures 2.III, 2.IV, and 2.VI). In the case of the *Galium* species, the potential expansion from a mid-continent refugium to a cross-continent disjunction is a migration pattern that we have not considered here, but somewhat resembles the hypothesis of a Driftless Area refugium. This is an exciting case study for future research because it adds complexity to the current models we have investigated.

Our SDMs included only climate factors and did not use any other ecological or geological data to avoid over-saturation of wide margins of error in historical models. Ecological and geological data might reveal a more restrictive historical projection than the ones we generated for *Galium palustre* and *Galium kamtschaticum* (Figures 2.III and 2.IV), but our conclusion of incongruence would likely still hold, given the mid-continent distributions of *Galium*. Discounting the *Galium* outliers, a more restrictive projection might limit the suitable niche space to the west, indicating a more congruent pattern of western-restriction pre-Pleistocene. This should be tested further with additional environmental variables.

Our SDMs were derived from paleoclimate models that were based on a general circulation model (MIROC-ESM). However, there has been ongoing discussion about incorporation of regional/local models of climate data (Ludwig et al. 2018). The integration of

both kinds of models has been suggested to make the representation of physical processes and climate feedbacks more realistic than when using GCM alone, but there is variation in how much added value there is based on the climate variable (Ludwig et al. 2018). An additional consideration for SDMs is the locality data used to describe a species' current range. It has been shown that the lack of intra-specific information or sampling across intra-specific groups can bias climate niche models (D'Amen et al. 2012). In our models we excluded localities that fell outside of North America, and ideally we would also test these models using the full breadth of climate space for each species. However, we argue that describing the full realized niche (and adaptive potential) of a species is most beneficial when projecting the niche onto future climate projections. Since the species already occupies this niche range, we have more confidence that it can occupy it in the future. However, for historical projections, it is less intuitive to project the most recent adaptive potential from across all potentially divergent geographic regions onto a limited historical range.

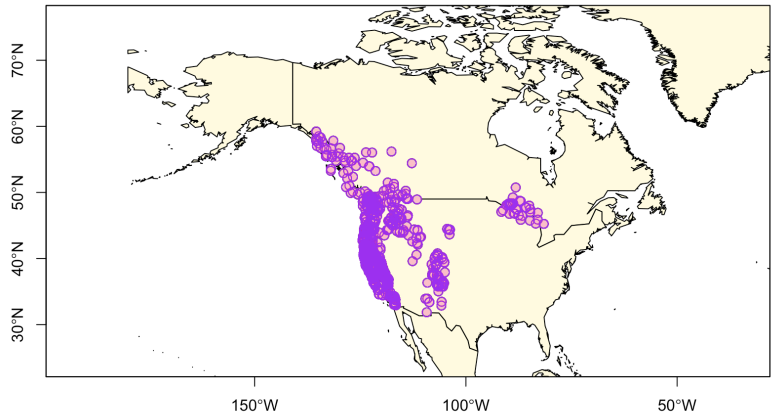
Major geological events in North America include the rise of the Rocky Mountain Range between 80 and 55 Ma, the aridification of the Great Plains from forest to grassland 25 Ma, and the more recent history of the Pleistocene glacial cycles 2.58 Ma-11,700 Ya. We focused our SDM analysis on the recent geological history of the Pleistocene at the species rank, but this leaves room for discussion on the history of disjunct genera, which might incorporate more complex models of migration and speciation.

CONCLUSION

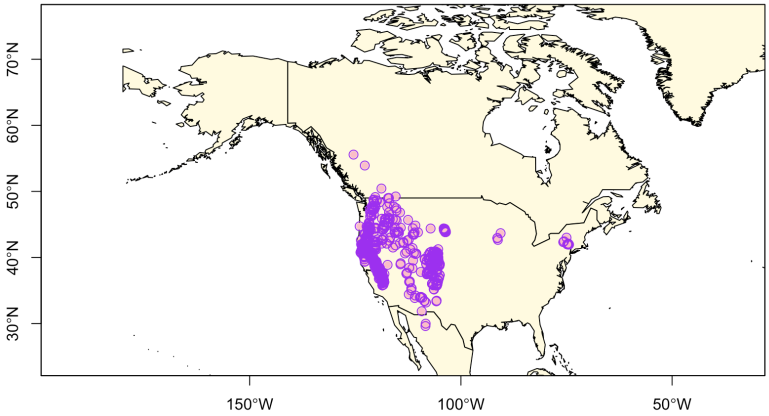
The phylogeography of disjunct distributions can be teased apart with multiple lines of genetic, temporal, geological, paleo-ecological, and environmental evidence. In addition to

studies that use single taxa as case studies, comparative phylogeographic approaches are important for understanding regional patterns of biodiversity. We have described analytical approaches for unveiling the individual phylogeographic histories of the WNA-GLR disjunct taxa. We present an updated list of the vascular plant taxa of this distribution, which we hope to be a starting point for future comparative research. We have shown based on historical projections of niche models that there is pseudo-congruence in the biogeographical histories of these species, and we present avenues for future research on additional migration models to describe this enigmatic distribution.

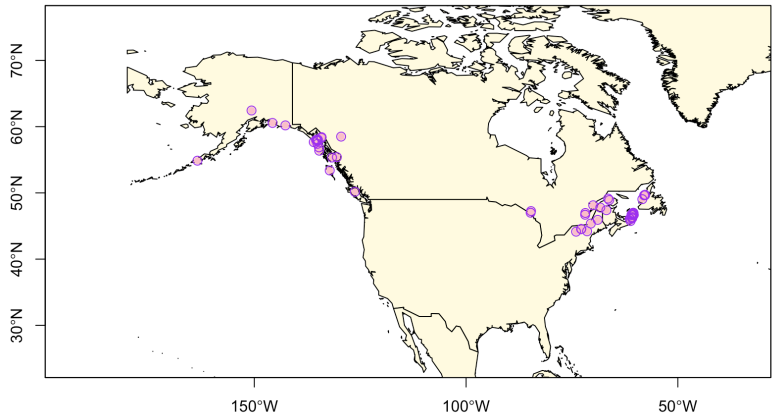
I: Disjunct in GLR
Rubus parviflorus



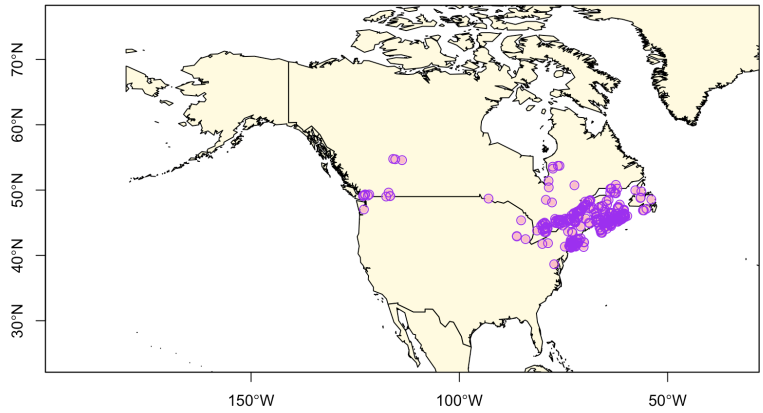
II: Disjunct in the GLR and disjunct in the Northeast
Aconitum columbianum



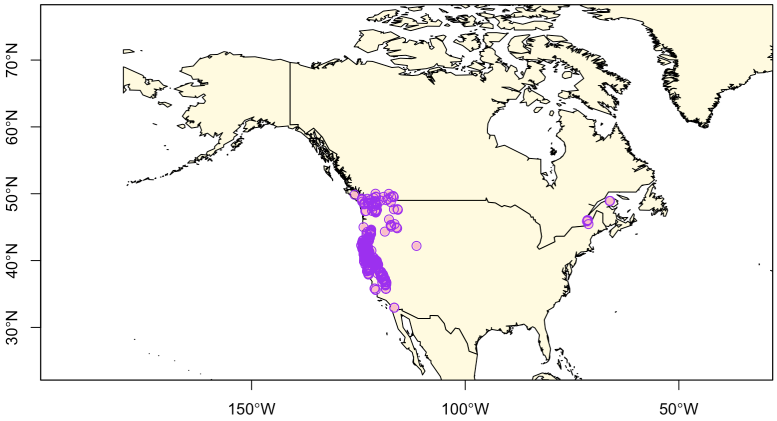
III: Disjunct in the GLR and disjunct in the St. Lawrence Seaway
Galium kamtschaticum



IV: Disjunct in the GLR + Appalachian affinity
Galium palustre



V: Disjunct in the Northeast
Aspidotis densa



VI: Disjunct in Mid-Canada
Dendrolycopodium hickeyi

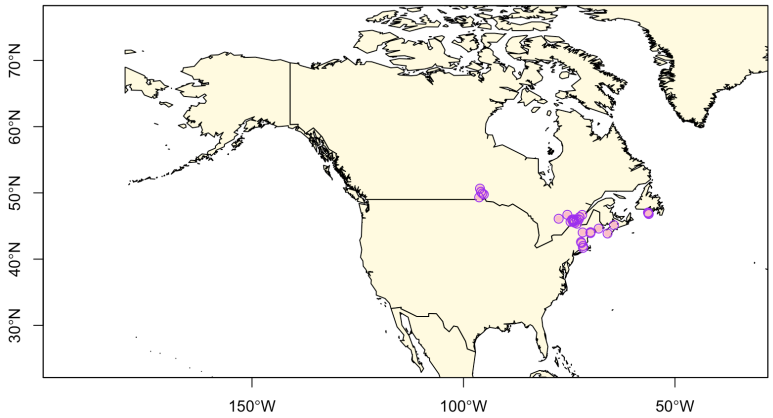


Fig. 1 (Above) Distribution patterns of the six categories: **I-** Disjunct in the GLR, **II-** Disjunct in the GLR and disjunct in the Northeast, **III-** Disjunct in the GLR and disjunct in the St. Lawrence Seaway, **IV-** Disjunct in the GLR + Appalachian affinity, **V-** Disjunct in the Northeast, and **VI-** Disjunct in Mid-Canada. Each map was generated from locality data for a representative species of that category. Locality data were downloaded from GBIF (see methods).

Table 1: A list of the 24 species that we selected for MaxEnt historical distribution mapping at the Last Glacial Maximum. Species are grouped by their disjunct category: **I-** Disjunct in the GLR, **II-** Disjunct in the GLR and disjunct in the Northeast, **III-** Disjunct in the GLR and disjunct in the St. Lawrence Seaway, **IV-** Disjunct in the GLR + Appalachian affinity, **V-** Disjunct in the Northeast, and **VI-** Disjunct in Mid-Canada. "Extra" column contains additional features if applicable: **SS:** stepping-stone occurrence, **NS_glacier:** disjunct occurrence north and south of the glacial line, **Potential:** continuous, but listed as historical, locally extirpated, or rare in between the Rocky Mountains at the GLR, **Ext_N:** extends northward on the East coast, and **E_disjunct:** a predominantly eastern taxon that seems disjunct in the West. Species were chosen based on the availability of way point data from GBIF.

FAMILY	TAXON	AUTHORITY	TYPE	EXTRA
Araliaceae	Oplopanax horridus	(Sm.) Miq.	I: GL	
Asteraceae	Adenocaulon bicolor	Hook.	I: GL	
Caprifoliaceae	Valeriana edulis	Nutt.	I: GL	SS
Ranunculaceae	Trollius laxus	Salisb.	I: GL	
Rosaceae	Rubus parviflorus	Nutt.	I: GL	SS
Dryopteridaceae	Polystichum lonchitis	(L.) Roth	II: GL, NE	SS, Ext_N
Orchidaceae	Piperia unalascensis	(Spreng.) Rydb.	II: GL, NE	SS
Ranunculaceae	Aconitum columbianum	Nutt.	II: GL, NE	SS
Rosaceae	Dryas drummondii	Richardson ex Hook.	II: GL,NE	
Ericaceae	Pterospora andromedea	Nutt.	III: GLR, SL	SS
Pteridaceae	Adiantum aleuticum	(Rupr.) C.A. Paris	III: GLR, SL	
Rubiaceae	Galium kamtschaticum	Steller ex Schult.	III: GLR, SL	
Juncaceae	Juncus articulatus	L.	IV: GLR + AP	
Phrymaceae	Mimulus moschatus	Douglas ex Lindl.	IV: GLR + AP	
Pteridaceae	Cryptogramma stelleri	(S.G. Gmel.) Prantl	IV: GLR + AP	Ext_N
Table 1 cont'd				
Rubiaceae	Galium palustre	L.	IV: GLR + AP	E_disjunct
Lycopodiaceae	Huperzia miyoshiana	(Makino) Ching	V: NE	
Poaceae	Festuca altaica	Trin. ex Ledeb.	V: NE	

Poaceae	Poa secunda	J. Presl	V: NE	
Pteridaceae	Aspidotis densa	(Brack.) Lellinger	V: NE	
Amaranthaceae	Atriplex glabriuscula	Edmondston	VI: Mid (IV)	
Caryophyllaceae	Spergularia canadensis	(Pers.) G. Don	VI: Mid (IV)	
Cyperaceae	Trichophorum clintonii	(A. Gray) S.G.Sm.	VI: Mid (III)	
Lycopodiaceae	Dendrolycopodium hickeyi	(W.H. Wagner, Beitel & R.C. Moran) A. Hain	VI: Mid (IV)	

Table 2: Our list of 114 taxa at the genus, species, and subspecific ranks that are part of the WNA-GLR disjunct distribution. Taxa are binned into categories **I-VI: I-** Disjunct in the GLR, **II-** Disjunct in the GLR and disjunct in the Northeast, **III-** Disjunct in the GLR and disjunct in the St. Lawrence Seaway, **IV-** Disjunct in the GLR + Appalachian affinity, **V-** Disjunct in the Northeast, and **VI-** Disjunct in Mid-Canada. "Extra" column contains additional features if applicable: **SS:** stepping-stone occurrence, **NS_glacier:** disjunct occurrence north and south of the glacial line, **Potential:** continuous, but listed as historical, locally extirpated, or rare in between the Rocky Mountains at the GLR, **Ext_N:** extends northward on the East coast, and **E_disjunct:** a predominantly eastern taxon that seems disjunct in the West.

I: Disjunct in GLR

FAMILY	TAXON	AUTHORITY	EXTRA
Amaranthaceae	Atriplex joaquiniana	A.Nelson	SS
Amaranthaceae	Sarcocornia pacifica	(Standl.) A.J. Scott	
Araliaceae	Oplopanax	(Torr. & A. Gray) Miq.	
Araliaceae	Oplopanax horridus	(Sm.) Miq.	
Asteraceae	Adenocaulon	Hook.	SS
Asteraceae	Adenocaulon bicolor	Hook.	SS
Asteraceae	Antennaria howellii subsp. howellii	Greene	SS
Asteraceae	Artemisia campestris subsp. borealis var. borealis	(Pall.) M.Peck	NS_glacier
Asteraceae	Askellia elegans	(Hook.) W.A.Weber	Potential
Asteraceae	Aster alpinus subsp. vierhapperi	Onno	NS_glacier
Caprifoliaceae	Valeriana edulis	Nutt.	SS
Crassulaceae	Rhodiola integrifolia	Raf.	SS
Cyperaceae	Carex geyeri	Boott	
Cyperaceae	Carex scirpoidea var. convoluta	K√°k.	
Equisetaceae	Equisetum telmateia	Ehrh.	
Ericaceae	Vaccinium membranaceum	Douglas ex Torr.	SS
Liliaceae	Prosartes hookeri	Torr.	SS
Malvaceae	Iliamna rivularis	(Douglas) Greene	
Plantaginaceae	Besseya	Rydb.	SS
Poaceae	Elymus spicatus	(Pursh) Gould	
Poaceae	Festuca occidentalis	Hook.	SS
Polemoniaceae	Polemonium occidentale	Greene	
Ranunculaceae	Trollius laxus	Salisb.	
Rhamnaceae	Ceanothus sanguineus	Pursh	
Rosaceae	Crataegus douglasii	Lindl.	Potential
Rosaceae	Rubus parviflorus	Nutt.	SS
Violaceae	Viola epipsila	Ledeb.	SS

II: Disjunct in the GLR and disjunct in the Northeast

FAMILY	TAXON	AUTHORITY	EXTRA
Amaranthaceae	Chenopodium leptophyllum	(Nutt. ex Moq.) B.D.Jacks.	
Dryopteridaceae	Polystichum lonchitis	(L.) Roth	SS, Ext_N
Orchidaceae	Piperia unalascensis	(Spreng.) Rydb.	SS
Ranunculaceae	Aconitum columbianum	Nutt.	SS
Ranunculaceae	Trollius	L.	
Rosaceae	Dryas drummondii	Richardson ex Hook.	

III: Disjunct in the GLR and disjunct in the St. Lawrence Seaway

FAMILY	TAXON	AUTHORITY	EXTRA
Amaranthaceae	Salicornia europaea	L.	
Apiaceae	Angelica lucida	L.	
Apiaceae	Ligusticum scoticum	L.	Ext_N
Apiaceae	Osmorhiza berteroi	DC.	SS
Aspleniaceae	Asplenium viride	Huds.	SS
Asteraceae	Arnica lanceolata	Nutt.	
Asteraceae	Aster alpinus	L.	
Cyperaceae	Eleocharis kamtschatica	(C.A.Mey.) Kom.	
Cyperaceae	Eleocharis mamillata	(H.Lindb.) H.Lindb.	
Cyperaceae	Eleocharis nitida	Fernald	
Cyperaceae	Eleocharis ovata	(Roth) Roem. & Schult.	
Cystopteridaceae	Cystopteris montana	(Lam.) Bernh. ex Desv.	Ext_N
Dryopteridaceae	Dryopteris expansa	(C. Presl) Fraser-Jenk. & Jermy	Ext_N
Dryopteridaceae	Dryopteris filix-mas	(L.) Schott	SS, Ext_N
Dryopteridaceae	Polystichum braunii	(Spenn.) Fée	
Ericaceae	Pterospora	Nutt.	SS
Ericaceae	Pterospora andromedea	Nutt.	SS
Ericaceae	Vaccinium ovalifolium	Sm.	
Fabaceae	Lathyrus japonicus subsp. maritimus	(L.)P.W.Ball	
Haloragaceae	Myriophyllum farwellii	Morong	
Iridaceae	Iris setosa	Pall. ex Link	
Juncaceae	Juncus hybridus	Brot.	
Onagraceae	Epilobium anagallidifolium	Lam.	Ext_N
Onagraceae	Epilobium hornemannii	Rchb.	SS, Ext_N
Onagraceae	Epilobium hornemannii ssp. hornemannii	Rchb.	SS, Ext_N

Onagraceae	Epilobium lactiflorum	Hausskn.	Ext_N
Onagraceae	Epilobium saximontanum	Hausskn.	SS
Ophioglossaceae	Botrychium ascendens	W.H. Wagner	
Ophioglossaceae	Botrychium lineare	W.H. Wagner	SS
Ophioglossaceae	Botrychium spathulatum	W.H. Wagner	
Orchidaceae	Goodyera oblongifolia	Raf.	SS
Orchidaceae	Listera convallarioides	(Sw.) Nutt. Ex Elliott	SS
Orchidaceae	Platanthera huronensis	Lindl.	SS
Plantaginaceae	Veronica tenella	All.	
Poaceae	Festuca subverticillata	(Pers.) E.B.Alexeev	Ext_N
Poaceae	Hierochloe alpina subsp. monticola	(Sw.) Roem. & Schult.	NS_glacier
Poaceae	Hordeum brachyantherum ssp. brachyantherum	Nevski	
Poaceae	Melica smithii	(Porter) Vasey	SS
Poaceae	Phleum alpinum	L.	SS, Ext_N
Polygonaceae	Rumex persicarioides	L.	
Pteridaceae	Adiantum aleuticum	(Rupr.) C.A. Paris	
Rosaceae	Argentina egedii ssp. egedii	(Wormsk.) Rydb.	Ext_N
Rubiaceae	Galium kamtschaticum	Steller ex Schult.	
Salicaceae	Salix calcicola	Fernald & Wiegand	Ext_N, E_disjunct
Salicaceae	Salix vestita	Pursh	Ext_N

IV: Disjunct in GLR + Appalachian affinity

FAMILY	TAXON	AUTHORITY	EXTRA
Cyperaceae	Carex echinata ssp. echinata	Murray	SS
Cyperaceae	Carex flava	L.	Potential
Juncaceae	Juncus articulatus	L.	SS
Ophioglossaceae	Botrychium lanceolatum	(S.G. Gmel.) Ångström	Ext_N
Phrymaceae	Mimulus moschatus	Douglas ex Lindl.	
Pteridaceae	Cryptogramma stelleri	(S.G. Gmel.) Prantl	
Rubiaceae	Galium palustre	L.	

V: Disjunct in the Northeast

FAMILY	TAXON	AUTHORITY	EXTRA
Asteraceae	Agoseris aurantiaca	(Hook.) Greene	SS
Asteraceae	Cirsium scariosum	(Poir.) Nutt.	
Asteraceae	Senecio pseudoarnica	Less.	
Brassicaceae	Draba incerta	Payson	

Brassicaceae	Erysimum inconspicuum	(S.Watson) MacMill.	
Cornaceae	Cornus suecica	L.	
Crassulaceae	Crassula saginoides	(Maxim.) M.Bywater & Wickens	
Cyperaceae	Carex macloviana	d'Urv.	Ext_N
Cyperaceae	Carex stylosa	C.A.Mey.	Ext_N
Cyperaceae	Trichophorum pumilum	(Vahl) Schinz & Thell.	
Dryopteridaceae	Polystichum scopulinum	(D.C. Eaton) Maxon	
Fabaceae	Astragalus robbinsii var. minor	(Hook.) Barneby	
Fabaceae	Oxytropis podocarpa	A.Gray	Ext_N
Lycopodiaceae	Huperzia miyoshiana	(Makino) Ching	
Ophioglossaceae	Botrychium pedunculosum	W.H. Wagner	
Poaceae	Festuca altaica	Trin. ex Ledeb.	
Poaceae	Poa eminens	J.Presl	
Poaceae	Poa secunda	J. Presl	
Pteridaceae	Aspidotis densa	(Brack.) Lellinger	
Ranunculaceae	Thalictrum alpinum	L.	Ext_N
Woodsiaceae	Athyrium americanum	(Butters) Maxon	Ext_N

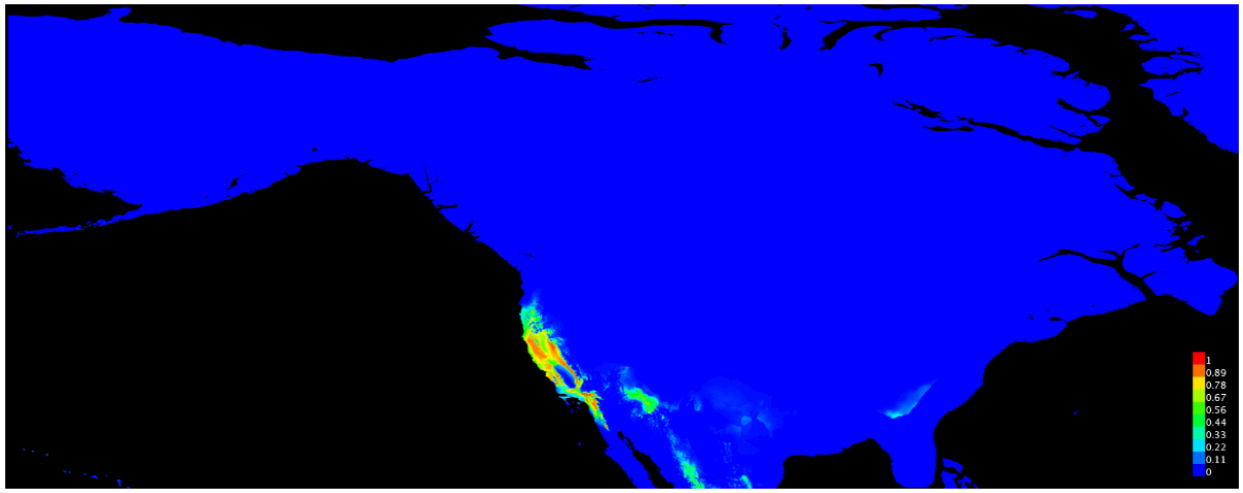
VI: Disjunct in Mid-Canada

FAMILY	TAXON	AUTHORITY	EXTRA
Amaranthaceae	Atriplex glabriuscula	Edmondston	Ext_N
Caryophyllaceae	Spergularia canadensis	(Pers.) G. Don	Ext_N
Cyperaceae	Trichophorum clintonii	(A.Gray) S.G.Sm.	
Gentianaceae	Gentianopsis barbata	(Froel.) Ma	
Lycopodiaceae	Dendrolycopodium hickeyi	W.H. Wagner	
Ophioglossaceae	Botrychium lanceolatum subsp. angustisegmentum	(Pease & A.H. Moore) R.T. Clausen	
Typhaceae	Sparganium glomeratum	(Laest. ex Beurl.) Beurl.	

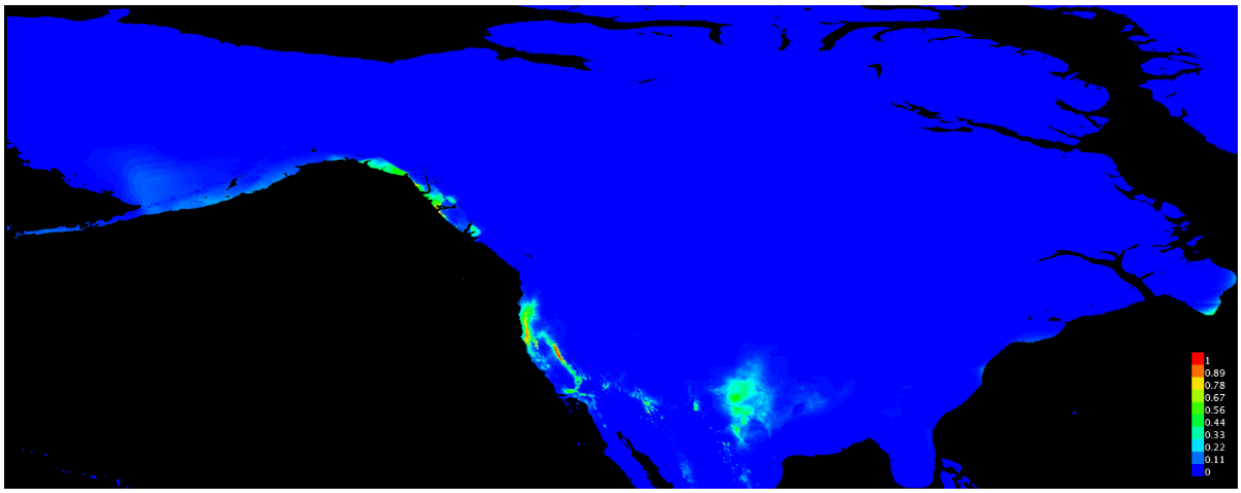
Fig. 2 (Below) The point-wise mean of 100 bootstrapped MaxEnt models applied to environmental layers from the LGM (~21,000 Ya) for 24 taxa. Maps are grouped by their disjunct category (see **Table 1** for information about the taxa). Locality data were collected from GBIF, environmental layers from BioClim were chosen for each taxon if they were uncorrelated (fell within -0.8 - 0.8) based on the Pearson correlation statistic. Warm and cool colors correspond to high and low suitability, respectively. Mean area under the receiver operating characteristic (AUC) is given as a measure of model performance.

Category I: Disjunct in the GLR

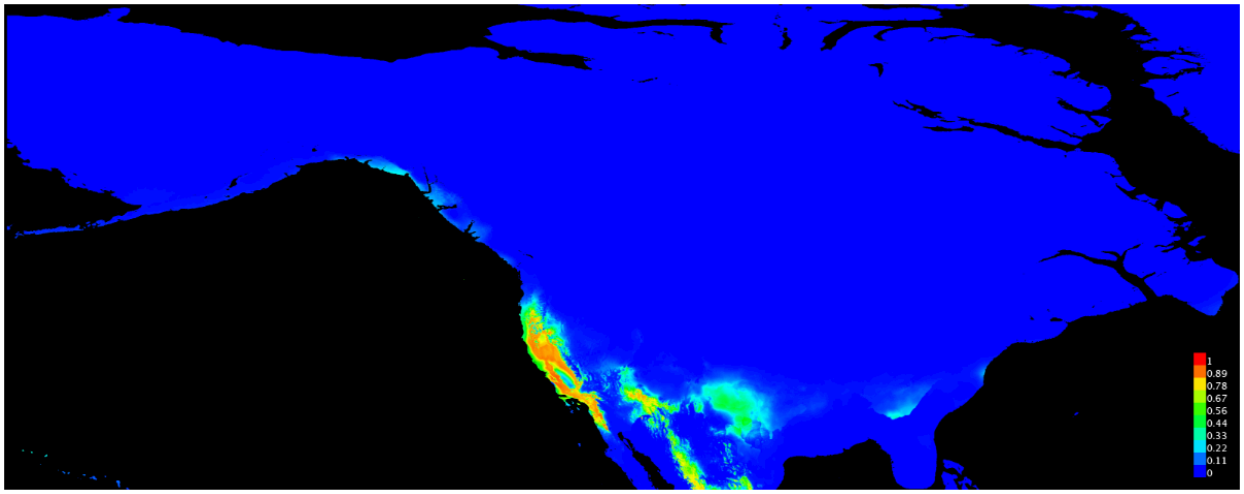
Adenocaulon bicolor (mean AUC = 0.973)



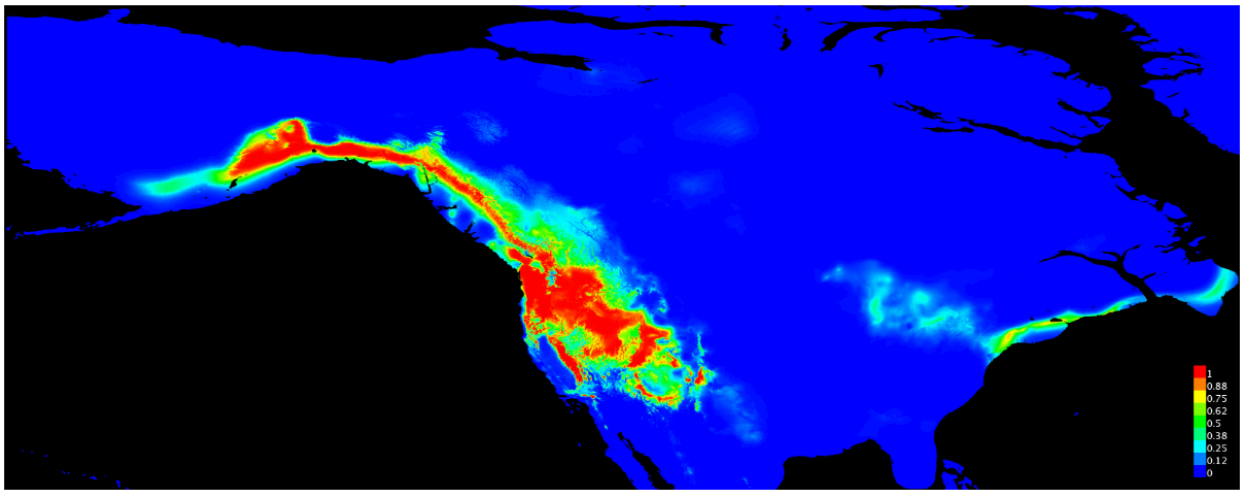
Oplopanax horridus (mean AUC = 0.968)



Rubus parviflorus (mean AUC = 0.937)

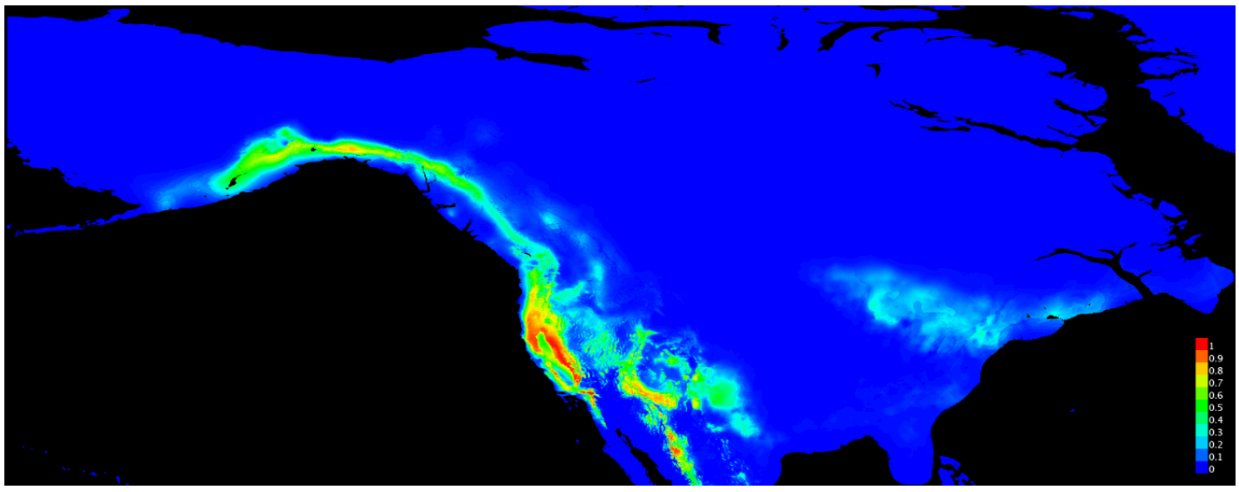


Trollius laxus (mean AUC = 0.969)

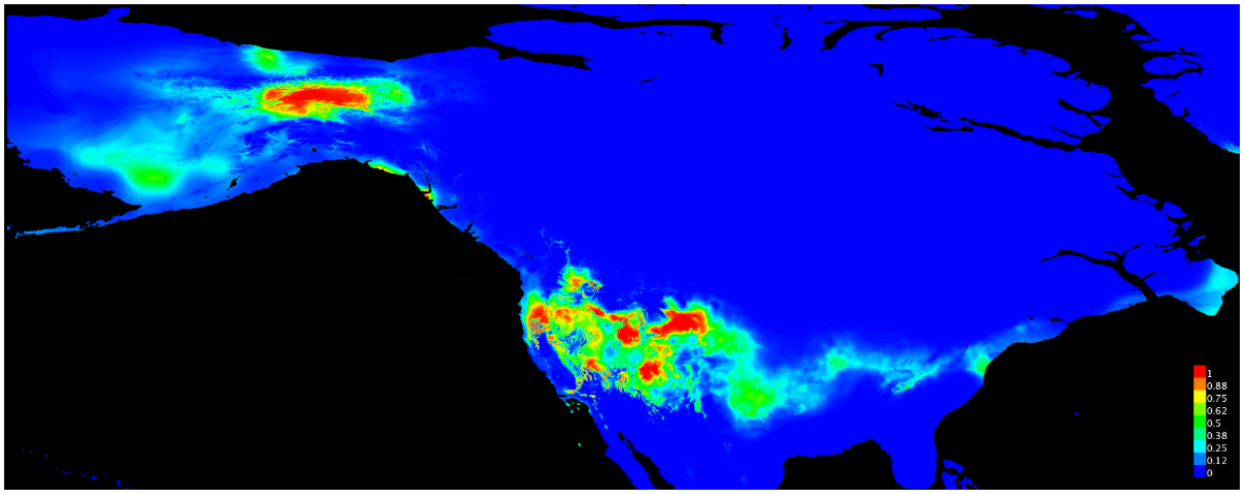


Category II: Disjunct in the GLR and disjunct in the Northeast

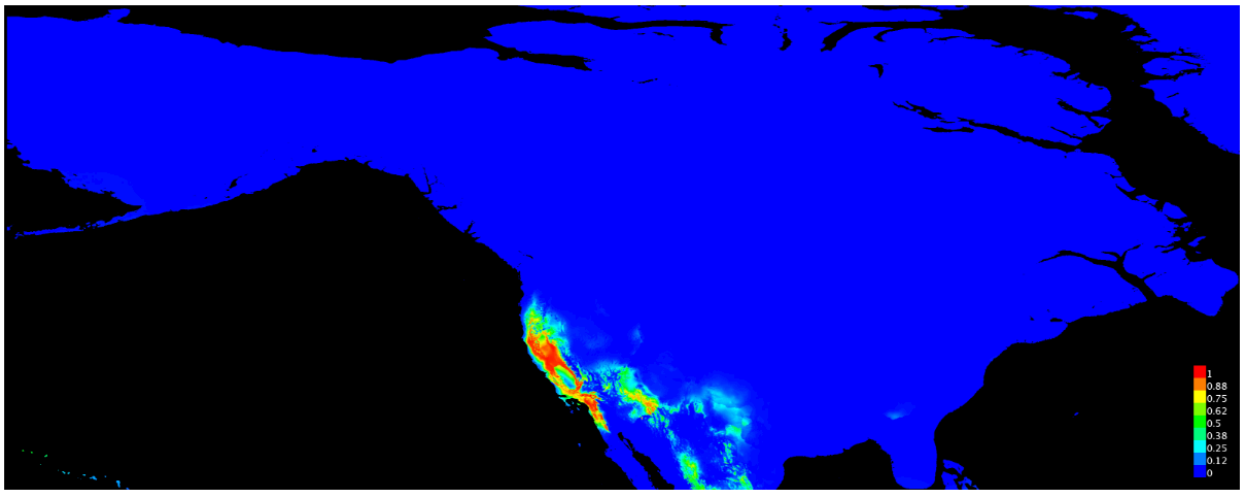
Aconitum columbianum (mean AUC = 0.954)



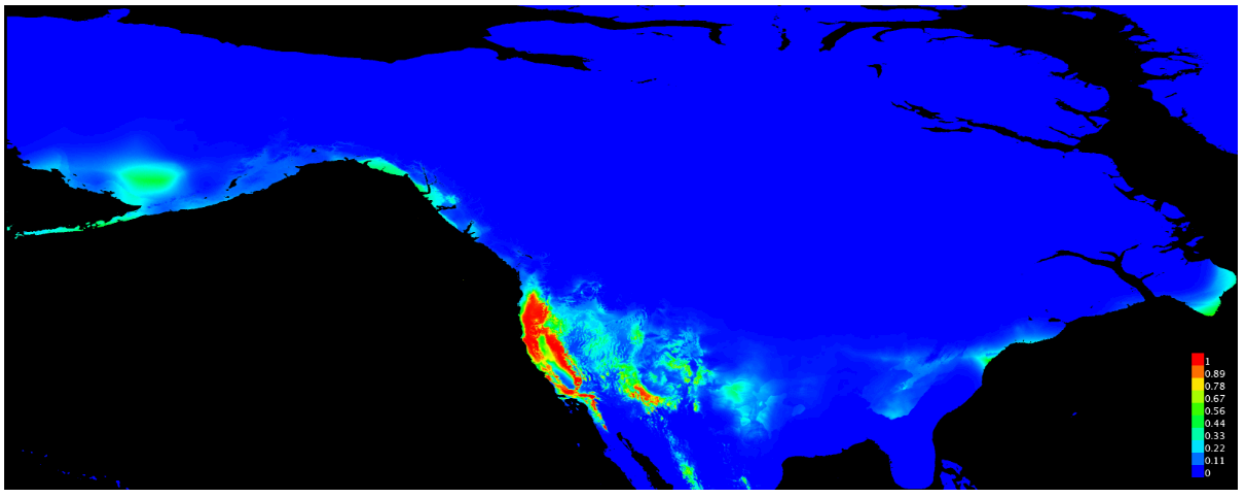
Dryas drummondii (mean AUC = 0.886)



Piperia unalascensis (mean AUC = 0.962)

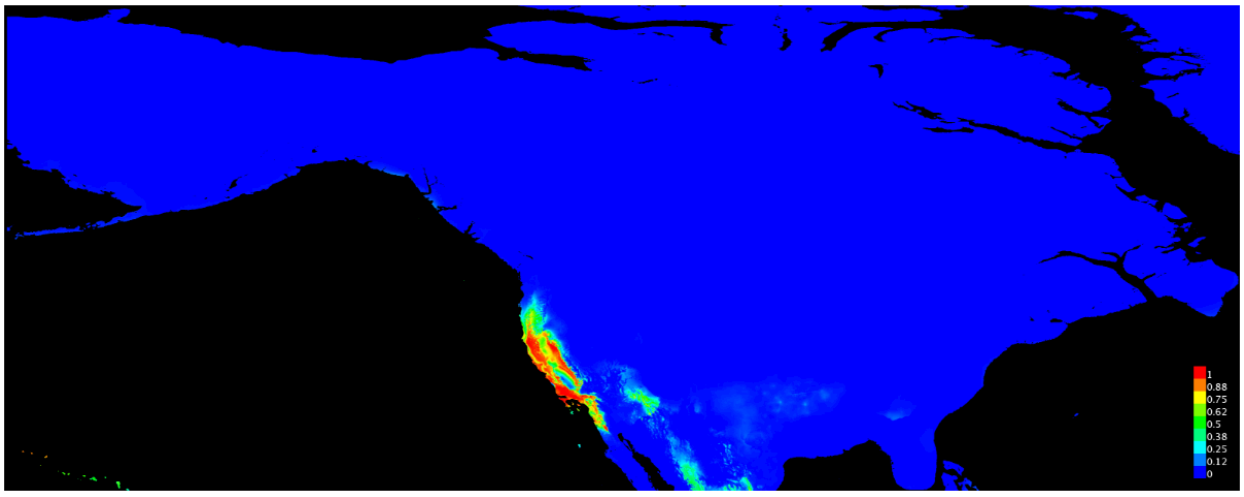


Polystichum lonchitis (mean AUC = 0.931)

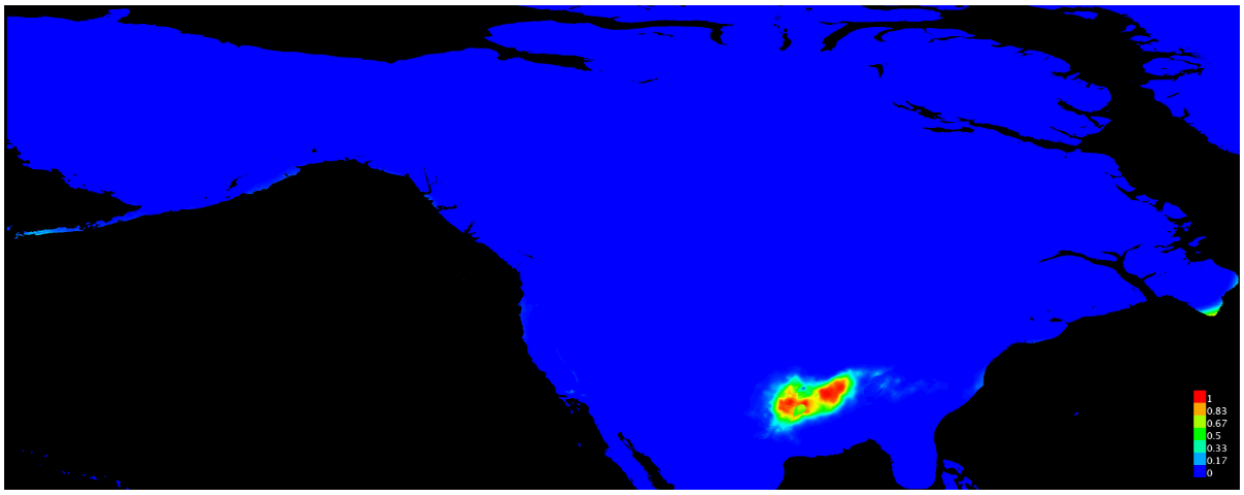


Category III: Disjunct in the GLR + St. Lawrence seaway

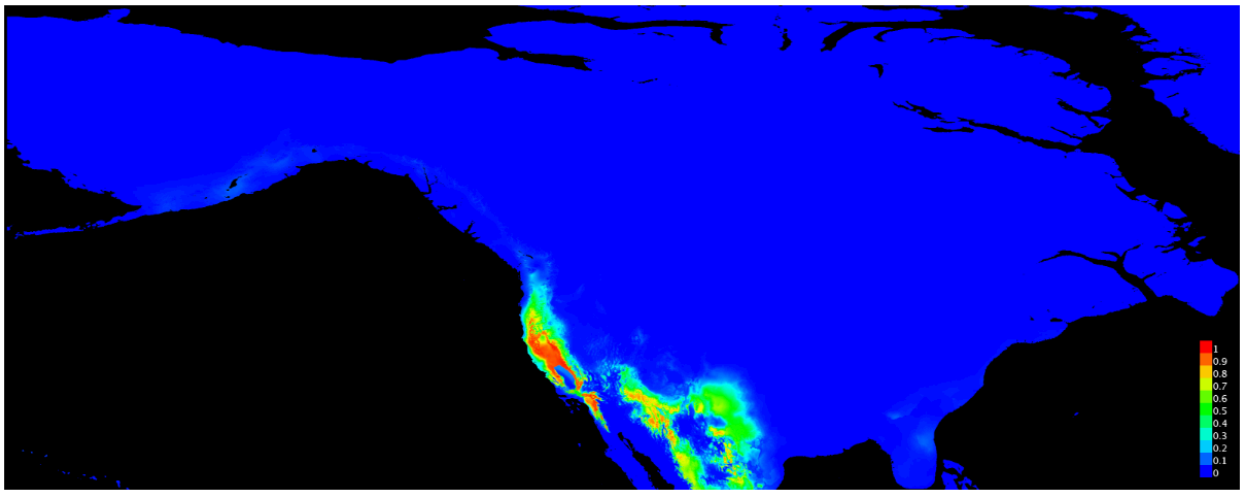
Adiantum aleuticum (mean AUC = 0.969)



Galium kamtschaticum (mean AUC = 0.975)

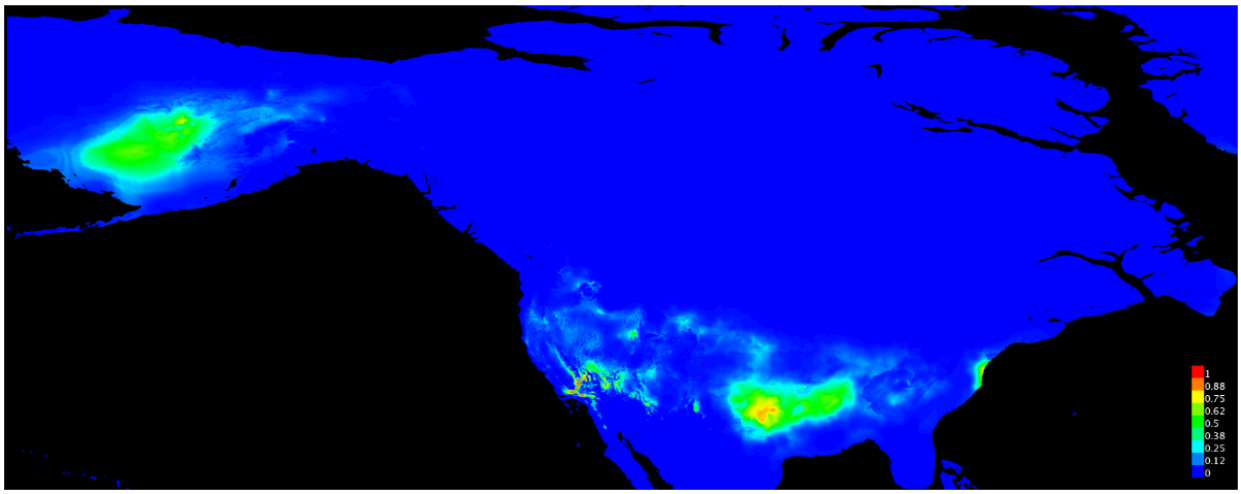


Pterospora andromedea (mean AUC = 0.959)

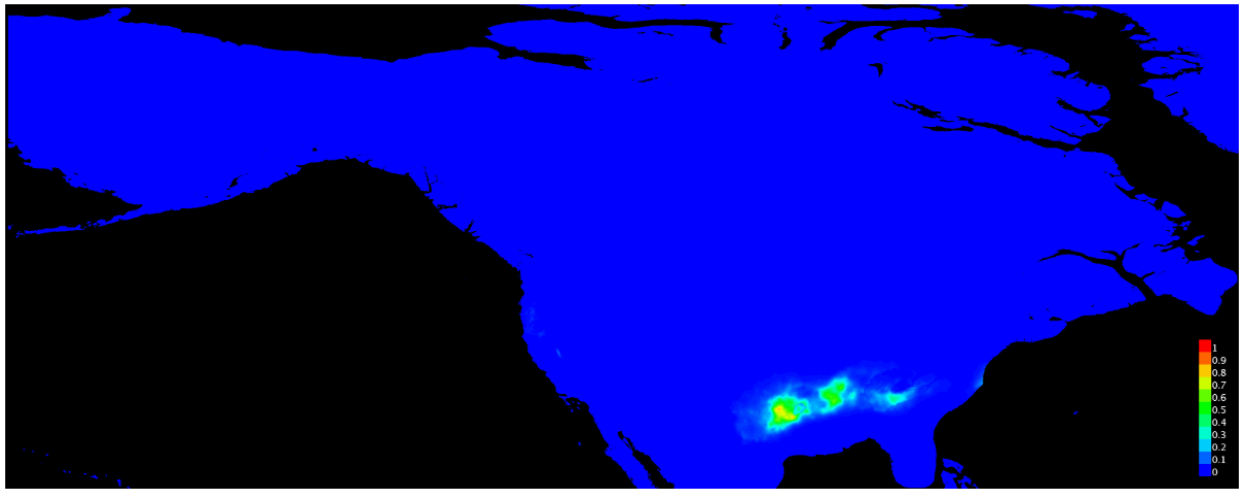


Category IV: Disjunct in the GLR + St. Lawrence seaway + Appalachian affinity

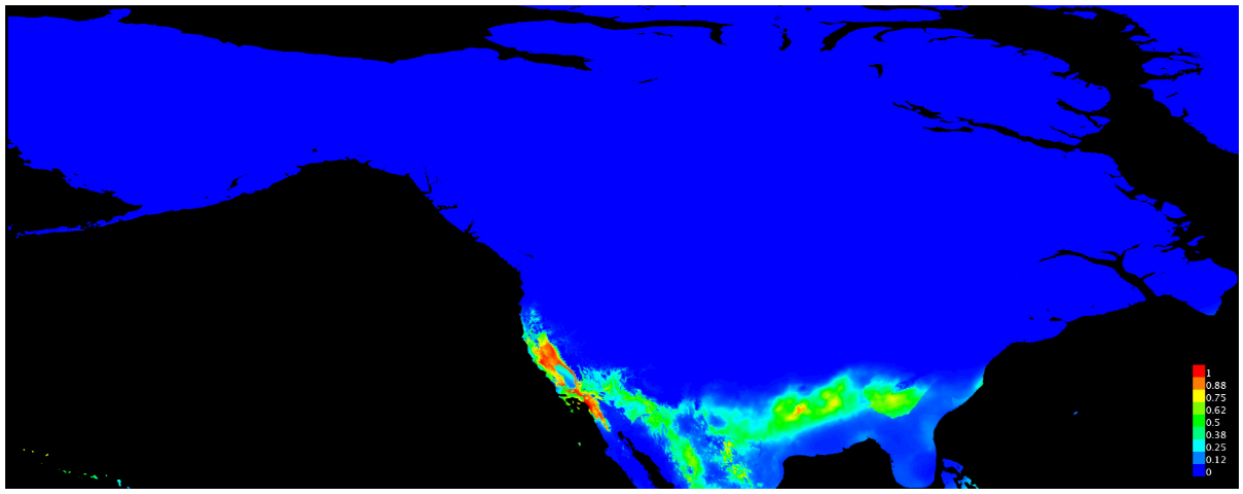
Cryptogramma stelleri (mean AUC = 0.869)



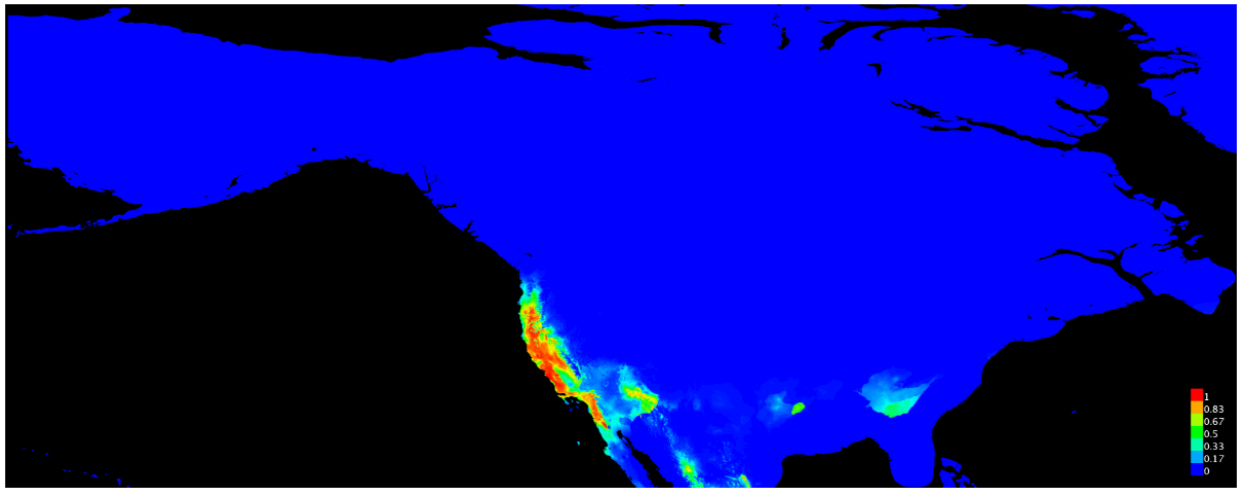
Galium palustre (mean AUC = 0.965)



Juncus articulatus (mean AUC = 0.926)

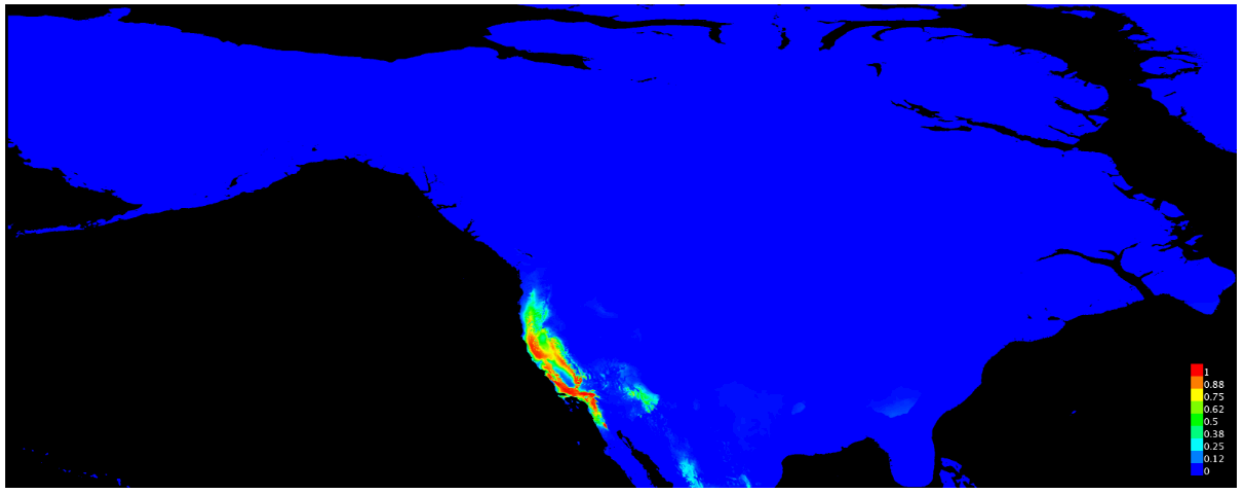


Mimulus moschatus (mean AUC = 0.966)

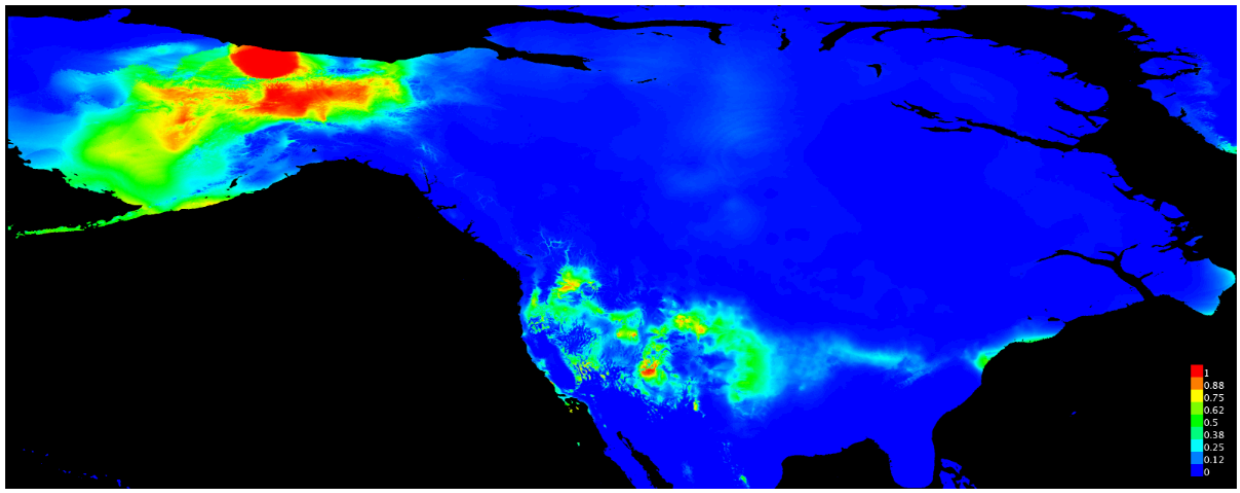


Category V: Disjunct in the Northeast

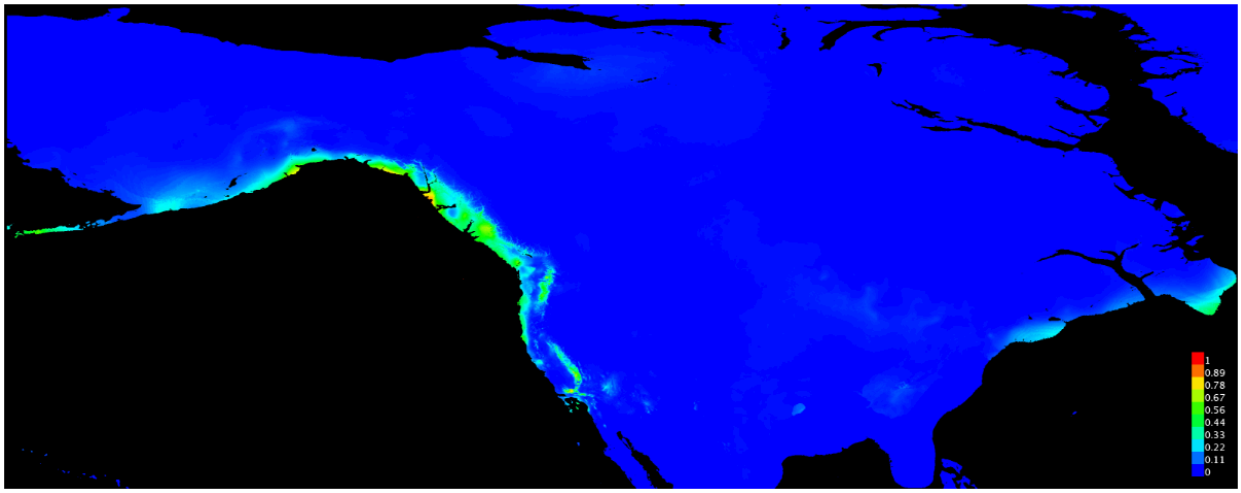
Aspidotis densa (mean AUC = 0.973)



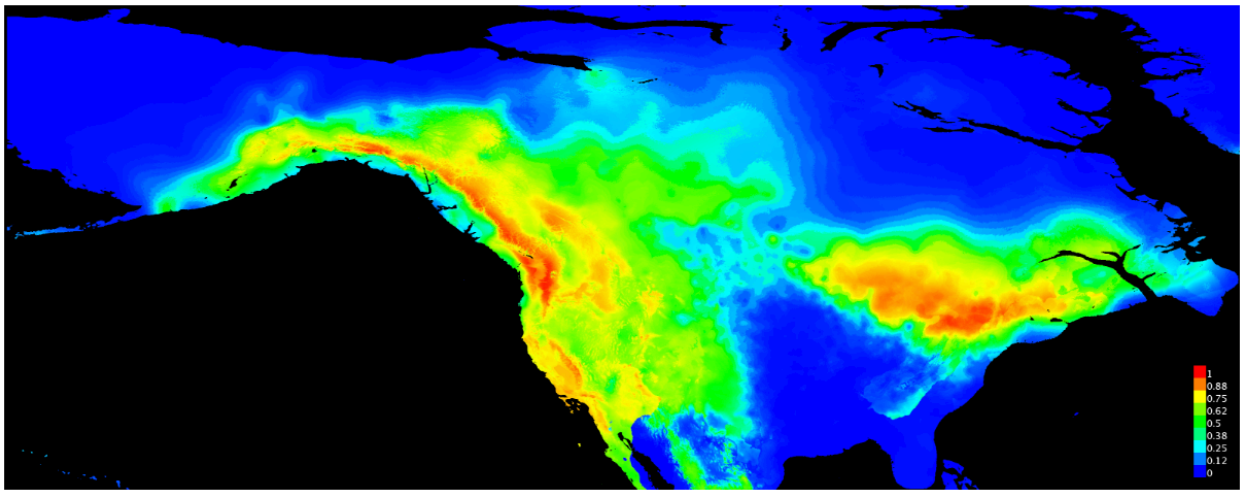
Festuca altaica (mean AUC = 0.886)



Huperzia miyoshiana (mean AUC = 0.969)

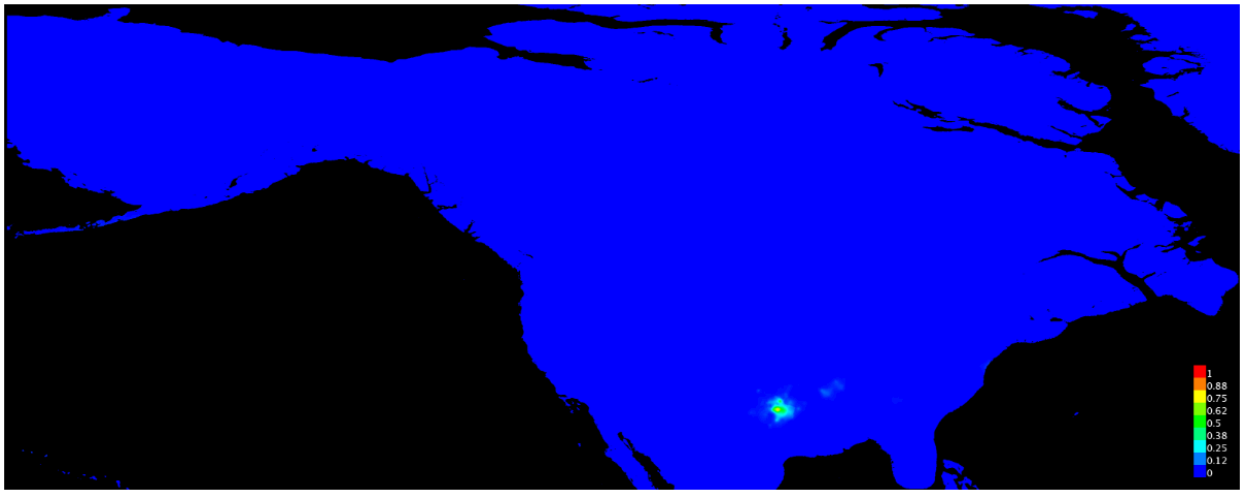


Poa secunda (mean AUC = 0.835)

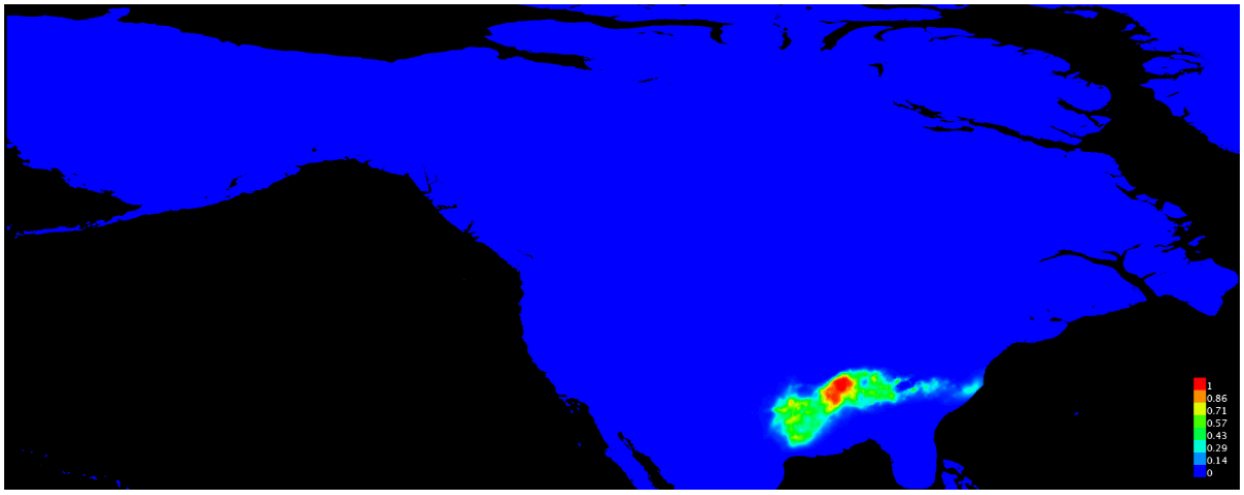


Category VI: Disjunct in Mid-Canada

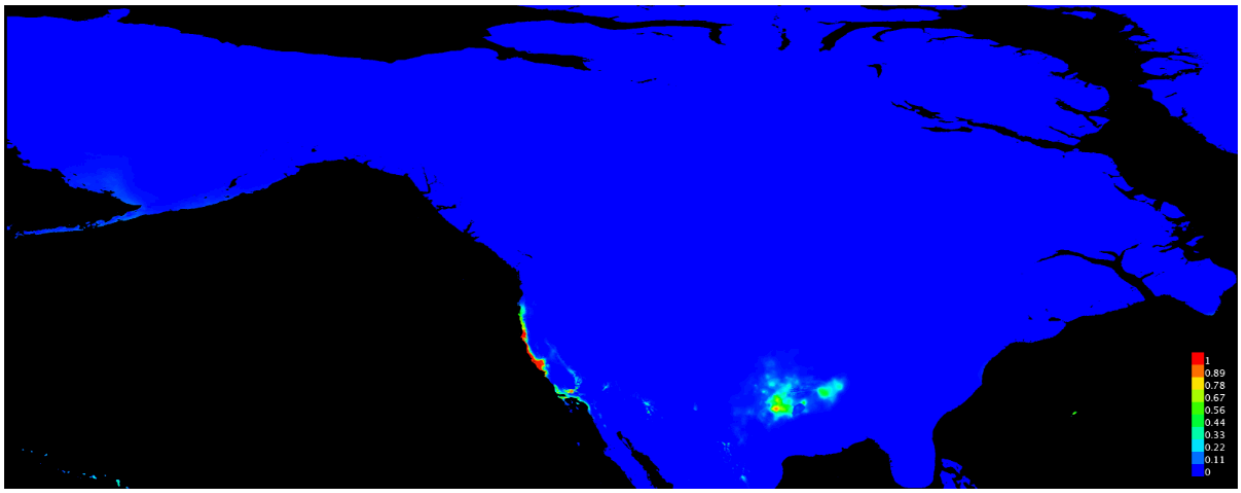
Atriplex glabriscula (mean AUC = 0.967)



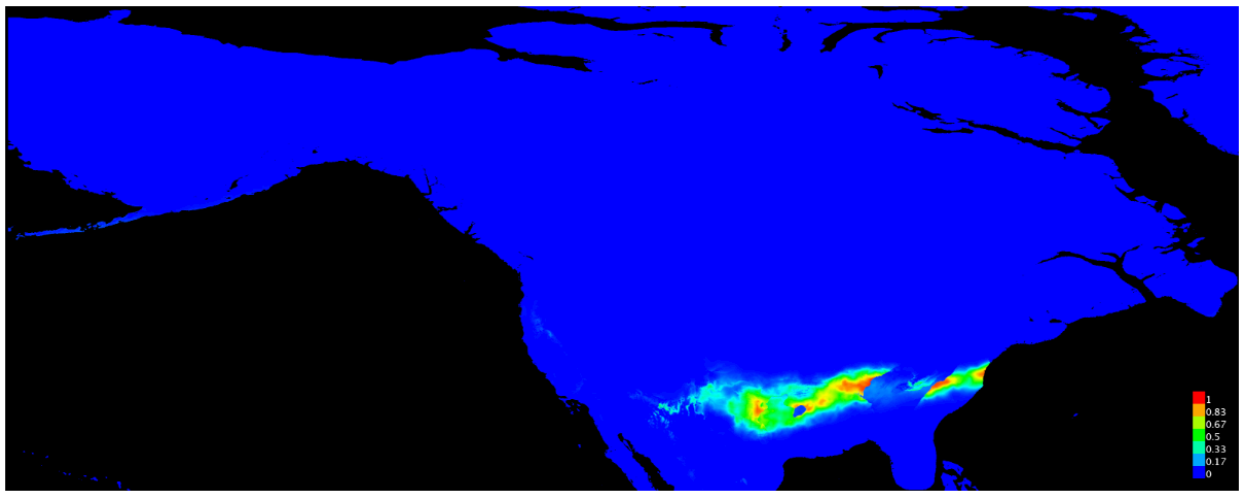
Dendrolycopodium hickeyi (mean AUC = 0.983)



Spergularia canadensis (mean AUC = 0.963)



Trichophorum clintonii (mean AUC = 0.958)



REFERENCES

Alden, W.C., 1904. *The Delavan lobe of the Lake Michigan glacier of the Wisconsin stage of glaciation and associated phenomena*. No. 34. US Government Printing Office.

Araújo, M.B. & Peterson, A.T., 2012. Uses and misuses of bioclimatic envelope modeling. *Ecology*, 93(7), pp.1527–1539.

Arbogast, B.S. & Kenagy, G.J., 2001. Comparative Phylogeography as an integrative approach to historical biogeography. *Journal of Biogeography*, 28(7), pp.819—825.

Ash, J.D., Givnish, T.J. & Waller, D.M., 2016. Tracking lags in historical plant species' shifts in relation to regional climate change. *Global Change Biology*, 23(3), pp.1305–1315.

Avise, J.C. 2000. Phylogeography: the history and formation of species, Harvard University Press, London.

Avise, J.C. et al., 1987, Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *annualreviews.org*

Baker, R.G., Van Zant, K.L. & Dulian, J.J., 2010. Three late-glacial pollen and plant Macrofossil assemblages from Iowa. *Palynology*, 4(1), pp.197–203.

Balloux, F. & Lugon-Moulin, N., 2002. The estimation of population differentiation with microsatellite markers. *Molecular Ecology*, 11(2), pp.155–165.

Beaumont, M.A., 2010. Approximate Bayesian computation in evolution and ecology. *Annu Rev Ecol Evol Syst*, 41(1), pp.379-406.

Beatty, G.E. & Provan, J., 2010. Refugial persistence and postglacial recolonization of North America by the cold-tolerant herbaceous plant Orthilia secunda. *Molecular Ecology*, 19(22), pp.5009–5021.

Bermingham, E. & Moritz, C., 1998. Comparative phylogeography: concepts and applications. *Molecular Ecology*, 7(4), pp.367–369.

Bivand, R. & Lewin-Koh, N., 2013. maptools: Tools for reading and handling spatial objects. *R package version 0.8-23*.

Iltis, H.H. 1965. The genus Gentianopsis (Gentianaceae): transfers and phytogeographic comments. *Sida, Contributions to Botany* 2, pp.129-154.

Bowen, B.W. et al., 2014. Phylogeography unplugged: comparative surveys in the genomic era. *Bulletin of Marine Science*, 90(1), pp.13–46.

Boyle, B. et al. 2013. The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics 14:16. doi:10.1186/1471-2105-14-16.

Braconnot et al, *The Paleoclimate Modeling Intercomparison Project contribution to CMIP5*, **CLIVAR Exchanges No. 56, Vol. 16, No.2, May 2011, pp 15-19**.

Brigham-Grette, J. & Gualtieri, L.M., 2003. Chlorine-36 and 14C chronology support a limited last glacial maximum across central Chukotka, northeastern Siberia, and no Beringian ice sheet. *Quaternary*, 59, pp.386–398.

Brink, D., 1982. Tuberous Aconitum (Ranunculaceae) of the Continental United States: Morphological Variation, Taxonomy and Disjunction. *Bulletin of the Torrey Botanical Club*, 109(1), p.13.

Brouillet L, et al., 2010+. Database of Vascular Plants of Canada (VASCAN). Online at http://data.canadensys.net/vascan and http://www.gbif.org/dataset/3f8a1297-3259-4700-91fc-acc4170b27ce, released on 2010-12-10. Version [xx]. GBIF key: 3f8a1297-3259-4700-91fc-acc4170b27ce. Data paper ID: doi: http://doi.org/10.3897/phytokeys.25.3100 [accessed on September 09, 2017]

Brunsfeld, S.J. et al., Comparative phylogeography of northwestern North America: a synthesis. *webpages.uidaho.edu*

Bryant, D. et al., 2012. Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Molecular Biology and Evolution*, 29(8), pp.1917–1932.

Cain, M.L., Milligan, B.G. & Strand, A.E., 2000. Long-Distance Seed Dispersal in Plant Populations. *American Journal of Botany*, 87(9), pp.1217–11.

Calabrese, J.M. et al., 2013. Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, 23(1), pp.99–112.

Cane, M.A. et al., 2006. Progress in Paleoclimate Modeling*. *Journal of Climate*, 19(20), pp.5031–5057.

Catling, P.M., 2009. Composition, Phytogeography, and Relict Status of the Vascular Flora of Alvars and Cliff Tops Southwest of Great Slave Lake, Northwest Territories, Canada. *Rhodora*, 111(946), pp.189–208.

Chamberlin, T. C. and R. D. Salisbury. 1906. I.—Geology: Earth History. Vol. II: Genesis-Paleozoic; pp. xxvi, 692, with 306 illustrations. Vol. III: Mesozoio-Cenozoic; pp. xi, 624, with 576 illustrations. John Murray, London. *Geological Magazine*, 3(10), p.472.

Chifman, J. & Kubatko, L., 2014. Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, 30(23), pp.3317–3324.

Cornuet, J.M. & Scholar, G.L.G., *Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data, Genet., 1996, vol. 144*,

Davey, J.W. et al., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7), pp.499–510.

Davey, J.W. et al., 2010. RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, 9(5-6), pp.416–423.

DeChaine, E.G. et al., 2013. Deep genetic divergence between disjunct Refugia in the Arctic-Alpine King's Crown, Rhodiola integrifolia (Crassulaceae). S. Lavergne, ed. *PLoS ONE*, 8(11), p.e79451.

Delcourt, P. A. & Delcourt, H. R. (1981) in *Geobotany II*, ed. Romans, R. C. (Plenum, New York), pp. 123–166.

Devlin-Durante, M.K. & Baums, I.B., 2017. Genome-wide survey of single-nucleotide polymorphisms reveals fine-scale population structure and signs of selection in the threatened Caribbean elkhorn coral, Acropora palmata. *PeerJ*, 5, p.e4077.

Donoghue, M.J. & Moore, B.R., 2003. Toward an integrative historical biogeography. *Integrative and Comparative Biology*, 43(2), pp.261–270.

Donoghue, M.J., Bell, C.D. & Li, J., 2001. Phylogenetic Patterns in Northern Hemisphere Plant Geography. *International Journal of Plant Sciences*, 162(S6), pp.S41–S52.

Dormann, C.F. et al., 2012. Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, 39(12), pp.2119–2131.

Dyke, A.S. & Prest, V.K., 1987. *Paleogeography of northern North America, 18 000-5 000 years ago*,

D'Amen, M., Zimmermann, N.E. & Pearman, P.B., 2012. Conservation of phylogeographic lineages under climate change. *Global Ecology and Biogeography*, 22(1), pp.93–104.

Eaton, D.A.R. et al., 2016. Misconceptions on Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants. *Systematic Biology*, 56, p.syw092.

Edwards, S.V. et al., 2016. Reticulation, divergence, and the phylogeography–phylogenetics continuum. *Proceedings of the National Academy of Sciences*, 113(29), pp.8025–8032.

Ehlers, J. & Gibbard, P.L., 2007. The extent and chronology of Cenozoic Global Glaciation. *Quaternary International*, 164-165, pp.6–20.

Eidesen, P.B. et al., 2007. Nuclear vs. plastid data: complex Pleistocene history of a circumpolar key species. *Molecular Ecology*, 16(18), pp.3902–3925.

Escobar García, P. et al., 2012. Extensive range persistence in peripheral and interior refugia characterizes Pleistocene range dynamics in a widespread Alpine plant species (Senecio carniolicus, Asteraceae). *Molecular Ecology*, 21(5), pp.1255–1270.

Fernald, M.L., 1925. Persistence of Plants in Unglaciated Areas of Boreal America. *Memoirs of the American Academy of Arts and Sciences*, 15(3), p.239.

Fick, S.E. & Hijmans, R.J., 2017. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12), pp.4302-4315.

Flora of North America Editorial Committee, eds. 1993+. Flora of North America North of Mexico. 20+ vols. New York and Oxford.

Friis, G. et al., 2016. Rapid postglacial diversification and long-term stasis within the songbird genus Junco: phylogeographic and phylogenomic evidence. *Molecular Ecology*, 25(24), pp.6175–6195.

Gavin, D.G. et al., 2014. Climate refugia: joint inference from fossil records, species distribution models and phylogeography. *New Phytologist*, 204(1), pp.37–54.

Givnish, T.J. & Renner, S.S., 2004. Tropical Intercontinental Disjunctions: Gondwana Breakup, Immigration from the Boreotropics, and Transoceanic Dispersal. *International Journal of Plant Sciences*, 165(S4), pp.S1–S6.

Givnish, T.J. et al., 2004. Ancient Vicariance or Recent Long-Distance Dispersal? Inferences about Phylogeny and South American–African Disjunctions in Rapateaceae and Bromeliaceae Based on ndhF Sequence Data. *International Journal of Plant Sciences*, 165(S4), pp.S35–S54.

Godbout, J. et al., 2005. A mitochondrial DNA minisatellite reveals the postglacial history of jack pine (Pinus banksiana), a broad-range North American conifer. *Molecular Ecology*, 14(11), pp.3497–3512.

Gugger, P.F. et al., 2011. Southward Pleistocene migration of Douglas-fir into Mexico: phylogeography, ecological niche modeling, and conservation of "rear edge" populations. 189, pp.1185–1199.

Guisan, A. & Rahbek, C., 2011. SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, 38(8), pp.1433–1444.

Hewitt, G.M., 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. (September 1994), pp.247–276.

Hickerson, M.J. et al., 2010. Phylogeography's past, present, and future: 10 years after Avise, 2000. *Molecular Phylogenetics and Evolution*, 54(1), pp.291–301.

Higgins, S.I. & Richardson, D.M., 1999. Predicting Plant Migration Rates in a Changing World: The Role of Long-Distance Dispersal. *The American Naturalist*, 153(5), pp.464–475.

Hijmans, R.J. & Elith, J., 2013. Species distribution modeling with R. *R package version 0.8-11*.

Hijmans, R.J. & van Etten, J., 2014. raster: Geographic data analysis and modeling. *R package version 2.8*.

Hipp, A.L. et al., 2018. Sympatric parallel diversification of major oak clades in the Americas and the origins of Mexican species diversity. *New Phytologist*, 217(1), pp.439–452.

Ho, S.Y.W. & SHAPIRO, B., 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources*, 11(3), pp.423–434.

Holliday, V. T., Knox, J. C., Running, G. L., IV, Mandel, R. D. & Ferring, C. R. (2002) in *The Physical Geography of North America*, ed. Orme, A. R. (Oxford Univ. Press, Oxford, U.K.), pp. 335–362.

Intergovernmental Panel on Climate Change (IPCC), www.ipcc.ch. 5th IPCC Report, 2013.

J, M.R. & G, V.E., 1981. Distributions of some western North American plants disjunct in the Great Lakes region. *Michigan Botanist*.

Jackson, S.T. & Overpeck, J.T., 2000. Responses of plant populations and communities to environmental changes of the late Quaternary. *Paleobiology*, 26(sp4), pp.194–220.

Jaramillo-Correa, J.P., Beaulieu, J. & Bousquet, J., 2004. Variation in mitochondrial DNA reveals multiple distant glacial refugia in black spruce (Picea mariana), a transcontinental North American conifer. *Molecular Ecology*, 13(9), pp.2735–2747.

Joy, J.B. et al., 2016. Ancestral Reconstruction. *PLoS computational biology*, 12(7), p.e1004763.

Kartesz, J.T., The Biota of North America Program (BONAP). 2015. *North American Plant Atlas*. (http://bonap.net/napa). Chapel Hill, N.C. [maps generated from Kartesz, J.T. 2015. Floristic Synthesis of North America, Version 1.0. Biota of North America Program (BONAP).

Kageyama, M., et al., 2018. PMIP4-CMIP6: the contribution of the Paleoclimate Modelling Intercomparison Project to CMIP6. *Geoscientific Model Development Discussions*, 11(3) pp. 1033-1057.

Lee-Yaw, J.A., Irwin, J.T. & Green, D.M., 2008. Postglacial range expansion from northern refugia by the wood frog, Rana sylvatica. *Molecular Ecology*, 17(3), pp.867–884.

Li, P. et al., 2013. Phylogeography of North American herbaceous Smilax (Smilacaceae): Combined AFLP and cpDNA data support a northern refugium in the Driftless Area. *American Journal of Botany*, 100(4), pp.801–814.

Li, R. & Wen, J., 2013. Phylogeny and Biogeography of Dendropanax (Araliaceae), an Amphi-Pacific Disjunct Genus Between Tropical/Subtropical Asia and the Neotropics. *Systematic Botany*.

Ludwig, P. et al., 2018. Perspectives of regional paleoclimate modeling. *Annals of the New York Academy of Sciences*, 5, p.735.

Marr, K.L., Allen, G.A. & Hebda, R.J., 2008. Refugia in the Cordilleran ice sheet of western North America: chloroplast DNA diversity in the Arctic-alpine plant Oxyria digyna. *Journal of Biogeography*, 35(7), pp.1323–1334.

Marth, G.T. et al., 2004. The Allele Frequency Spectrum in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in Three Large World Populations. *Genetics*, 166(1), pp.351–372.

Meyer, E.M., Peterson, A.T. & Hargrove, W.W., 2004. Ecological niches as stable distributional constraints on mammal species, with implications for Pleistocene extinctions and climate change projections for biodiversity. *Global Ecology and Biogeography*, 13(4), pp.305–314.

Miller, N.G., 1979. Boreal and western North American plants in the late Pleistocene of Vermont. *JSTOR*

Narum, S.R. et al., 2013. Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, 22(11), pp.2841–2847.

Nimis, P.L. et al., 1998. A multivariate phytogeographic analysis of plant diversity in the Putorana Plateau (N. Siberia). *Opera Botanica*, 136.

Nogués-Bravo, D., 2009. Predicting the past distribution of species climatic niches. *Global Ecology and Biogeography*, 18(5), pp.521–531.

Orr, H.A., 1995. The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics*, 139(4), pp.1805–1813.

Page, R.D.M., 1990. Temporal Congruence and Cladistic Analysis of Biogeography and Cospeciation. *Systematic Zoology*, 39(3), p.205.

Papadopoulou, A. & Knowles, L.L., 2016. Toward a paradigm shift in comparative phylogeography driven by trait-based hypotheses. *Proceedings of the National Academy of Sciences*, 113(29), pp.8018–8024.

Pettengill, J.B. & Moeller, D.A., 2012. Phylogeography of speciation: allopatric divergence and secondary contact between outcrossing and selfing Clarkia. *Molecular Ecology*, 21(18), pp.4578–4592.

Phillips, A. & Davis, M., 2006. *Tags for Identifying Languages*, RFC Editor.

Poncet, V. et al., 2013. Phylogeography and niche modelling of the relict plant Amborella trichopoda (Amborellaceae) reveal multiple Pleistocene refugia in New Caledonia. *Molecular Ecology*, 22(24), pp.6163–6178.

Provan, J. & Beatty, G.E., 2011. Phylogeographic analysis of North American populations of the parasitic herbaceous plant Monotropa hypopitys L. reveals a complex history of range expansion from multiple late glacial refugia. *Journal of Biogeography*, 38(8), pp.1585–1599.

Provan, J. & Beatty, G.E., 2014. Phylogeographical analysis of two cold-tolerant plants with disjunct Lusitanian distributions does not support in situsurvival during the last glaciation H. P. Comes, ed. *Journal of Biogeography*, 41(11), pp.2185–2193.

Provan, J. & Bennett, K.D., 2008. Phylogeographic insights into cryptic glacial refugia. *Trends in Ecology & Evolution*, 23(10), pp.564–571.

Pusateri, W.P., Iowa, D.R.J.O.T.1993, Habitat and distribution of plants special to Iowa's Driftless Area. *scholarworks.uni.edu*

QGIS Development Team (2018). QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org.

Radosavljevic, A. & Anderson, R.P., 2013. Making better Maxent models of species distributions: complexity, overfitting and evaluation M. Araújo, ed. *Journal of Biogeography*, 41(4), pp.629–643.

Rebernig, C.A. et al., 2010. Multiple Pleistocene refugia and Holocene range expansion of an abundant southwestern American desert plant species (Melampodium leucanthum, Asteraceae). *Molecular Ecology*, 19(16), pp.3421–3443.

Rovito, S.M. & Schoville, S.D., 2017. Testing models of refugial isolation, colonization and population connectivity in two species of montane salamanders. *Heredity*, 119(4), pp.265–274.

Rowe, K.C. et al., 2004. Surviving the ice: Northern refugia and postglacial colonization. *Proceedings of the National Academy of Sciences*, 101(28), pp.10355–10359.

Samonds, K.E. et al., 2012. Spatial and temporal arrival patterns of Madagascar's vertebrate fauna explained by distance, ocean currents, and ancestor type. *Proceedings of the National Academy of Sciences*, 109(14), pp.5352–5357.

Satler, J.D. & Carstens, B.C., 2016. Phylogeographic concordance factors quantify phylogeographic congruence among co-distributed species in the Sarracenia alatapitcher plant system. *Evolution*, 70(5), pp.1105–1119.

Schmitt, T., 2007. Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology*, 4(1), p.11.

Schneeweiss, G.M. & Schönswetter, P., 2010. The wide but disjunct range of the European mountain plant Androsace lactea L. (Primulaceae) reflects Late Pleistocene range fragmentation and post-glacial distributional stasis. *Journal of Biogeography*, 18(10).

Selkoe, K.A. & Toonen, R.J., 2006. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, 9(5), pp.615–629.

Servedio, M.R. et al., The role of reinforcement in speciation: theory and data. *annualreviews.org*

Shafer, A.B.A. et al., 2010. Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Molecular Ecology*, 19(21), pp.4589–4621.

Smith, H.T.U., 1949. Periglacial Features in the Driftless Area of Southern Wisconsin. *The Journal of Geology*, 57(2), pp.196–215.

Soltis, D.E. et al., 1997. Chloroplast DNA intraspecific phylogeography of plants from the Pacific Northwest of North America. *Plant Systematics and Evolution*, 206(1-4), pp.353–373.

Soltis, D.E. et al., 2006. Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, 15(14), pp.4261–4293.

Spalink, D. et al. Spatial phylogenetics reveals evolutionary constraints on the assembly of a large regional flora. *In press*.

Stubbs, R.L. et al., 2018. Pseudo-parallel patterns of disjunctions in an Arctic-alpine plant lineage. *Molecular Phylogenetics and Evolution*, 123, pp.88–100.

Sueyoshi, T. et al., 2013. Set-up of the PMIP3 paleoclimate experiments conducted using an Earth system model, MIROC-ESM. *Geoscientific Model Development*, 6(3), pp.819–836.

Swofford et al., *PAUP* Phylogeny Analysis Using Parsimony (* and other methods), version 40b10*.

Taberlet, P. et al., 1998a. Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology*, 7(4), pp.453–464.

Taberlet, P. et al., 1998b. Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology*, 7(4), pp.453–464.

Thuiller, W. et al., 2009. BIOMOD - a platform for ensemble forecasting of species distributions. *Ecography*, 32(3), pp.369–373.

Thule, E.H.S.B.-L.A.1937, *Outline of the history of Boreal and Arctic biota during the Quaternary Period*,

USDA, NRCS. 2018. The PLANTS Database (http://plants.usda.gov, 26 June 2018). National Plant Data Team, Greensboro, NC 27401-4901 USA.

Vitt, P. et al., 2010. Assisted migration of plants: Changes in latitudes, changes in attitudes. *Biological Conservation*.

Wang, L. et al., 2009. History and evolution of alpine plants endemic to the Qinghai-Tibetan Plateau: Aconitum gymnandrum (Ranunculaceae). *Molecular Ecology*, 18(4), pp.709–721.

Watanabe, S. et al., 2011. MIROC-ESM: model description and basic results of CMIP5-20c3m experiments. *Geoscientific Model Development Discussions*, 4(2), pp.1063–1128.

Wen, J. & BOND, S.M.I., 2009. Evolution of the Madrean–Tethyan disjunctions and the North and South American amphitropical disjunctions in plants. *Journal of Systematics and Evolution*, 47(5), pp.331–348.

Wen, J. et al., 2013. Biogeography: Where do we go from here?, 62(October), pp.912–927.

Whittaker, R.J. et al., 2013. The geographical distribution of life and the problem of regionalization: 100 years after Alfred Russel Wallace. *Journal of Biogeography*, 40(12), pp.2209–2214.

Wiens, J.J., 2004. What Is Speciation and How Should We Study It? *The American Naturalist*, 163(6), pp.914–923.

Xiang, Q.Y.J. & Soltis, D.E., 2001. Dispersal-Vicariance Analyses of Intercontinental Disjuncts: Historical Biogeographical Implications for Angiosperms in the Northern Hemisphere. *International Journal of Plant Sciences*, 162(S6), pp.S29–S39.

CHAPTER 2: Molecular evidence reveals the Black Hills did not serve as a stepping stone for western North American and Great Lakes disjunct *Rubus parviflorus* (Rosaceae)

ABSTRACT

The flora of the Great Lakes region (GLR) contains species that are more abundant and widespread west of the Rocky Mountains (see Chapter 1), and these disjunct, unique species of the GLR have been of interest to botanists for decades (Marquis and Voss 1951). *Rubus parviflorus* Nutt. (thimbleberry, Rosaceae) is native to North America, and exhibits this distribution with an additional occurrence in the Black Hills of South Dakota, a region which has been considered a potential dispersal "stepping stone" for disjunct populations. Here we test this stepping stone model of *R. parviflorus* with phylogenetic dating, phylogeographic analysis, and population genetics. Our results show that *R. parviflorus* predates the last glacial cycles of the Pleistocene and was subject to the forces of advancing and retreating ice sheets. We show that the Black Hills populations did not serve as a dispersal stepping stone from the West; in fact, they share a more recent common ancestry with western populations. We also show that the GLR populations are more closely related to those from British Columbia in northwest North America. This finding supports a relationship previously proposed based on morphological data, and points to the occurrence of a more northern post-glacial dispersal route. Our population genetic analysis demonstrates that the more isolated GLR and Black Hills populations are not lower in genetic diversity than western North American populations, reinforcing the scenario of vicariance of a post-glacial route rather than persistence in a glacial refugium or a long-distance dispersal event which would have had resulted in a genetic bottleneck.

INTRODUCTION

Botanists have long been interested in the geographic distributions of plants across North America because they speak to phenomena such as trait variation and plasticity, species histories, dispersal vs. vicariance, and geological/ecological forces (e.g. Hultén 1937, Shreve 1942, Axelrod 1950, Stebbins 1950, 1970, Teeri and Stowe 1976, Soltis 2006, Baldwin 2014). One starting point for teasing apart these phenomena is to investigate regional patterns of biodiversity. By examining a regional flora, we can ask questions such as, how did the species arrive there, what forces impacted their broader distributions, and what, if any, forces impede their persistence there? The flora of the Great Lakes Region (GLR) in North America is of particular interest with regard to these questions because it contains over 100 enigmatic, co-occurring disjunct taxa from western North America (Marquis and Voss 1981, Chapter 1). Here we focus on one example from this disjunct distribution, *Rubus parviflorus* (Nutt.) (Rosaceae, subg. *Anoplobatus*), that has an additional occurrence in the Black hills of South Dakota, to illuminate its phylogeographic history, uncover the forces that have impacted its distribution, assess its future persistence in the GLR, and deepen our understanding of the GLR flora as a whole.

Studies of plant phylogeography in North America have to contend with the Pleistocene glacial cycles. These global glacial cycles caused many biogeographic interruptions across plants and animals (Deevey 1949, Schmitt 2007, Shafer et al. 2010), and the non-uniform, marching and receding ice sheets in North America formed pockets of refugia (e.g. northern Alaska and the "Driftless Area" of Wisconsin and Iowa, North America south of the ice front) and broke up otherwise continuous distributions of species (Hultén 1937, Nimis et al. 1998, Swenson and Howard 2005, Beatty and Provan 2010). The disjunct species of the GLR in particular, are

suggested to be the result of re-colonization from local Pleistocene glacial refugia in the Driftless Area, or of new, post-glacial dispersal as ice retreated (Iltis 1965, Marquis and Voss 1981). In the case of *R. parviflorus*, there are additional populations in the Black Hills of South Dakota, which might be remnants of a dispersal "stepping stone" to the GLR.

To test the plausibility of these Pleistocene-based hypotheses for *R. parviflorus* we first need to determine its age. Recent papers on the Rosaceae family have provided age estimates for *Rubus* (Xiang et al. 2017, Zhang et al. 2017), but the placement of *Rubus* within Rosoideae has been problematic. It has been placed as sister to the Colurieae clade (*Geum*, *Fillugia*, *Waldsteinia*), sister to the Roperculina clade (*Rosa*, Agrimonieae, and Potentilleae), which is sister to the Colurieae clade, or sister to Colurieae + Roperculina (Eriksson et al. 1998, Smedmark and Eriksson 2002, Erikkson et al. 2003, Potter et al. 2007). Because support values in combined ITS and chloroplast analyses have consistently been lower than for each data set alone, it has been suggested that there is genomic incongruence within *Rubus* (Eriksson et al. 2003). This signature of this genomic incongruence has persisted as genetic data sets have expanded and Maximum Likelihood and Bayesian statistics have been implemented. Zhang et al. (2017) places three monophyletic *Rubus* species sister to Roperculina with high support based on plastome data, and Xiang et al. (2017) places four monophyletic *Rubus* species sister to Colurieae + Roperculina based on 113 concatenated, low-copy nuclear genes. This genomic incongruence is not surprising for *Rubus*, given a high frequency of hybridization within the genus, which could affect the placement of its sister clade. The most recent phylogeny of *Rubus* as a whole used ITS in a parsimony framework and was largely unresolved (Alice and Campbell 1999). This phylogeny placed *R. parviflorus* sister to *R. odoratus* within the *Anoplobatus* clade with high support. It placed the *Anoplobatus* clade sister to the major *Rubus* clades and

embedded *Dalibarda* within *Rubus*. We propose to expand this *Rubus*-wide phylogeny to date the stem and crown of *R. parviflorus*, testing its persistence through the Pleistocene glacial cycles in North America.

Once we place *R. parviflorus* in the context of the Pleistocene, we can assess its phylogeographic history. Previous morphological analysis of *R. parviflorus* (Fassett 1941) revealed testable phylogeographic patterns. By comparing which varieties were found in different geographic regions Fassett (1941) showed that the Black Hills populations had a stronger affinity with the western populations, and that the GLR populations had a stronger affinity with populations from Oregon and the Sierra Nevada. By comparing the number of varieties found in each region, Fassett (1941) showed that the GLR populations exhibited similar variation as do their western counterparts, but at higher uniformity. Based on the geographic affinities, the comparable amount of variation in the GLR, and the cool, mesic habitat that *R. parviflorus* currently occupies, Fassett (1941) suggested that the GLR populations came across Canada during a cool, humid post-glacial period. The fruit of *R. parviflorus* is an aggregate berry and it is animal dispersed, however, making long-distance dispersal unlikely. There is some evidence that pockets of deciduous forest persisted in the Driftless Area of Wisconsin and Iowa in addition to the more well-recognized tundra habitat during the Last Glacial Maximum (LGM) (Jackson and Overpeck 2000). Had they existed, these pockets could have supported *R. parviflorus* in a refugium. The hypotheses that Fassett proposed remain to be tested with genetic data. The prominent questions are whether the Black Hills populations fit a "stepping stone" model of dispersal, and whether the GLR populations show signatures of a genetic bottleneck, indicating a founder event or persistence in a glacial refugium. Recent advances in sequencing technology are making SNP data easier and cheaper to generate, and these data can illuminate

population structure, speciation events, geographic variation, and biogeographic history at the level from within species to within genera (Wagner et al. 2012, Eaton and Ree 2013, Dillenberger and Kadereit 2017). We use genetic and genomic methods in *R. parviflorus* to address these questions and evaluate the species migration history and genetic diversity in the GLR.

METHODS

Sampling: We collected leaf samples from 15-20 individuals in 19 populations across the North American range of *R. parviflorus*. We did not include European populations of *R. parviflorus* because the European variety is introduced and cultivated (Clapham et al. 1962). Sampling was completed within five field seasons (2013-2017) and included three collections from collaborators. Because *R. parviflorus* grows in patches and determination of individuals is often obscured by its rhizomatous growth pattern, leaf tissue was collected at 5-meter intervals along a linear transect to avoid sampling potential clones. Samples were silica-dried. One herbarium voucher specimen was collected for each population and placed in the Wisconsin State Herbarium (WIS). We obtained leaf tissue from an individual *R. odoratus* from Ontario to avoid any potential hybrid sample from the range overlap in the GLR, although historical hybridization cannot be ruled out (see results and discussion). *Rubus odoratus* was used as an outgroup for rooting trees.

DNA Extraction: All samples were disrupted with a Tissuelyser (Qiagen, Valencia, CA, USA) using tungsten-carbide beads, and DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA) following the manufacturer's protocol, or a modified CTAB protocol

(Modified from Doyle and Doyle 1990). DNA integrity was confirmed by presence of high-mass bands on a 0.8% agarose gel, and DNA quantification was conducted with the Qubit dsDNA High-Sensitivity Assay Kit (Invitrogen, USA) using a Qubit Reader (Thermofisher, USA) or the AccuBlue High Sensitivity dsDNA Kit (Biotium, USA) using a plate reader at UW-Madison Biotechnology Center.

Rubus *chronogram*: *Multi-gene alignments:* We used PyPHLAWD (Smith and Brown 2018) to extract all available sequences from the Plant and Fungal NCBI database and build clusters of loci across *Rubus*, retaining one accession per species. Our clustering specifications were kept at the default settings, including sequence length >= 600, clustering length limit = 0.65, and percent identity = 20. We used clusters of ITS (ITS1 + 5.8S + ITS2 + 26S) and 7 chloroplast DNA (cpDNA) regions (*matK*, *trnL-trnF*, *rbcL*, *rpl20-rps12*, *trnG-trnS*, *rpl16*, *ndhF*) that had at least 30 species, and the cluster FASTA files were aligned using MUSCLE in Geneious v.10.1 (http://www.geneious.com, Kearse et al. 2012). We added three outgroups (*Fragaria virginiana*, *Rosa arvensis*, and *Agrimonia pilosa*), all from the Roperculina clade, to allow the outgroup to be monophyletic. Species were removed from the concatenated chloroplast alignment if they had > 50% missing data compared to the ungapped alignment consensus sequence.

Sanger Sequencing: We amplified ITS for *R. parviflorus* samples from each of the major geographic regions. Sequencing was conducted at the UW-Madison Biotech Center on an ABI 3730x1 DNA Analyzer. We built assemblies, edited sequences, and re-aligned our sequences to the nrITS alignment in Geneious.

Chronogram calibration: We used a secondary crown date for *Rubus*: 23 Mya at the Oligocene-Miocene boundary based on the Xiang et al. (2017) Rosaceae phylogeny of 148

species which used 113 concatenated low-copy nuclear genes and 19 fossil calibrations. This phylogeny places *Rubus* sister to Colurieae + Roperculina. We used a secondary stem date for *Rubus* (57.84 Mya [95% HPD 56.93-65.66]) based on the Zhang et al. (2017) Rosaceae cpDNA phylogeny of 1-tip per 87 genera, and 1 secondary and 6 fossil calibrations. This stem age is based on divergence from the Roperculina clade, which is sister to *Rubus* with high support. We did not test the secondary date from the Chin et al. (2014) Rosaceae phylogeny because of the low support within Rosoideae.

Phylogenies: We ran the concatenated cpDNA data set and the ITS data sets separately and combined with partitions using RAxML v.8.2.10 (Stamatakis 2014) and MrBayes v.3.2.6 (Ronquist et al. 2012) on the CIPRES platform (Miller et al. 2010). In RAxML we used the GTR model of nucleotide substitution under a Γ model of rate heterogeneity with alpha estimated, and we used the rapid hill-climbing algorithm. In MrBayes, we were required to specify one outgroup, *Agrimonia pilosa*, and we ran a GTR model with 1 million generations and 25% burn-in. As the incongruence in the placement of *Rubus* between the highly supported topologies in Xiang et al. 2017 and Zhang et al. 2017 might be due to different histories in the chloroplast and nuclear genomes, we ran each data set separately with its respective calibration in BEAST v.2.4.6 (Bouckaert et al. 2014) on the CIPRES platform. We also ran a partitioned analysis on the combined data set using both calibrations despite their dates being derived from conflicting topologies. We justified the use of both dates in the same analysis because of the short branch that splits *Rubus*, Colurieae and Roperculina. BEAST analyses were run with GTR + Γ, a relaxed lognormal clock model, and a yule birth model. Tree likelihood calculation was set to treat ambiguities as equally likely. Secondary dates were given a normal distribution and the outgroups, *Rubus*, and *R. parviflorus* were each constrained to be monophyletic. We ran 10

million generations, evaluated the posterior distributions of the priors in Tracer v.1.6.0 (Rambaut et al. 2018), and calculated the maximum clade credibility tree using TreeAnnotator v.2.4.6 (Bouckaert et al. 2014) with 25% burn-in.

Genotyping-by-sequencing: We selected 4-9 individuals per population for genotyping-by-sequencing and included one sample of *R. odoratus*. Sample library prep and sequencing were conducted at the UW-Madison Biotechnology Center using the *ApeKI* restriction digest enzyme and pooling samples 48-plex on 3 lanes of the Illumina Hi-Seq 2500 platform.

Bioinformatics: We ran cutadapt v.1.13 (Martin 2011) to trim off common adapters at the 3' end and retain only full-length reads. We ran ipyrad v.0.7.13 (Eaton 2014) steps 1 and 2 to demultiplex samples allowing 0 mismatches in barcodes, remove barcodes, then filter out reads with any low-quality bases (phred Q score < 20). We ran cutadapt again to trim off the 4-base pair overhang at the 5' end, resulting in 89bp read length. We removed populations with very low mean read coverage, and 14 populations remained in the final dataset (Table 1). We assembled reads by running ipyrad and Stacks (Catchen et al. 2013) on the Condor computing cluster at the Center for High Throughput Computing facility at UW-Madison. We ran ipyrad in a Docker container (Merkel 2014) to ensure access to required dependencies.

Optimizing assembly parameters: We tested 3 influential clustering parameters as described in Paris et al. (2017) and Rochette and Catchen (2017) (Stacks), and in Anderson et al. (2017) (PyRAD): minimum depth to make a statistical base call (m), maximum number of mismatches allowed when forming a locus within an individual (M), and maximum number of mismatches allowed when aligning loci between individuals (n). We tested values 1-10 for the minimum depth to make a statistical base call. Results from this revealed that our data were at

lower than 20X coverage at the default value (m = 2) (Appendix 1) and as expected, coverage was higher when requiring a higher minimum depth to retain a locus. Based on suggestions from the Stacks developer (pers. comm, Stacks Googlegroup) data with lower coverage should not be forced to a higher depth, and m should be kept at 2 or 3 to allow for statistical calls. We used m = 3, and tested M = n for values 1-6. We chose M = n = 3. The m, M, and n parameter values were selected based on maximizing the number of loci formed, stabilizing the proportion of loci recovered for n-SNPs/locus, and minimizing the proportion of heterozygotes at each base pair (Appendix 1).

Phylogeography: To resolve the relationships between populations of *R. parviflorus* and uncover its phylogeographic history, we estimated phylogenies using concatenated SNP data with Maximum Likelihood (ML) in RAxML and a multi-species coalescent approach in SVDQuartets v.1.0 (Chifman and Kubatko 2014) in PAUP v.4.0a150 (Swofford 2002).

The issue of how to handle allelic data from genomic data sets for use in phylogenetics is currently being debated (Kates et al. 2018, Andermann et al. 2018). Use of allelic data is not an issue for coalescent approaches, which assess variation and topology at each locus individually. However, it is a potential issue for concatenated loci from *de novo* data sets, which is the common approach for obtaining a ML tree with GBS data (i.e., Dillenberger and Kadereit 2017, Anderson et al. 2017). Although algorithms for allele phasing within loci and among gene copies have been implemented in various assembly programs (i.e., HybPiper, Johnson et al. 2016; PyRAD, Eaton 2014; and Stacks, Catchen et al. 2013), allele phasing across concatenated loci in *de novo* data sets is still an unresolved issue. One approach to overcome this for GBS data is to reduce allelic data to one summary sequence. Heterozygous sites can be coded with IUPAC

ambiguity codes, which can result in a highly ambiguous data matrix for highly heterozygous individuals (Lemmon et al. 2009). Alternatively, individuals can be grouped by population and only loci that are fixed-within and variant-among populations are retained. This second approach reduces the amount of ambiguity, but also significantly reduces the variation available for phylogenetic inference. There are other potential ways to handle across-locus phasing, for example parametrically sampling an allele at each locus based on its frequency within a population. For our ML analyses we tested using ambiguity codes for each individual (INDIV), and we took a majority-rule consensus approach (CONS), which is less conservative than the fixed-within and variant-among populations approach. Both of these approaches eliminate the issue of phasing alleles across concatenated loci, and in addition, the majority-rule approach reduces ambiguity.

RAxML: Because raw, concatenated SNP data have been shown to overestimate branch lengths (Leaché et al. 2015), we chose to implement the Lewis ascertainment bias correction in RAxML to account for unobserved invariant sites. This correction requires that the data contain no invariant sites, and it interprets ambiguities as either of the two base calls, rendering some of them invariant. To get the INDIV data set, we ran the populations module in Stacks on all samples to get an output of loci present in at least 75% of individuals, concatenating only one SNP per locus. We removed any samples that had >50% missing data and any samples that were the only representative from a population. Because the populations module includes Ns as variant sites, we changed Ns to gaps and masked invariant sites in Geneious. The remainder of the ambiguous base calls were not considered invariant by the ascertainment correction, so we left them in. We changed the gaps back to Ns and ran RAxML on the CIPRES platform with the Lewis ascertainment bias correction, the GTR model of nucleotide substitution under a Γ model

of rate heterogeneity with alpha estimated, and the rapid hill-climbing algorithm. We conducted 100 GTRCAT rapid bootstrap replicates. For the CONS data set, we modified the INDIV data set and extracted the majority rule consensus sequence for each population. We combined them without realigning since they were all of uniform length. We replaced Ns with gaps then masked invariant sites, and because some ambiguous sites were called as invariant, we also masked ambiguous sites. This resulted in a very short alignment, so we also ran the CONS alignment without masking and without the ascertainment bias correction. We replaced gaps with Ns and ran both data sets in RAxML with the same settings.

SVDQuartets: To account for any incomplete lineage sorting (ILS), we used SVDQuartets in Paup to estimate the history of each locus individually and obtain a multi-species coalescent tree. This program is robust to missing data, unlike other multi-species coalescent programs that use SNP data such as SNAPP (Bryant et al. 2012). It is comparable to ASTRAL-2 (Mirarab and Warnow 2015) and NJst (Liu and Yu 2011) for estimating the correct tree with SNP data in the face of low ILS (Chou et al. 2015). It also handles ambiguities by proportionally distributing the probability of allelic variants in the Q-matrix of each quartet, across all sites. This allows us to include real allelic ambiguity. We ran populations with and without the outgroup to get outputs of loci present in at least 75% of individuals, retaining one SNP per locus. For the data set without the outgroup we ran SVDQuartets in Paup with each sample as a lineage, which looks at quartets across all individuals, and we ran it with multiple samples assigned to a locality population, which samples only once from a population for each quartet. We rooted the SVDQuartets tree without the outgroup using the topology from the RAxML analysis. For the data set with the outgroup we ran SVDQuartets assigning samples to populations. For each analysis, we used the multi-species coalescent tree model to evaluate all

possible quartets and used the QFM quartet assembly to summarize quartets and obtain the bootstrap majority rule tree (Reaz et al. 2014). We distributed heterozygote ambiguity instead of masking it and we ran 100 bootstrap replicates on subsets of 100,000 quartets.

Population structure and genetic diversity: To estimate population structure and identify potential admixture between populations, we used Structure (Pritchard et al. 2000). This program uses a Bayesian model to leverage the assumptions of Hardy-Weinberg equilibrium and linkage equilibrium between loci to construct probabilities of populations assignment for each individual, given a number of populations (K). We chose this method over a principal-components analysis (PCA) approach for visualizing structure because it explicitly estimates global ancestry, which is not easily interpretable from PCA coordinates (Novembre and Stephens 2008). We ran populations in Stacks without the outgroup and retained only loci that were present in 100% of the populations and in at least 75% of individuals within each population. We retained one SNP per locus to maximize independence between SNP loci. We ran fastStructure with the logistic setting to detect any subtle structure that might otherwise be swamped out by the structure from large geographic isolation between populations (Raj et al. 2014). We tested K values 2-10 with 10^6 generations and ran 5 iterations. Following Barker et al. (2017) we tested consistency between runs of the same K value using CLUMPP v 1.1.2 (Jackobsson and Rosenberg 2007) to estimate the highest value of pairwise similarity (H' value). In CLUMPP we used the Greedy algorithm with 1000 random input orders for K 2-7 and the LargeKGreedy algorithm for K 8-10. For each of the five iterations, we ran the chooseK module in fastStructure, which reports the K values that give the best marginal likelihood score and the number of populations used to explain the structure in the data. We selected the most common best K value reported from the five

iterations and plotted the results of the mean population assignment determined in CLUMPP for that K value. We plotted the results using distruct2.2 (courtesy of Chhatre, 2016).

To test the hypothesis of a population bottleneck due to a founder event or restriction in a Driftless Area refugium, we measured the genetic diversity of the populations in the GLR. We looked at nucleotide diversity (π), expected and observed heterozygosity (H_e and H_o), the inbreeding coefficient (F_{IS}), and the number of private alleles for each population. Although the site frequency spectrum and Tajima's D are common ways to detect bottleneck events (Luikart and Cornuet 1998, Marth et al. 2004), we did not use these methods due to our uncertainty in interpreting the alternative to the null hypothesis of neutral evolution. We also used these genetic diversity metrics to assess whether there are threats to future persistence of the populations in the GLR or the Black Hills. We ran the populations module in Stacks, grouping samples by their collection locality and retaining only loci that were present in 100% of the populations and at least 75% of individuals within a population. We ran diversity statistics in the populations module using kernel-smoothing and 100 bootstrap replicates. The F_{IS} statistic was calculated with the Bonferroni correction using a sliding window with a base p-value of 0.05. This correction is more conservative than a p-value correction, so it reduces Type I errors, but may introduce more Type II errors. We performed unpaired, two-sample t-tests to compare mean values of each statistic between populations in the GLR and populations in the west as well as populations in the Black Hills and populations in the west.

RESULTS

Rubus *chronogram:* The chloroplast concatenated alignment was 7520bp with 110 tips and the ITS alignment was 924bp with 212 tips. The combined alignment was 8444bp with 216 tips. The

RAxML and MrBayes analyses for ITS and the chloroplast separately revealed high support for the Anoplobatus clade containing *R. parviflorus*, *R. odoratus*, *R. deliciosus*, *R. neomexicanus*, and *R. trilobus* with weak support across the rest of the tree (not shown). The combined data set showed similarly high support for the Anoplobatus clade (Appendix 2a-b). The dates from the separate cpDNA and ITS BEAST analyses were comparable to the combined data set, which had higher support values, so we present only the tree from the combined data set (Fig. 1). The crown age of *R. parviflorus* is pre-Pleistocene and ~6.25 Mya (3.3-9.5 Mya 95% HPD). The stem age is ~7.5 Mya (3.75-10.9 Mya 95% HPD). Support values for individual *R. parviflorus* populations are very low, so these ages are not interpretable.

GBS Bioinformatics: With optimized parameters, our samples had a mean read depth of 13X (range: 6X-35X). The coverage inconsistency is not associated with population assignment or date of extraction (results not shown). Because sample concentrations are standardized before pooling, it is more likely due to quality of the DNA sample, or the digestion of a large genome that can result in shallow read depths.

Phylogeography: *RAxML:* After filtering samples with >50% missing data and samples that were the only representative of a population and masking invariant sites, the INDIV data set contained 27,783 concatenated SNPs. After similar filtering and masking invariant and ambiguous sites, the CONS data set contained 2,628 concatenated SNPs. Without masking, the CONS data set contained 28,512 concatenated SNPs. The INDIV data set resulted in monophyly of only a few of our locality populations, but it resolved three major clades with high support (Fig. 2). Notably, the Colorado, Black Hills and Idaho populations are supported as a clade sister

to the rest of the populations. The western populations furthest south in California and Oregon form a clade sister to the remainder of the Pacific northwest (PNW) and north northwest (NNW) populations, which have high support for monophyly. And finally, the populations from the GLR are a highly supported clade embedded within the NNW populations. The CONS data set with masking of invariant and ambiguous sites and ascertainment bias correction revealed a similar pattern with high support for the same clades, but low support throughout the rest of the tree, and low support for non-monophyly of the GLR populations (Fig. 3a). Low support values for this data set might be the result of the strict filtering process that removed ambiguous base calls and reduced the length of the data set. However, the CONS data set that used all SNPs without masking recovered the monophyly of the GLR populations and had high support for the three major clades, but there was little improvement for the support of the rest of the topology (Fig. 3b).

SVDQuartets: The data sets without any masking contained 28,512 and 29,350 SNPs with and without the outgroup, respectively. For the population assignment trees, there were minor shifts in the topology when the outgroup was added (not shown). The topology was similar between the individual and population trees that did not include the outgroup (not shown). Given that we do not expect the topology to change from adding an outgroup unless there are extraneous issues with the outgroup, we present the population assignment tree without the outgroup, but rooted based on the RAxML topology (Fig. 4). This tree has a different topology than the RAxML trees, and the low support values likely reflect conflict between SNPs. The ancestral range is in Washington, and there are two major clades: one with dispersal to British Columbia and a dispersal to Oregon and then the Black Hills, and one with divergence between populations in British Columbia and the GLR. If the major branches with low support

were collapsed and rearranged, our inference of the history of these populations would not change.

Population structure and genetic diversity: Population assignments were inconsistent between the five iterations of each K value, except K = 2 (Appendix 3). The chooseK module reported a best K of 3, 4, and 5, and K = 4 was chosen for three of the iterations. We chose to plot the mean population assignments for K = 4 (Fig. 5), but K = 3 and K = 5 show the same major patterns (not shown). The population assignments roughly correspond to three major groups: a Black Hills, Colorado, and Idaho group (Black Hills group), a group of the western populations (Western group), and a group of the GLR populations (GLR group). The fourth population assignment corresponds with a signal of admixture across all sampled populations except the Black Hills. There is a strong signal of admixture between the Black Hills group and the western group, and a separate signal of admixture between the western group and the GLR populations.

Each of the genetic diversity metrics (π, H_e, H_o, F_{IS}, private alleles, and percent polymorphic loci) were lower in the GLR than in the West, but none were significant except for the number of private alleles (Table 2a-c). Comparing the Black Hills with the West, the inbreeding coefficient was the only metric that was significantly lower in the Black Hills.

DISCUSSION

Implications of age of* Rubus parviflorus*: The biogeographic question of when and where *R. parviflorus* diverged in North America plays a role in determining the timing and likelihood of a restricted or widespread pre-Pleistocene distribution. The crown age of *R. parviflorus* dates to ~6.25 Mya (3.3-9.5 Mya 95% HPD), indicating a pre-Pleistocene origin, but not ruling out a

mid-Pleistocene origin. This confirms that we need to consider the glacial cycles in our analysis of the *R. parviflorus* disjunct distribution. This date also rules out the hypothesis that vicariance of northern-dispersed *R. parviflorus* was caused by the aridification of the Great Plains (~25 Mya), which was hypothesized by Fassett (1941). Its sister species, *R. odoratus*, is also native to North America, but it is eastern-restricted. This type of species-pair distribution between Western and Eastern North America was noted by Iltis (1965) who said that it is of particular interest when sister species post-glacially migrate far enough from their restricted ranges to overlap and sometimes hybridize with each other. He noted that this overlap almost exclusively occurs in the glaciated regions of Northeastern United States. In Michigan, which was totally glaciated, *R. parviflorus* and *R. odoratus* are known to hybridize (Fassett 1941). The date of the divergence between *R. parviflorus* and *R. odoratus* allows for this scenario of sister species migration, where either species was restricted with subsequent post-glacial re-connection.

Incongruence using* R. odoratus *as an outgroup: Low support in the RAxML CONS trees is possibly the result of conflicting histories between SNP loci. This might be due to incomplete lineage sorting (ILS) or hybridization. We expected that sampling of *R. odoratus* from a cultivated collection in British Columbia would circumvent the issue of inadvertently sampling a recent hybrid from the GLR where these two species overlap. We also made efforts pre-sequencing to exclude populations that were in the conspicuous zone of overlap in Michigan in order to avoid any recent hybrids. However, these efforts do not preclude the possibility that there was historical hybridization between *R. odoratus* and *R. parviflorus* that is reducing support for the RAxML topology. The effect of this potential hybridization is also reflected in the SVDQuartet analyses with changes in topology and decreased support values when the

outgroup is included. The SVDQuartet approach looks at individual loci and summarizes all quartets. If there was any hybridization, the conflicting histories between loci would make it difficult to cleanly summarize quartets, and bootstrap replicates, which subsample loci, and in our case individuals, would not always result in the same topology. In addition, the amount of missing data in the outgroup is substantial due to the small proportion of outgroup samples and the filtering process during GBS assembly. Missing data has been shown to have an impact on topology estimates for coalescence methods (Hime 2017). To get around this, effort should be made during GBS assembly filtering to include only loci that are also included in the outgroup, even though this will reduce the size of the dataset. In the case of *R. parviflorus*, this filtering might not result in significant improvement to the topology stability because background hybridization will continue to be an issue. Based on our results, we consider the SVDQuartet analysis without the outgroup to be the more accurate representation of the population histories.

Stepping stone model: Remarkably, our results exactly mirror the results that Fassett (1941) found based on the number and makeup of the morphological varieties in each region. Our SVDQuartet analysis shows that the Black Hills populations were not a dispersal stepping stone from western North America to the GLR. Instead, we show that the Black Hills populations have a more recent common ancestor with the Idaho and Colorado populations, suggesting an ancestral range among these regions. In addition, the Black Hills, Idaho, and Colorado populations share recent common ancestry with the northern California and Oregon populations, suggesting an ancestral range within these broader regions, as Fassett had previously concluded.

In addition, British Columbia would have been largely covered by ice at the Last Glacial Maximum, suggesting that the two dispersal events to British Columbia (Fig. 4) occurred post-

glacially as the ice was receding. However, there is some phylogeographical and geological evidence of refugia within the ice sheets in southwestern Alberta and northern British Columbia, as well as an ice-free corridor, reviewed in Shafer et al. (2010), that might have allowed these northward dispersals to have occurred during ice cover or even earlier. This scenario can be tested in the future with denser population sampling from this region and finer-scale demographic modeling and dating analyses. The more complicated scenario in which *R. parviflorus* dispersed north earlier, survived in a glacial refugium in Alaska or Beringia (Hultén 1937, Nimis et al. 1998, Shafer et al. 2010), and re-colonized British Columbia would also be consistent with the noted dispersals to British Columbia. However, the absence of long branches on the northern populations (Fig. 4) and the lack of *R. parviflorus* further north into Alaska makes this alternative scenario less likely. Population dating analyses might be able to tease apart these competing scenarios. Estimating the generation time of perennial species is difficult and there is not a good genome-wide mutation rate, so divergence dating in a coalescence model might not be very feasible. However, reduced representation genomic data can be informative on time scales as deep as 50-60 Mya (Eaton et al. 2016), so there is promise for future research on the divergence dates of these populations. We might be able to use GBS data of the Anoplobatus clade plus other *Rubus* taxa to reconstruct a BEAST chronogram using secondary time calibrations from a more robust *Rubus*-wide chronogram.

Finally, the GLR populations are sister to two of the British Columbia populations, suggesting a more northern dispersal route to the GLR. Although long distance dispersal is less likely based on the fleshy aggregate fruit type of *R. parviflorus* and absence of west-east animal migrations (Marquis and Voss 1981), we cannot rule this out based on phylogenetic evidence alone. Alternatively, this range expansion might have occurred as periglacial migration along a

receding glacier because in the case of their own post-glacial dispersal from the south, the British Columbia populations that are sister to the GLR populations would otherwise not have existed. Finally, considering the scenario of refugial isolation of the British Columbia populations amidst glacial cover, dispersal out of refugial pockets may have been hindered by ice, but dispersal from an ice-free corridor is plausible. This possibility of older British Columbia populations means that we cannot rule out an older dispersal route from British Columbia to the GLR without additional demographic modeling of population bottlenecks in the British Columbia region.

Genetic structure: The genetic structure of the populations of *R. parviflorus* tells a similar story as does the SVDQuartet analysis. The Black Hills group has a distinct population structure (Fig. 5, in purple), but there is admixture between the Black Hills group and the western populations (Fig. 5, in red), which further supports their phylogenetic ancestry. In addition, the primary genotypes of the Wisconsin populations (Fig. 5, in blue) are the same as the primary genotypes in the West, and there is additional signal of admixture (Fig. 5, in red) between these populations, which is the same admixed genotype shared with the Black Hills group. The Michigan populations have a distinctive population structure (Fig. 5, in green) but share some elements of the genotype shared between Wisconsin and the West (Fig. 5, in blue). The primary genotype of the Black Hills group (Fig. 5, in purple) is admixed with the group in the West and the Michigan populations. We propose that both of the signals of admixture between the GLR populations and the Black Hills group (Fig. 5, in red and purple) are indications that these populations established later and brought along genotypes from the western North American populations that already contained admixed genotypes from the Black Hills group. Finally, the lack of any admixture (Fig. 5, in red) in the South Dakota Black Hills population, indicated by its

complete purple color, suggests isolation of this population and a lack of gene flow between it and other surrounding populations. This pattern is reinforced with the higher signal of inbreeding when comparing the Black Hills with western North America.

Genetic diversity: The low number of private alleles found in the GLR might be indicative of the younger age of the populations in the GLR or the effect of a long-distance dispersal that reduced the number of low-frequency alleles. Other genetic diversity metrics, such as low heterozygosity or high F_{IS}, did not reveal any significant signatures of population bottleneck in the GLR (Table 2b). This indicates that the GLR populations are probably not the result of re-colonization from a limited source area after glacial retreat, nor the result of a long-distance dispersal founder event. Since the genetic signature of population bottlenecks can be detected (Luikart and Cornuet 1998), explicit tests of bottleneck events such as Tajima's D should be done to confirm this. The genetic diversity measurements likewise did not indicate any immediate threats to the populations in the GLR or the Black Hills when comparing them with the western populations. However, when comparing the Black Hills population with the western North American populations, the inbreeding coefficient was significantly higher in the Black Hills (Table 2c). The increased homozygosity could be due to isolation of the Black Hills populations and minimal gene flow entering the population.

A caveat to consider when evaluating our measurements is that our final sampling post-filtering had one instance of as few as 4 individuals per population. However, it has been suggested that as few as 3 or 4 individuals is enough to capture genetic patterns within a population for comparison across landscapes (Prunier et al. 2013). In addition, even though our sampling scheme did not cover the entire breadth of the populations, we were consistent in the

spacing of sampling from population to population, which we expect to mitigate some of the issues when comparing populations of different sizes.

CONCLUSION

By evaluating the phylogeographic history of *R. parviflorus* in a molecular framework, we show that its genetic history mirrors its morphological variation (Fassett 1941). We uncovered the dispersal history of these populations with some general inferences of the age of population divergences. In particular, we show that the Black Hills were not a dispersal stepping stone to the Great Lakes Region, and we suggest that the GLR populations were likely the result of a more northern peri-glacial dispersal route. This post-glacial migration is suggested to be common for species that have more eastern-restricted sister species, such as *R. odoratus*, with which there is overlap in glaciated regions of Northeastern United States (Iltis 1965). The comparable levels of genetic diversity in the GLR compared to the western North American populations shows that the GLR populations are not necessarily more vulnerable to climate change based on lower adaptive potential, despite being isolated from the more widespread range in the West. However, potential suitable niche space in the future is moving asynchronously with current species localities in the GLR (Ash et al. 2017, Spalink et al. in press), and the persistence of *R. parviflorus* in the GLR should be explored further with species distribution models.

In addition to our conclusions on the phylogeographic history and the persistence of *R. parviflorus* in the GLR, our data uncovers issues with ILS, hybridization and missing data. Our results show evidence of incongruence between loci in *R. parviflorus*, perhaps caused by ILS. Incomplete lineage sorting might be expected at the population level, and the short branches between populations further suggests that this could be a factor in the history of these

populations. Because of this incongruence, we propose that the population tree should be assessed with alternative concordance methods such as BUCKy (Larget et al. 2010). The evidence of ILS also highlights the issues involved with concatenating SNPs for ML analyses. The cause of the topological differences that we found is still not certain, however, and analyses should be run with reduced amounts of missing data in the outgroup. Here we identify the underlying problem with outgroup selection for phylogeographic analysis in groups like *Rubus* that have rampant hybridization. This issue might be overcome in the future, since GBS data have been shown to work across larger time scales, allowing for outgroup selection that is not likely involved in hybridization.

Additional taxa of the Western North America-GLR disjunct distribution should be assessed for their phylogeographic histories to determine whether the species of this distribution share the same history. Additional analysis of plants with different seed dispersal or environmental sensitivities might enable a broader meta-analysis of this distribution, which is important for continued understanding of the forces that drive the floristic elements in North America.

Table 1: List of populations sampled for genotyping-by-sequencing, their locality, and the number of samples per population in the final, cleaned data set. The population names have the state/province appended to the beginning for easier interpretation of their general location.

Name	Locality	Number of samples in final data set
WA_BridgeCreek	48.503826, -120.716707	4
WI_BruleS	46.632370, -91.599356	7
OR_Willamette	44.395234, -122.137903	6
MI_Windigo	47.936917, -89.135196	4
WA_Cascade	48.474824, -121.077042	7
CO_South	37.918661, -104.972692	8
WI_Door	44.927170, -87.184555	8
MI_Drummond	45.927780,-83.536890	6
BC_Hixton	53.445760, -122.563396	7
ID	43.627235, -115.987111	10
CA_Klamath	41.942098, -123.146643	5
BC_McLeod	54.903991,-122.935450	6
BC_MtRob	53.057330,-119.212983	8
SD_BlackHills	44.452483, -103.863067	7

Table 2a. The genetic diversity metrics for each population based on all loci that were present in 100% of the populations and at least 75% of individuals within a population: number of private alleles, percent of loci that were polymorphic, observed and expected heterozygosity, nucleotide diversity (Pi), and the inbreeding coefficient (F_{IS}). Diversity statistics were measured with kernel-smoothing and 100 bootstrap replicates. F_{IS} was calculated with the Bonferroni correction with a base p-value of 0.05. **2b.** The comparison between the GLR and western North American populations and **2c.** The comparison between the Black Hills and western North American populations. P-values are from unpaired, two-sample t-tests.

Table 2a

Pop_ID	# of Sites	# Private Alleles	% Polymorphic Loci	Coded Region
WA_BridgeCreek	1614259	304	0.22722	West
WI_BruleS	1614259	259	0.29475	GLR
WA_Cascade	1614259	642	0.36017	West
CO_South	1614259	249	0.14409	West
WI_Door	1614259	301	0.30187	GLR
MI_Drummond	1614259	43	0.13734	GLR
BC_Hixton	1614259	516	0.36271	West
ID	1614259	394	0.2201	West
CA_Klamath	1614259	1122	0.3378	West
BC_McLeod	1614259	420	0.31185	West
BC_MtRob	1614259	401	0.33439	West
SD_BlackHills	1614259	227	0.11541	BH
OR_Willamette	1614259	1028	0.37187	West
MI_Windigo	1614259	63	0.23119	GLR

Table 2a (cont'd)

Pop_ID	Obs Het	Exp Het	Pi	F_{IS}
WA_BridgeCreek	0.00106	0.00081	0.00094	-0.00023
WI_BruleS	0.00104	0.0009	0.00097	-0.00011
WA_Cascade	0.00112	0.00105	0.00115	0.00009
CO_South	0.00081	0.00057	0.00061	-0.00041
WI_Door	0.00104	0.00094	0.00101	-0.00008
MI_Drummond	0.00097	0.00053	0.00058	-0.00072
BC_Hixton	0.0011	0.00104	0.00113	0.00008
ID	0.001	0.00079	0.00083	-0.00038
CA_Klamath	0.00108	0.00102	0.00114	0.00013
BC_McLeod	0.00102	0.00095	0.00104	0.00006
BC_MtRob	0.00112	0.001	0.00107	-0.00011
SD_BlackHills	0.00102	0.00054	0.00058	-0.00083

OR_Willamette	0.00111	0.00106	0.00117	0.00015
MI_Windigo	0.0012	0.00085	0.00098	-0.00039

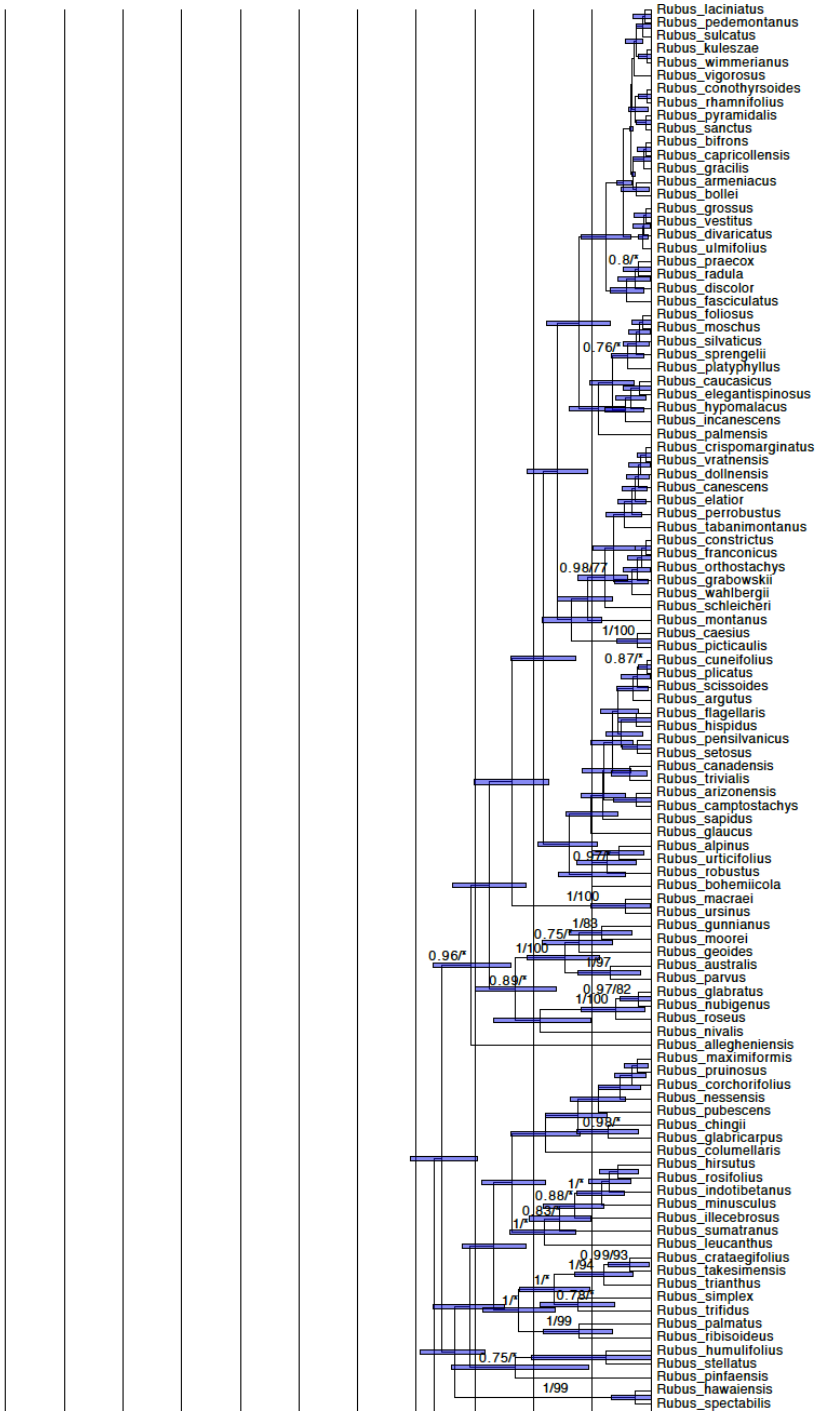
Table 2b

	Mean GLR Value	**Mean Western Value**	**p-value**
Private Alleles	166.5	564	0.03479**
% Polymorphic Loci	0.2412875	0.2966889	0.2684
Obs Het	0.0010625	0.00104667	0.7936
Exp Het	0.000805	0.00092111	0.2846
Pi	0.000885	0.00100889	0.3046
F_{IS}	-0.000325	-0.0000689	0.1094

Table 2c

	Mean Black Hills Value	**Mean Western Value**	**p-value**
Private Alleles	227	564	0.3353
% Polymorphic Loci	0.11541	0.2966889	0.06414
Obs Het	0.00102	0.00104667	0.804
Exp Het	0.00054	0.00092111	0.06069
Pi	0.00058	0.00100889	0.06072
F_{IS}	-0.00083	-0.0000689	0.01152**

Fig. 1 (starts on following page)



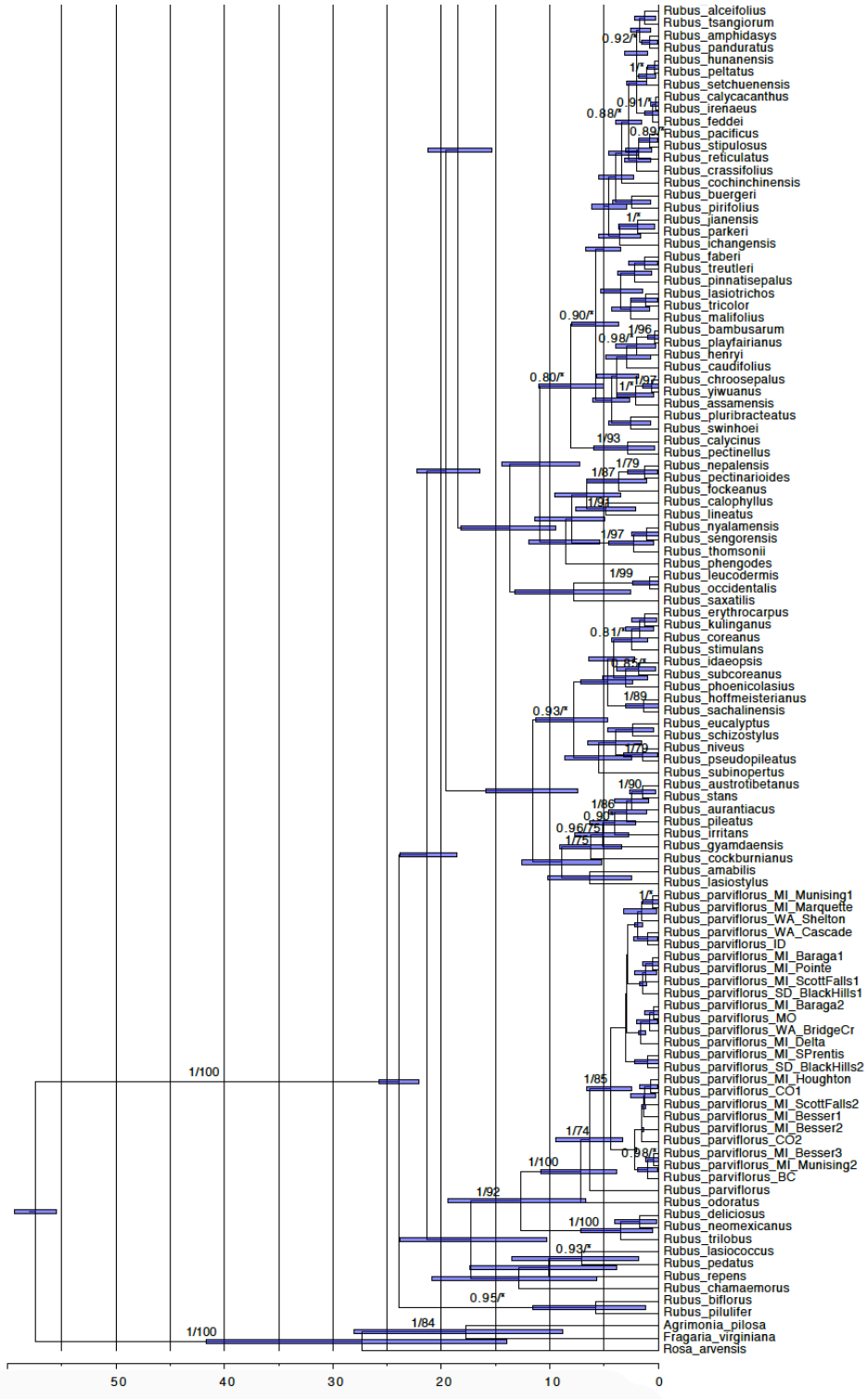


Fig. 1 BEAST chronogram of *Rubus* from the nrITS + 7 cpDNA partitioned data set. Secondary dates for the crown (23 Mya) and stem (57.84 Mya) age of *Rubus*. Node bars indicate 95% HPD, time axis is in millions of years. Posterior probabilities and ML bootstrap support are shown for PP values >= 0.75 (PP/bootstrap). Asterisks indicate that the ML tree did not include this topology or that the ML bootstrap support was < 75.

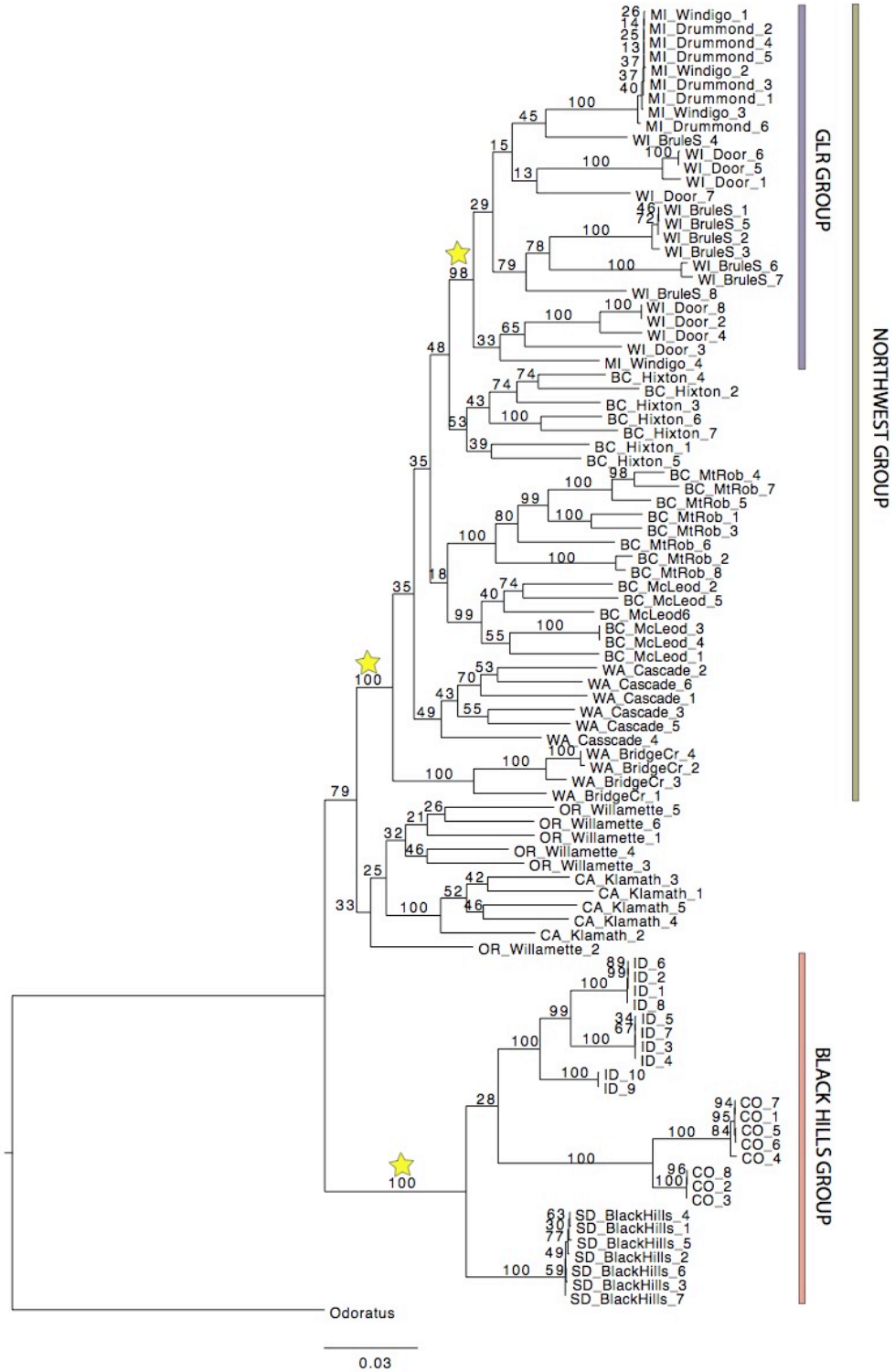
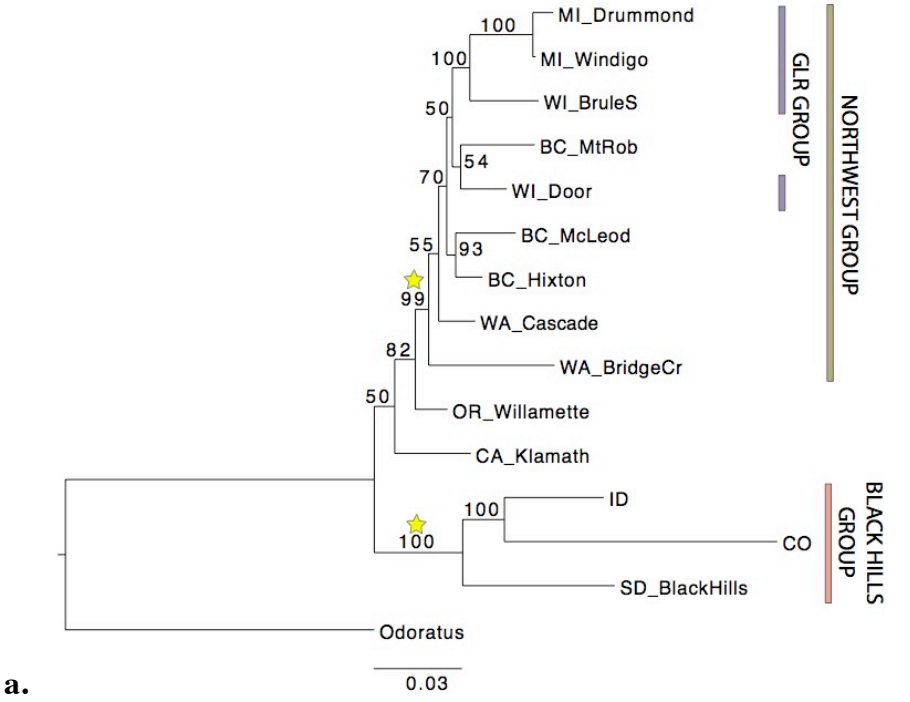
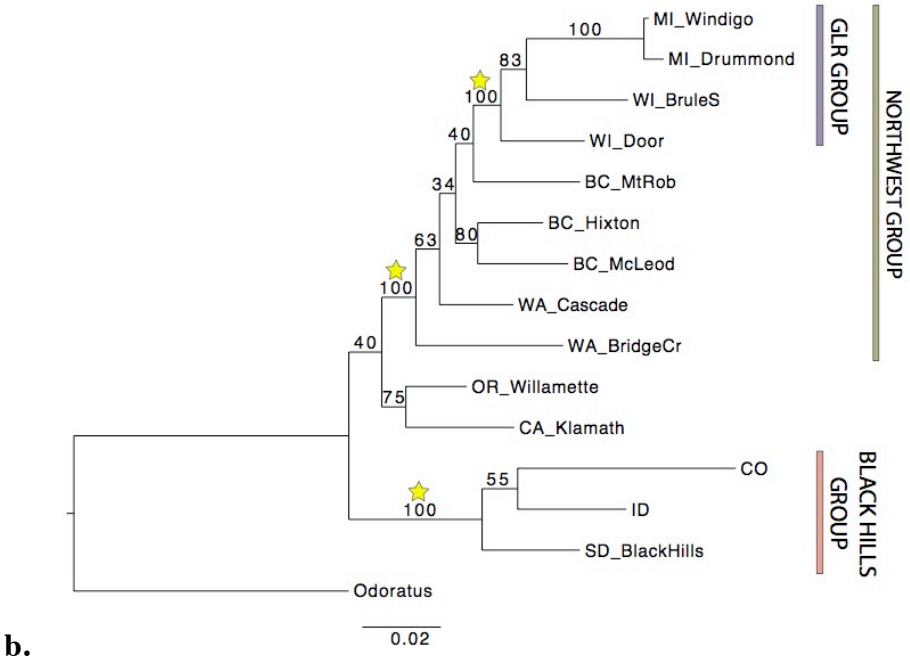


Fig. 2 RAxML tree estimated with GTR + Γ using the INDIV data set of 27,783 concatenated SNPs and implementing the Lewis ascertainment bias correction. Bootstrap values are above the branches. Stars indicate highly supported major clades. Major clades are: brown = central North American, Black Hills group, green = north, northwestern group (plus GLR), and purple is the monophyletic Great Lakes Region (GLR) group.



a.



b.

Fig. 3a RAxML tree using the CONS data set of 2,628 concatenated, majority rule SNPs for each site within a population, and implementing the Lewis ascertainment bias correction. **3b** RAxML tree using the CONS unmasked data set of 28,512 concatenated, majority rule SNPs for each site within a population, without ascertainment bias correction. Both trees were estimated with GTR + Γ and bootstrap values are reported above the branches. Stars indicate highly supported major clades. Major clades are: brown = central, Black Hills group, green = north northwestern group (plus GLR), and purple is the non-monophyletic GLR group.

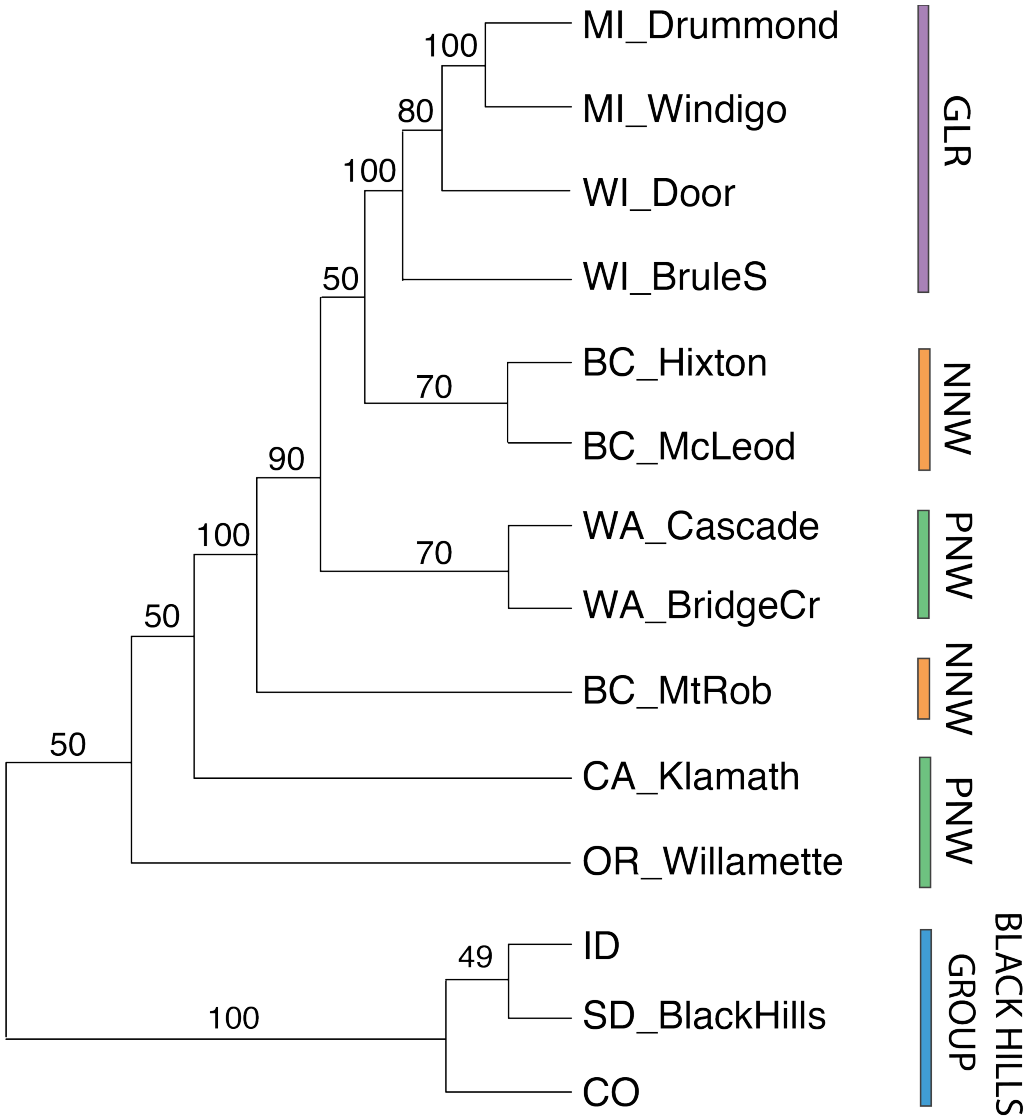


Fig. 4 PAUP SVDQuartet coalescence tree using the CONS, unmasked data set of 28,512 SNPs, without *R. odoratus*. Tree was rooted using the RAxML topology (Fig. 3). Bootstrap values are from 100 bootstrap replicates on 100,000 sub-sampled quartets. Low support for some of the clades might indicate conflict between SNPs.

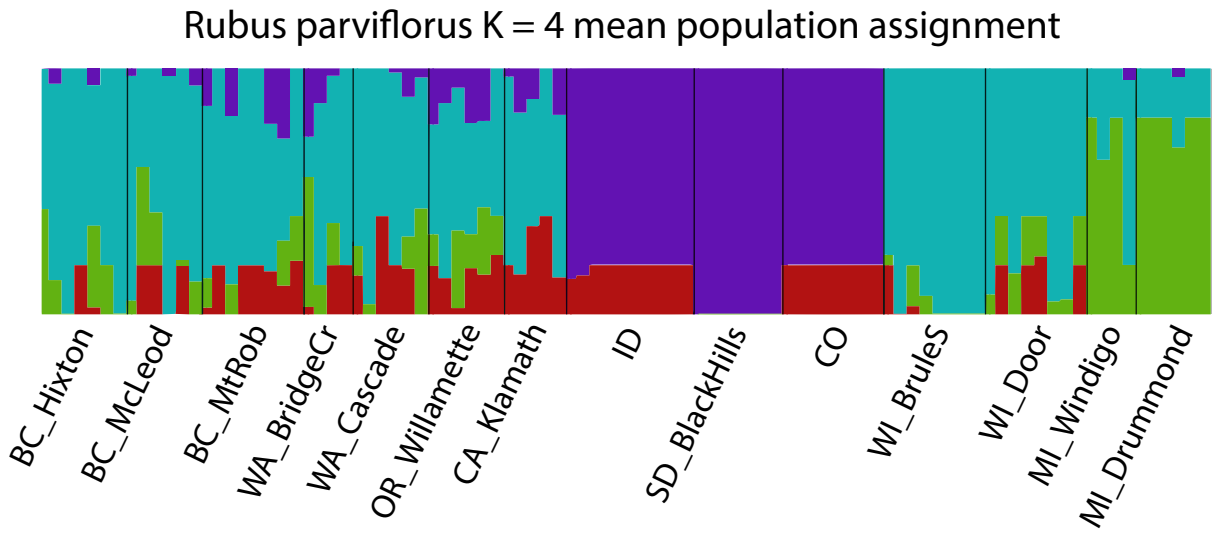


Fig. 5 Result from fastStructure with K = 4, plotted using distruct22. Mean population clusters were generated in CLUMPP, averaging across five iterations of fastStructure using K = 4. Populations in the plot are generally ordered West to East.

REFERENCES

Alice, L.A. & Campbell, C.S., 1999. Phylogeny of Rubus (Rosaceae) based on nuclear ribosomal DNA internal transcribed spacer region sequences. 86(1), pp.81–97.

Andermann, T. et al., 2018. Allele Phasing Greatly Improves the Phylogenetic Utility of Ultraconserved Elements. S. Renner, ed. *Systematic Biology*, 16, p.S5.

Anderson, B.M. et al., 2017. Genotyping-by-Sequencing in a Species Complex of Australian Hummock Grasses (Triodia): Methodological Insights and Phylogenetic Resolution G. Sun, ed. *PLoS ONE*, 12(1), pp.e0171053–34.

Association, D.S.S. et al., *PAUP* Phylogeny Analysis Using Parsimony (* and other methods), version 40b10*.

Axelrod, D.I. 1950. Evolution of desert vegetations in western North America. Carnegie Inst. Wash. Publ, 590, pp.215-306.

Baker, H.G., 1962. Flora of the British Isles. A. R. Clapham, T. G. Tutin, and E. F. Warburg. Cambridge University Press, New York, ed. 2, 1962. xlvii + 1269 pp. Illus. $13.50. *Science*, 138(3539), pp.506–506.

Baldwin, B.G., 2014. Origins of Plant Diversity in the California Floristic Province. *Annual Review of Ecology, Evolution, and Systematics*, 45(1), pp.347–369.

Barker, B.S. et al., 2017. Population genomic analyses reveal a history of range expansion and trait evolution across the native and invaded range of yellow star thistle (Centaurea solstitialis). *Molecular Ecology*, 26(4), pp.1131–1147.

Beatty, G.E. & Provan, J., 2010. Refugial persistence and postglacial recolonization of North America by the cold-tolerant herbaceous plant Orthilia secunda. *Molecular Ecology*, 19(22), pp.5009–5021.

Iltis, H.H. 1965. The genus Gentianopsis (Gentianaceae): transfers and phytogeographic comments. *Sida, Contributions to Botany* 2, pp.129-154.

Bouckaert, R. et al., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. A. Prlic, ed. *PLoS computational biology*, 10(4), p.e1003537.

Bryant, D. et al., 2012. Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Molecular Biology and Evolution*, 29(8), pp.1917–1932.

Catchen, J. et al., 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22(11), pp.3124–3140.

Chhatre, 2016, http://www.crypticlineage.net/pages/distruct.html

Chifman, J. & Kubatko, L., 2014. Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, 30(23), pp.3317–3324.

Chin, S.-W. et al., 2014. Diversification of almonds, peaches, plums and cherries - molecular systematics and biogeographic history of Prunus (Rosaceae). *Molecular Phylogenetics and Evolution*, 76, pp.34–48.

Chou, J. et al., 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics*, 16(Suppl 10), p.S2.

Deevey Jr, E.S., 1949. Biogeography of the Pleistocene: Part I: Europe and North America. *Geological Society of America Bulletin*, 60(9), pp. 1315-1416.

Dillenberger, M.S. & Kadereit, J.W., 2017. Simultaneous speciation in the European high mountain flowering plant genus Facchinia (Minuartia s.l., Caryophyllaceae) revealed by genotyping-by-sequencing. *Molecular Phylogenetics and Evolution*, 112, pp.23–35.

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

Eaton, D.A.R. & Ree, R.H., 2013. Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). *Systematic Biology*, 62(5), pp.689–706.

Eaton, D.A.R. et al., 2016. Misconceptions on Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants. *Systematic Biology*, 56, p.syw092.

Eriksson, T., Donoghue, M.J. & Hibbs, M.S., 1998. Phylogenetic analysis ofPotentilla using DNA sequences of nuclear ribosomal internal transcribed spacers (ITS), and implications for the classification ofRosoideae (Rosaceae). *Plant Systematics and Evolution*, 211(3-4), pp.155–179.

Eriksson, T., Hibbs, M.S. & Yoder, A.D., 2003. The phylogeny of Rosoideae (Rosaceae) based on sequences of the internal transcribed spacers (ITS) of nuclear ribosomal DNA and the trnL/F region of chloroplast DNA.

Fassett, N.C., 1941. Mass Collections: Rubus Odoratus and R. parviflorus. *Annals of the Missouri Botanical Garden*, 28(3), p.299.

Hime, P.M. 2017. Dissertation: Genomic perspectives on amphibian evolution across multiple phylogenetic scales. University of Kentucky.

J, M.R. & G, V.E., 1981. Distributions of some western North American plants disjunct in the Great Lakes region. *Michigan Botanist*.

Jackson, S.T. & Overpeck, J.T., 2000. Responses of plant populations and communities to environmental changes of the late Quaternary. *Paleobiology*, 26(sp4), pp.194–220.

Jakobsson, M. & Rosenberg, N.A., 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23(14), pp.1801–1806.

Johnson, M.G. et al., 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in plant sciences*, 4(7), p.1600016.

Kates, H.R. et al., 2018. Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of Artocarpus. *American Journal of Botany*, 105(3), pp.404–416.

Kearse, M. et al., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), pp.1647–1649.

Larget, B.R. et al., 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, 26(22), pp.2910–2911.

Leaché, A.D. et al., 2015. Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. *Systematic Biology*, 64(6), pp.1032–1047.

Lemmon, A.R. et al., 2009. The Effect of Ambiguous Data on Phylogenetic Estimates Obtained by Maximum Likelihood and Bayesian Inference. *Systematic Biology*, 58(1), pp.130–145.

Liu, L. & Yu, L., 2011. Estimating species trees from unrooted gene trees. *Systematic Biology*, 60(5), pp.661–667.

Luikart, G. & Cornuet, J.-M., 1998. Empirical Evaluation of a Test for Identifying Recently Bottlenecked Populations from Allele Frequency Data. *Conservation Biology*, 12(1), pp.228–237.

Marth, G.T. et al., 2004. The Allele Frequency Spectrum in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in Three Large World Populations. *Genetics*, 166(1), pp.351–372.

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

Merkel, D., 2014. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014(239), p2.

Miller, M.A., Pfeiffer, W. & Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In 2010 Gateway Computing Environments Workshop (GCE). IEEE, pp. 1–8.

Mirarab, S. & Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12), pp.i44–52.

Nimis, P.L. et al., 1998. A multivariate phytogeographic analysis of plant diversity in the Putorana Plateau (N. Siberia). *Opera Botanica*, 136.

Novembre, J. & Stephens, M., 2008. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5), pp.646–649.

Paris, J.R., Stevens, J.R. & Catchen, J.M., 2017. Lost in parameter space: a road map for stacks S. Johnston, ed. *Methods in Ecology and Evolution*, 8(10), pp.1360–1373.

Potter, D. et al., 2007. Phylogeny and classification of Rosaceae. *Plant Systematics and Evolution*, 266(1-2), pp.5–43.

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

Prunier, J.G. et al., 2013. Optimizing the trade-off between spatial and genetic sampling efforts in patchy populations: towards a better assessment of functional connectivity using an individual-based sampling scheme. *Molecular Ecology*, 22(22), pp.5516–5530.

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

Rambaut, A. et al., 2018. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. E. Susko, ed. *Systematic Biology*, 22, p.341.

Reaz, R., Bayzid, M.S. & Rahman, M.S., 2014. Accurate Phylogenetic Tree Reconstruction from Quartets: A Heuristic Approach R. Wu, ed. *PLoS ONE*, 9(8), pp.e104008–13.

Rochette, N.C. & Catchen, J.M., 2017. Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Publishing Group*, 12(12), pp.2640–2659.

Ronquist, F. et al., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3), pp.539–542.

Schmitt, T., 2007. Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology*, 4(1), pp.11–13.

Shafer, A.B.A. et al., 2010. Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Molecular Ecology*, 19(21), pp.4589–4621.

Shreve, F., 1942. The desert vegetation of North America. *The Botanical Review*, 8(4), pp.195–246.

Smedmark, J.E. & Eriksson, T., 2002. Phylogenetic relationships of Geum (Rosaceae) and relatives inferred from the nrITS and trnL-trnF regions. *Systematic Botany*, 27(2), pp.303-317.

Smith, S.A. & Brown, J.W., 2018. Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany*, 105(3), pp.302–314.

Soltis, D.E. et al., 2006. Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, 15(14), pp.4261–4293.

Spalink, D. et al. Spatial phylogenetics reveals evolutionary constraints on the assembly of a large regional flora. *In press*.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), pp.1312–1313.

Stebbins, G.L., 1950. Variation and evolution in plants. Geoffrey Cumberlege, London.

Stebbins, G.L., 1970. Variation and evolution in plants: progress during the past twenty years. In *Essays in evolution and genetics in honor of Theodosius Dobzhansky* (pp. 173-208). Springer, Boston, MA.

Swenson, N.G. & Howard, D.J., 2005. Clustering of contact zones, hybrid zones, and phylogeographic breaks in North America. *The American Naturalist*, 166(5), pp.581–591.

Teeri, J.A. & Stowe, L.G., 1976. Climatic patterns and the distribution of C4 grasses in North America. *Oecologia*, 23(1), pp.1–12.

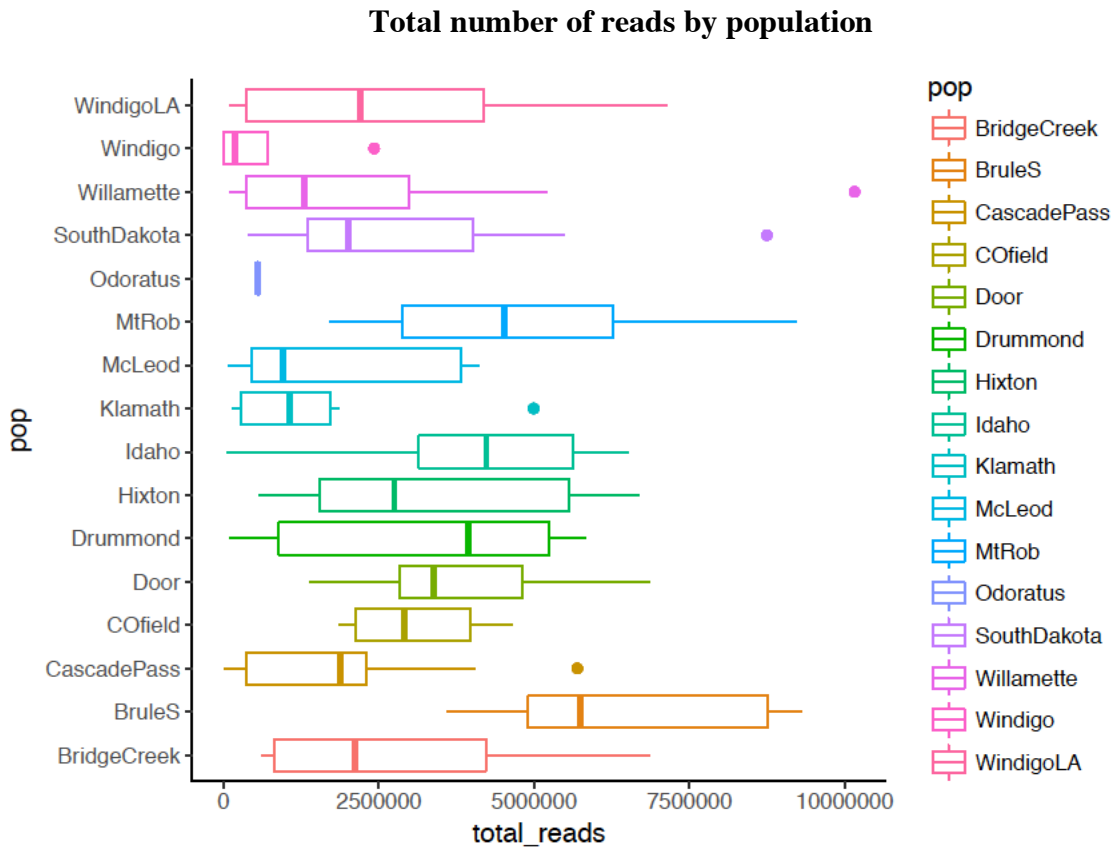
Thule, E.H.S.B.-L.A.1937, *Outline of the history of Boreal and Arctic biota during the Quaternary Period*,

Wagner, C.E. et al., 2012. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, 22(3), pp.787–798.

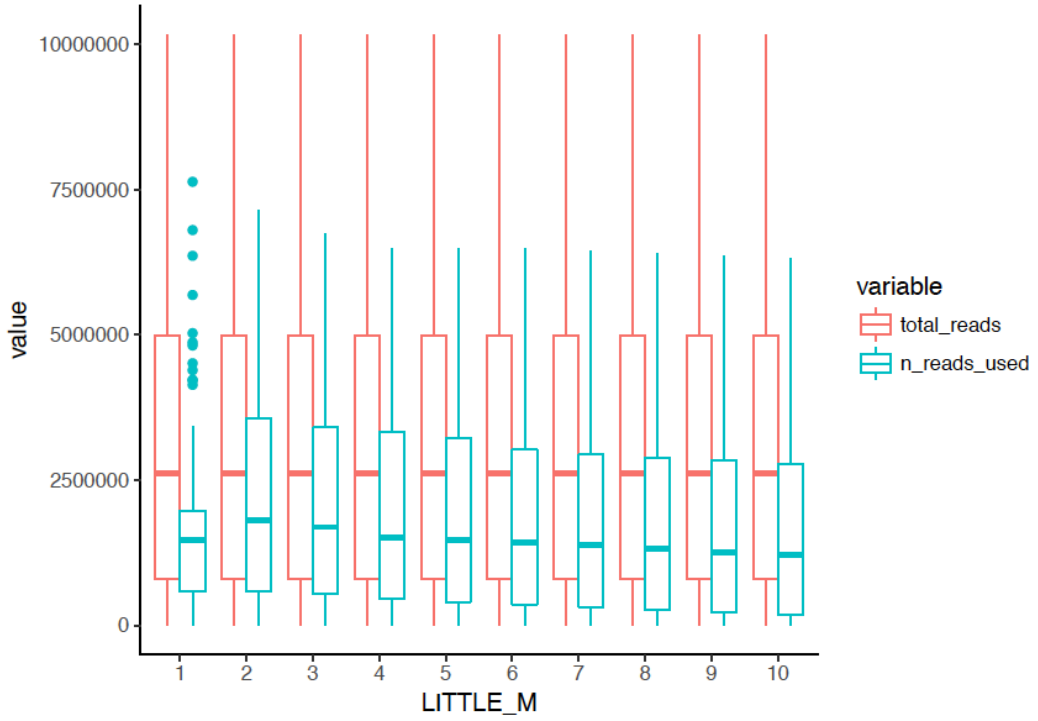
Xiang, Y. et al., 2017. Evolution of Rosaceae Fruit Types Based on Nuclear Phylogeny in the Context of Geological Times and Genome Duplication. *Molecular Biology and Evolution*, 34(2), pp.262–281.

Zhang, S.-D. et al., 2017. Diversification of Rosaceae since the Late Cretaceous based on plastid phylogenomics. *New Phytologist*, 214(3), pp.1355–1367.

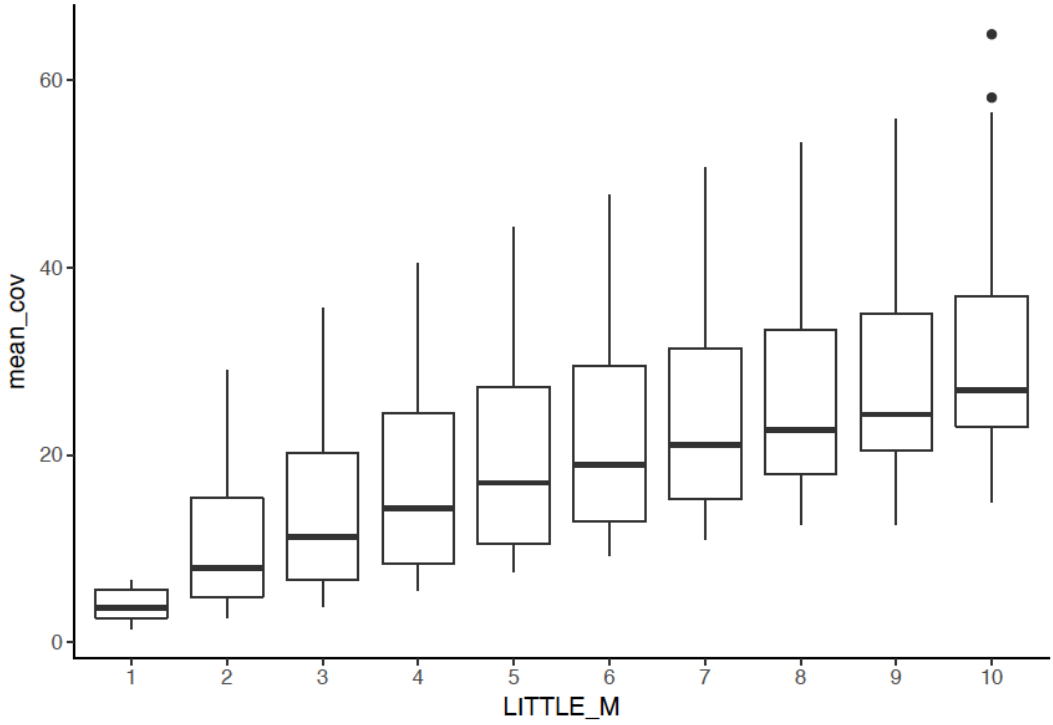
Appendix 1: Results from Stacks assembly optimization.



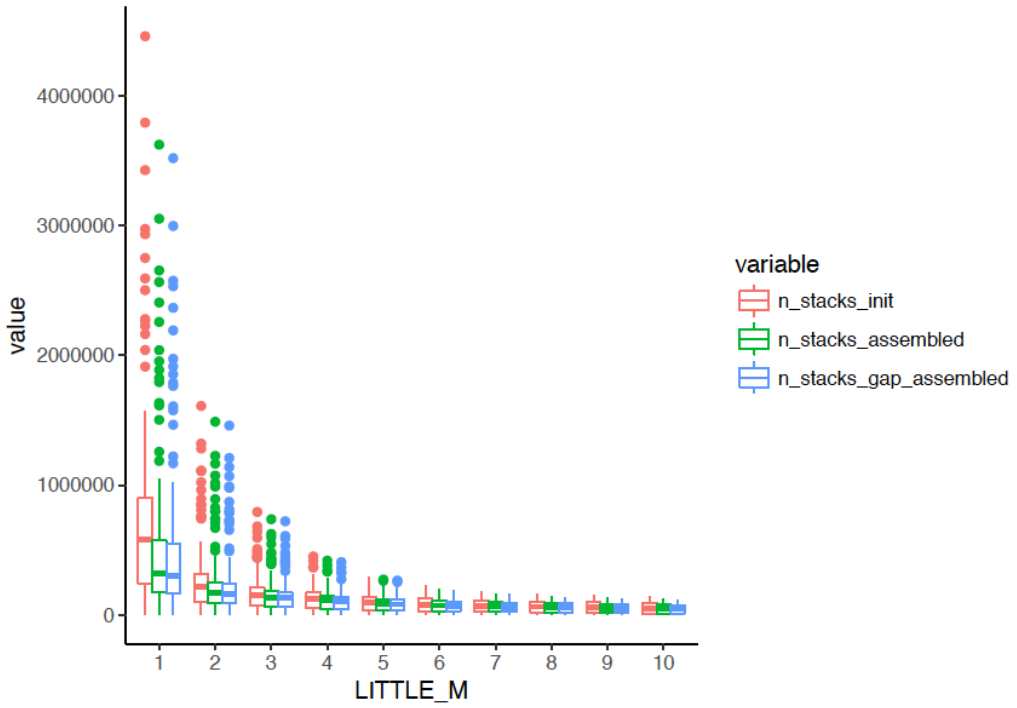
Number of reads used based on 'm' assembly parameter



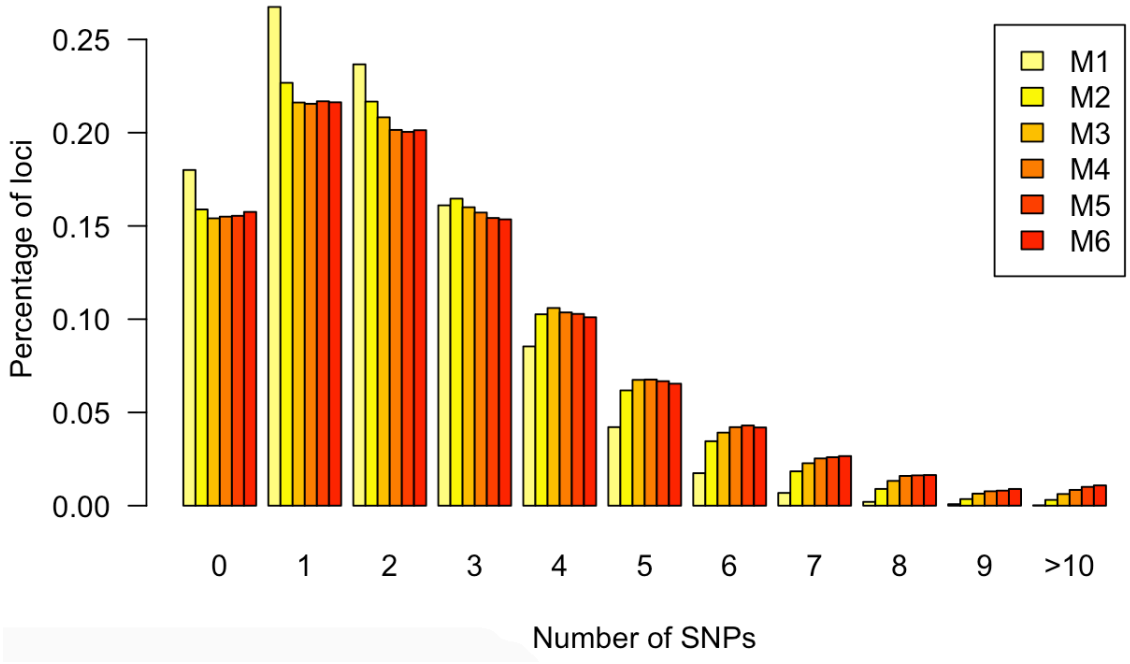
Mean coverage per locus based on 'm' assembly parameter



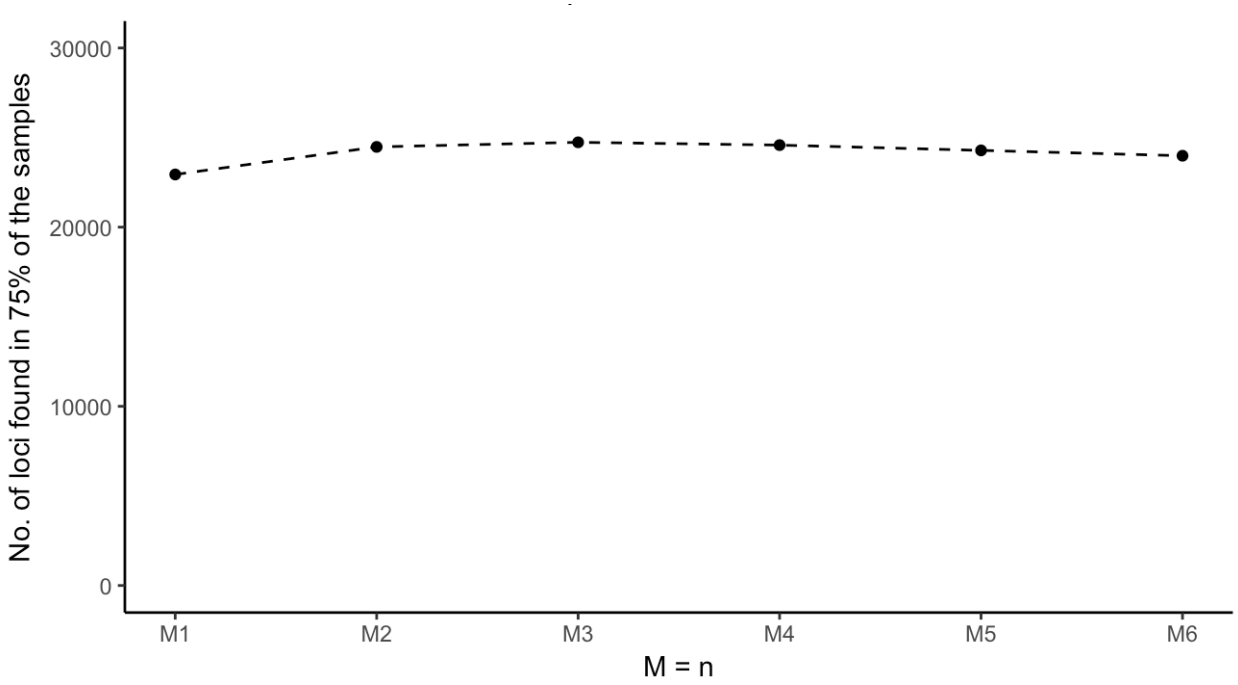
Number of loci (stacks) formed based on 'm' assembly parameter



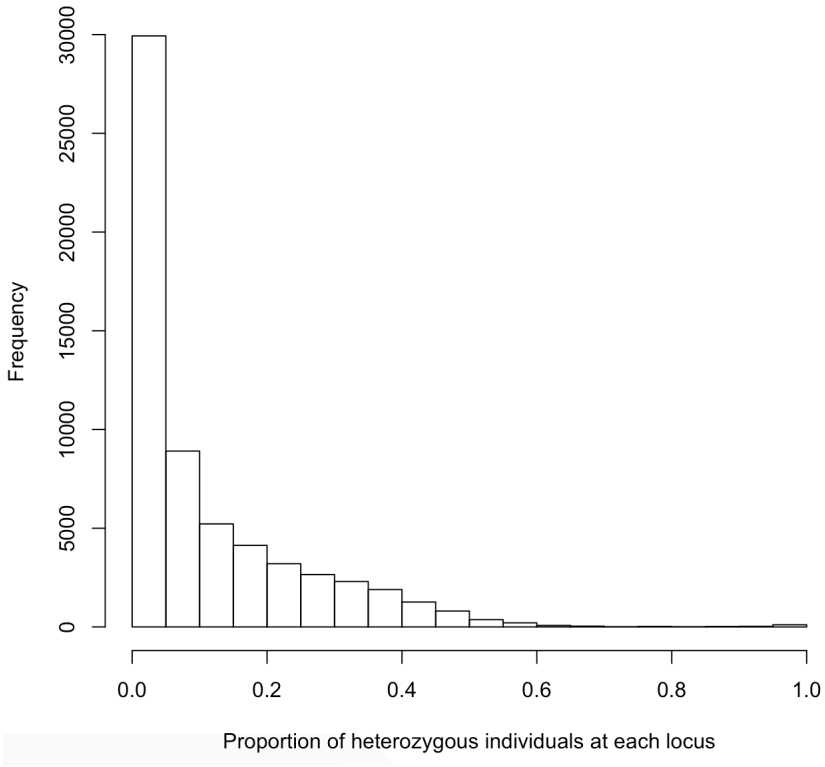
Proportion of loci with n-SNPs based on "M" assembly parameter



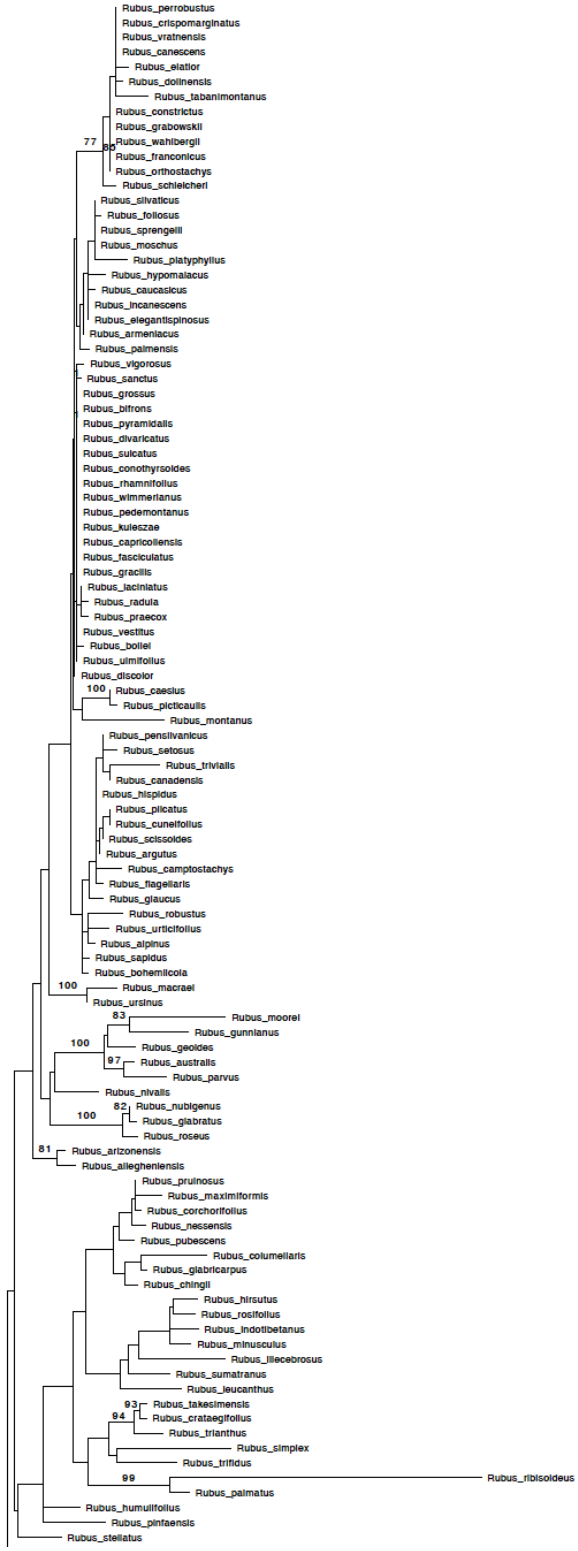
Number of loci in 75% of samples for each value of M==n

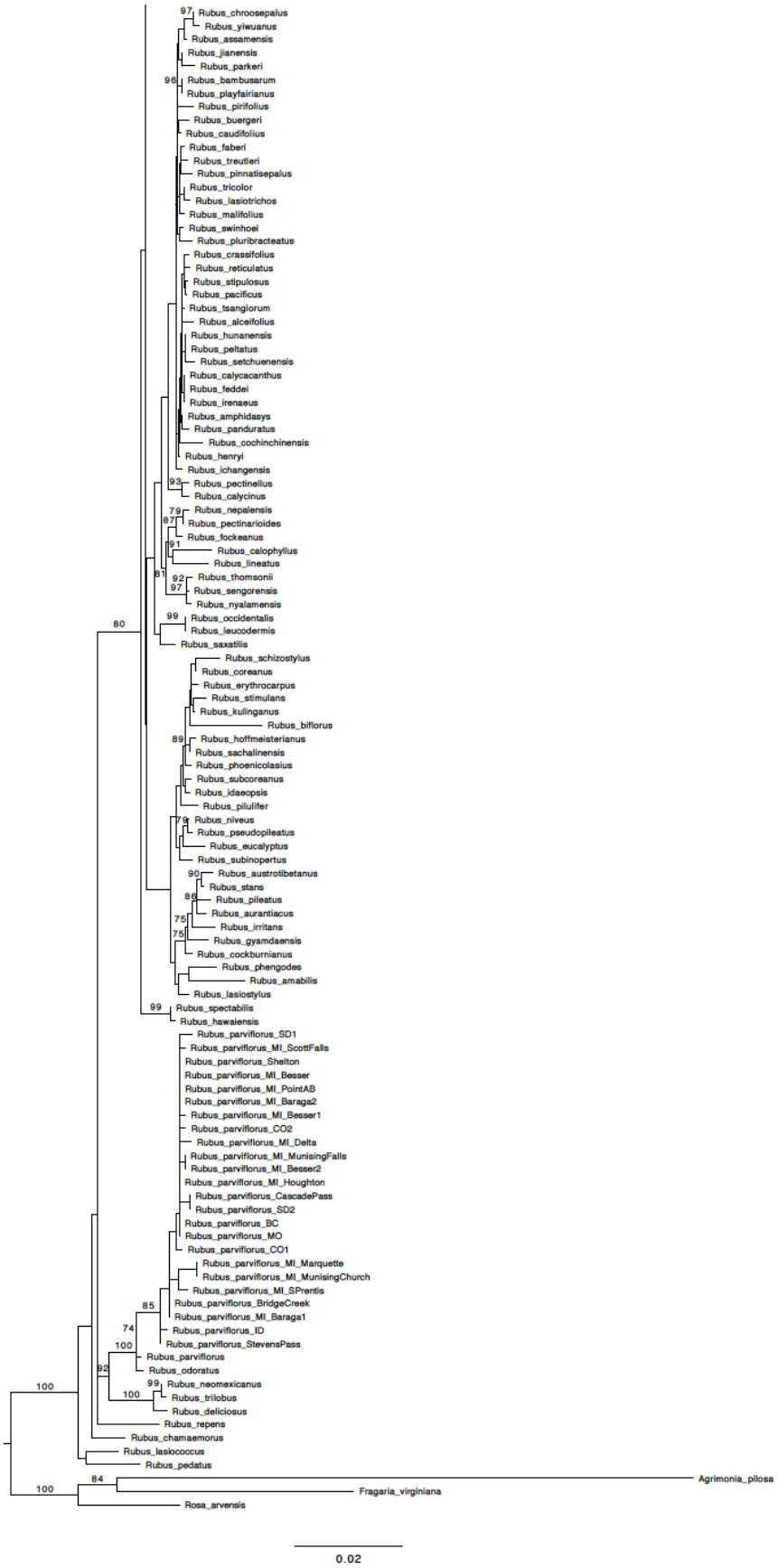


Proportion of heterozygotes at each locus (SNP) for M = n = 3

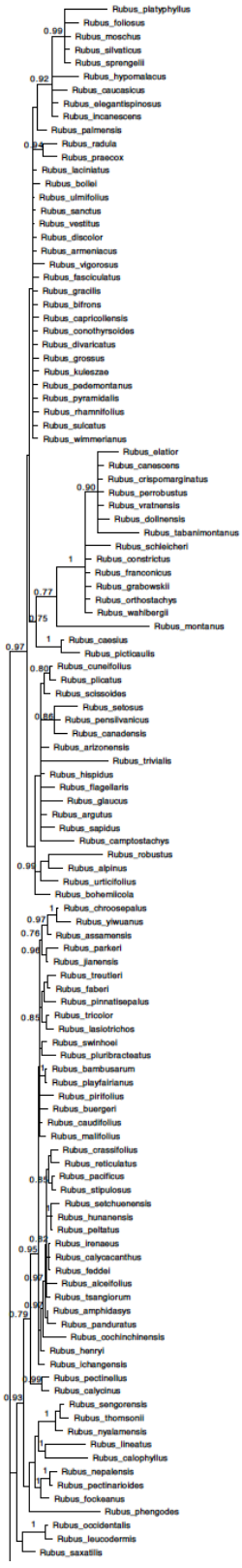


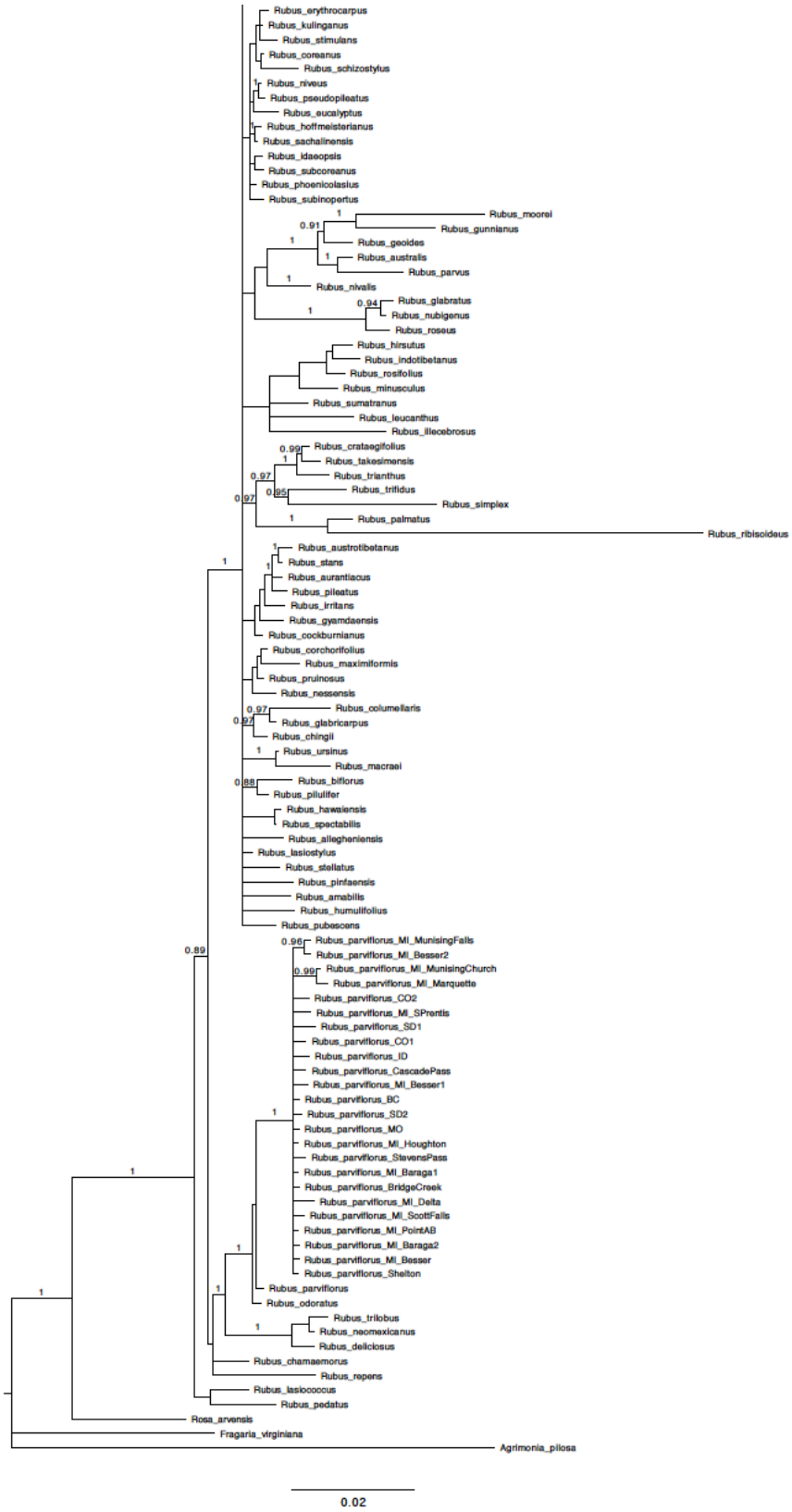
Appendix 2a: *Rubus* nrITS + cpDNA RAxML tree using GTR + Γ and 100 bootstrap replicates. Only bootstrap support >= 75 is shown.




Rubus_chroosepalus
Rubus_yiwuanus
Rubus_assamensis
Rubus_jianensis
Rubus_parkeri
Rubus_bambusarum
Rubus_playfairianus
Rubus_pirifolius
Rubus_buergeri
Rubus_caudifolius
Rubus_faberi
Rubus_treutleri
Rubus_pinnatisepalus
Rubus_tricolor
Rubus_lasiotrichos
Rubus_malifolius
Rubus_swinhoei
Rubus_pluribracteatus
Rubus_crassifolius
Rubus_reticulatus
Rubus_stipulosus
Rubus_pacificus
Rubus_tsangiorum
Rubus_alceifolius
Rubus_hunanensis
Rubus_peltatus
Rubus_setchuenensis
Rubus_calycacanthus
Rubus_feddei
Rubus_irenaeus
Rubus_amphidasys
Rubus_panduratus
Rubus_cochinchinensis
Rubus_henryi
Rubus_ichangensis
Rubus_pectinellus
Rubus_calycinus
Rubus_nepalensis
Rubus_pectinarioides
Rubus_fockeanus
Rubus_calophyllus
Rubus_lineatus
Rubus_thomsonii
Rubus_sengorensis
Rubus_nyalamensis
Rubus_occidentalis
Rubus_leucodermis
Rubus_saxatilis
Rubus_schizostylus
Rubus_coreanus
Rubus_erythrocarpus
Rubus_stimulans
Rubus_kuilinganus
Rubus_biflorus
Rubus_hoffmeisterianus
Rubus_sachalinensis
Rubus_phoenicolasius
Rubus_subcoreanus
Rubus_idaeopsis
Rubus_pilulifer
Rubus_niveus
Rubus_pseudopileatus
Rubus_eucalyptus
Rubus_subinopertus
Rubus_austrotibetanus
Rubus_stans
Rubus_pileatus
Rubus_aurantiacus
Rubus_irritans
Rubus_gyamdaensis
Rubus_cockburnianus
Rubus_phengodes
Rubus_amabilis
Rubus_lasiostylus
Rubus_spectabilis
Rubus_hawaiensis
Rubus_parviflorus_SD1
Rubus_parviflorus_MI_ScottFalls
Rubus_parviflorus_Shelton
Rubus_parviflorus_MI_Besser
Rubus_parviflorus_MI_PointAB
Rubus_parviflorus_MI_Baraga2
Rubus_parviflorus_MI_Besser1
Rubus_parviflorus_CO2
Rubus_parviflorus_MI_Delta
Rubus_parviflorus_MI_MunisingFalls
Rubus_parviflorus_MI_Besser2
Rubus_parviflorus_MI_Houghton
Rubus_parviflorus_CascadePass
Rubus_parviflorus_SD2
Rubus_parviflorus_BC
Rubus_parviflorus_MO
Rubus_parviflorus_CO1
Rubus_parviflorus_MI_Marquette
Rubus_parviflorus_MI_MunisingChurch
Rubus_parviflorus_MI_SPrentis
Rubus_parviflorus_BridgeCreek
Rubus_parviflorus_MI_Baraga1
Rubus_parviflorus_ID
Rubus_parviflorus_StevensPass
Rubus_parviflorus
Rubus_odoratus
Rubus_neomexicanus
Rubus_trilobus
Rubus_deliciosus
Rubus_repens
Rubus_chamaemorus
Rubus_lasiococcus
Rubus_pedatus
Agrimonia_pilosa
Fragaria_virginiana
Rosa_arvensis
0.02

Appendix 2b: *Rubus* nrITS + cpDNA MrBayes consensus tree from 10 million generations with 25% burn-in, specifying *Agrimonia pilosa* as the outgroup. Only PP >= 75% is shown.




Rubus_erythrocarpus
Rubus_kuilinganus
Rubus_stimulans
Rubus_coreanus
Rubus_schizostylus
Rubus_niveus
Rubus_pseudopileatus
Rubus_eucalyptus
Rubus_hoffmeisterianus
Rubus_sachalinensis
Rubus_idaeopsis
Rubus_subcoreanus
Rubus_phoenicolasius
Rubus_subinopertus
Rubus_moorei
Rubus_gunnianus
Rubus_geoides
Rubus_australis
Rubus_parvus
Rubus_nivalis
Rubus_glabratus
Rubus_nubigenus
Rubus_roseus
Rubus_hirsutus
Rubus_lindotibetanus
Rubus_rosifolius
Rubus_minusculus
Rubus_sumatranus
Rubus_leucanthus
Rubus_illecebrosus
Rubus_crataegifolius
Rubus_takesimensis
Rubus_trianthus
Rubus_trifidus
Rubus_simplex
Rubus_palmatus
Rubus_ribisoideus
Rubus_austrotibetanus
Rubus_stans
Rubus_aurantiacus
Rubus_pileatus
Rubus_irritans
Rubus_gyamdaensis
Rubus_cockburnianus
Rubus_corchorifolius
Rubus_maximiformis
Rubus_pruinosus
Rubus_nessensis
Rubus_columellaris
Rubus_glabricarpus
Rubus_chingii
Rubus_ursinus
Rubus_macraei
Rubus_biflorus
Rubus_pilulifer
Rubus_hawaiensis
Rubus_spectabilis
Rubus_alleghaniensis
Rubus_lasiostylus
Rubus_stellatus
Rubus_pinfaensis
Rubus_amabilis
Rubus_humulifolius
Rubus_pubescens
Rubus_parviflorus_MI_MunisingFalls
Rubus_parviflorus_MI_Besser2
Rubus_parviflorus_MI_MunisingChurch
Rubus_parviflorus_MI_Marquette
Rubus_parviflorus_CO2
Rubus_parviflorus_MI_SPrentis
Rubus_parviflorus_SD1
Rubus_parviflorus_CO1
Rubus_parviflorus_ID
Rubus_parviflorus_CascadePass
Rubus_parviflorus_MI_Besser1
Rubus_parviflorus_BC
Rubus_parviflorus_SD2
Rubus_parviflorus_MO
Rubus_parviflorus_MI_Houghton
Rubus_parviflorus_StevensPass
Rubus_parviflorus_MI_Baraga1
Rubus_parviflorus_BridgeCreek
Rubus_parviflorus_MI_Delta
Rubus_parviflorus_MI_ScottFalls
Rubus_parviflorus_MI_PointAB
Rubus_parviflorus_MI_Baraga2
Rubus_parviflorus_MI_Besser
Rubus_parviflorus_Shelton
Rubus_parviflorus
Rubus_odoratus
Rubus_trilobus
Rubus_neomexicanus
Rubus_deliciosus
Rubus_chamaemorus
Rubus_repens
Rubus_lasiococcus
Rubus_pedatus
Rosa_arvensis
Fragaria_virginiana
Agrimonia_pilosa
0.02

Appendix 3: H' consistency values from CLUMPP, comparing population assignments across five iterations of each K value. We tested K = 2-10. Higher H' values indicate higher consistency. Only K2 was consistent, using a 0.8 cut-off.

	H'	**CLUMPP Algorithm**
K2	0.999	Greedy
K3	0.542	Greedy
K4	0.435	Greedy
K5	0.435	Greedy
K6	0.485	Greedy
K7	0.508	Greedy
K8	0.435	LargeKGreedy
K9	0.568	LargeKGreedy
K10	0.484	LargeKGreedy

CHAPTER 3: The devil is in the detail: phylogeography of disjunct *Oplopanax horridus* (Araliaceae) rejects the hypothesis of a founder event or refugium bottleneck in the Great Lakes Region, pointing to alternative modes of migration

ABSTRACT

The flora of the Great Lakes Region (GLR) is comprised of confluent floristic elements from adjacent geographic regions. It has additional unique elements that are disjunct from farther regions of North America, and some of these disjuncts are particularly striking because of their restricted isolation even within the GLR (See also, Chapter 4). Here we take an in-depth look at one case study of this strict disjunct distribution, *Oplopanax horridus* (Sm.) Miq. (Devil's-club, Araliacaeae, Asian Palmate Clade), to assess its biogeographic origin, mechanism of entry into the GLR, and vulnerability to future climate change. *Oplopanax horridus* is a member of the ginseng family (Araliaceae) and it is native to western North America with disjunct populations in the islands of Lake Superior. We used phylogenetics, population structure, and measures of genetic diversity to test these phylogeographic questions. Results indicate that *O. horridus* diverged from its Asian sister species before the Pleistocene, ~4 Mya, and that its ancestral range was in the Pacific Northwest, south of the glacial boundary at the Last Glacial Maximum. In western North America, Devil's-club dispersed northward, and demographic modeling supports dispersal from British Columbia to the islands of Lake Superior instead of from its more southern ancestral range, implying a post-glacial age of the GLR populations. Lake Superior populations do not show signs of a bottleneck event, strongly suggesting a migration route other than long-distance dispersal or persistence during the Pleistocene in a midwestern refugium.

These results do not rule out the possibility of human-mediated dispersal, which should be investigated further in an ethnobotanical framework.

INTRODUCTION

Oplopanax horridus (Sm.) Miq. (Devil's club) is one of a suite of taxa that are disjunct between western North America and the Great Lakes Region (GLR) (Marquis and Voss 1981, Chapter 1). Devil's-club is a unique case of this disjunct pattern because it is abundant and widespread west of the Rocky Mountains, which is characteristic of co-distributed taxa, but its distribution in the GLR is restricted to islands in Lake Superior. This incongruity in abundance and range between the west and the GLR strongly suggests an ancestral western North American range with dispersal to the GLR. *Oplopanax horridus* does not exhibit characters that would make it an obvious candidate for long-distance dispersal, however, since it produces bright red, fleshy drupes in mid-summer that are primarily bear and deer dispersed. The unique biogeographical distribution and life history of Devil's-club raises important questions: where did the GLR populations come from, when did they arrive in the GLR, and are the GLR populations at risk of extirpation given their isolation and rapid climate change?

The larger context of the genus *Oplopanax* sheds some light on the history of *O. horridus*. *Oplopanax horridus* is part of the Asian Palmate Clade of the ginseng family (Araliaceae). The Asian Palmate Clade consists of 18 Asian genera (Wen et al. 2001, Plunkett et al. 2004): 13 are restricted to Asia, one extends into Europe, three extend into Central and South America, while *Oplopanax* is the only genus that extends into North America. Of the three species of *Oplopanax*, *O. horridus* is the only North American species, and *O. elatus* and *O. japonicus* are restricted to East Asia. Because of its unique North American distribution

compared to its two sister species, *O. horridus* is suggested to have arisen from an Old World to New World migration, most likely across the Bering land bridge between Asia and North America (Hopkins 1967, Wen and Zimmer 1996, Wen et al. 1998, Wen 1999). This migration further supports the hypothesis that the ancestral range of *O. horridus* was in western North America.

The stem age of *O. horridus* is also required to appropriately determine the historical geological and ecological context that resulted in its phylogeographic history and current disjunct distribution. Specifically, it will inform our incorporation of the Pleistocene glacial cycles, which were a predominating force that determined and shifted plant distributions in North America (Deevey 1949). During the Last Glacial Maximum (LGM), ~22,000-18,000 years ago (Martinson et al. 1987, Yokoyama et al. 2000), the Laurentide and Cordilleran Ice Sheets in North America reached ~40° N (Ehlers and Gibbard 2007). Thus if *O. horridus* is older than the Pleistocene, the routes and timing of dispersal within North America were limited.

To date, the age of *O. horridus* has not been tested, and currently published chronograms that include *Oplopanax* do not include *O. horridus* with its sister species. In addition, chronograms of Araliaceae have revealed different histories between the chloroplast and the nuclear genomes. Previous papers revealed a hard polytomy in the early ancestry of the Asian Palmate Clade, most likely resulting from a radiation and reticulate evolution in the clade (Wen et al. 2001, Mitchell and Wen 2004, Plunkett et al. 2004). It is hypothesized that this early radiation led to the incongruence between chloroplast (cpDNA) and nuclear ribosomal (nrITS) markers. The incongruence is quite striking in the genus *Oplopanax*, which appears either sister to the rest of the clade and dated to ~69 Ma based on six cpDNA markers, or embedded within the clade and dated to ~30 Ma based on nrITS (Valcarcel et al. 2014). Another age estimate of

Oplopanax places it at 7.37 Ma (Wen et al. 2010 based on unpublished data), but it is unclear how this date was estimated. Recent papers have investigated the divergence of specific lineages within Araliaceae (Li and Wen 2013, Nicolas and Plunkett 2014, Valcarcel et al. 2014), resulting in published genetic markers across the family that can be harnessed to build a chronogram including *O. horridus* using a super-matrix approach.

Once an age is established, phylogeographic hypotheses can be tested. Based on evidence from a broader analysis of the western North America-GLR distribution (Marquis and Voss 1981, Chapter 1), there are three competing Pleistocene refugia hypotheses for how these disjunct populations arose: 1. Dispersal to the GLR along the edge of a receding glacier followed by local extirpation in the intermediate region due to restriction of cool moist climates to the GLR, 2. A widespread range pre-Pleistocene followed by survival in a Driftless Area refugium with subsequent recolonization into formerly glaciated Michigan and Ontario, and 3. Arrival in the GLR via long-distance dispersal. The highly restricted GLR populations of *O. horridus* with the remainder of the species in western North America poses an analytical challenge: how can multiple lines of evidence, incorporating timing, phylogenetics, population genetics and demographic modeling, tease apart the phylogeographic history of the disjunct Devil's-club? There are two alternative explanations for the *O. horridus* distribution that we do not strongly consider here. First, an east (Great Lakes) to west (western North America) migration for *O. horridus* is possible, but not plausible given the east Asian range of its sister taxa and its abundance in the west. Second, the possibility of human dispersal arises from the known medicinal and spiritual use of Devil's-club (Turner 1982, Gottesfeld 1992, Lantz et al. 2004). *Oplopanax horridus* is a rhizomatous, woody shrub with prickles along the length of the stems and leaf veins that produce an irritant, but its roots are harvested by at least 38 indigenous

linguistic groups to soothe ailments such as arthritis and influenza and for use in spiritual practices (Gottesfeld 1992, Lantz et al. 2004). This hypothesis of pre-settlement movement of Devil's-club by Native Americans should be addressed in the future with ethnobotanical research.

The phylogeographic history of *O. horridus* can also shed light on the sustainability of the isolated populations in the GLR. Recent studies have shown that the suitable climate space of taxa of the GLR is shifting dramatically based on projected climate models (Ash et al. 2017, Spalink et al. in press). Long-distance dispersal to the GLR or historical survival in a Driftless Area refugium would negatively impact the genetic diversity of these populations, potentially reducing their adaptive potential. This could render them more vulnerable to biotic and abiotic changes in the future. The diminishment or even loss of Devil's-club from its current GLR distribution has implications for cultural sustainability as well, since its use in medicinal and spiritual practices has been documented wherever it is found, including the populations in the GLR (Turner 1982, Lantz et al. 2004). Here we test the genetic diversity to look for geographic patterns and evidence of historical bottlenecks to assess future persistence in the GLR.

METHODS

Sampling: We collected leaf samples from 15-20 individuals in 15 populations across the North American restricted geographic range of *Oplopanax horridus*. Sampling was completed within four field seasons (2014-2017). Because *O. horridus* has a rhizomatous growth pattern, leaf tissue was collected at 5-meter intervals along a linear transect to avoid sampling potential clones. Leaf tissue was silica-dried. One herbarium voucher specimen was collected for each population, and they reside in the Wisconsin State Herbarium (WIS). We obtained leaf tissue of

the two remaining species restricted to east Asia, *O. japonicus* and *O. elatus* (courtesy of Jun Wen, Smithsonian Institute), to serve as outgroups.

DNA Extraction: All samples were disrupted with a TissueLyser (Qiagen, Valencia, CA, USA) using tungsten-carbide beads, and DNA was extracted using the Qiagen DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA) following the manufacturer's protocol. DNA integrity was confirmed by presence of high-mass bands on a 0.8% agarose gel, and DNA quantification was conducted with the Qubit dsDNA High-Sensitivity Assay Kit (Invitrogen, USA) or the AccuBlue High Sensitivity dsDNA Kit (Invitrogen, USA).

***Araliaceae chronogram*:**

Multi-gene alignments: Because of the reported conflict between the nuclear and plastid genomes, we used loci from both genomes to get age estimates of *O. horridus* using a super-matrix approach. We used PyPHLAWD (Smith and Brown 2018) to extract all available sequences from the Plant and Fungal NCBI database and build clusters of loci across Araliaceae, retaining one accession per species. Our clustering specifications were kept at the default settings, including sequence length >= 600, clustering length limit = 0.65, and percent identity = 20. We retained clusters of nrITS (18S + ITS1 + 5.8S + ITS2 + 28S) and 12 chloroplast markers (*trnL+trnL-F+trnF*, *rpl16*, *rps16*, *atpB-rbcL*, *ndhF*, *matK*, *rbcL*, *trnT-trnD*, *rpl32-trnL*, *rps16-trnK*, *ndhF-rpl32*, *trnS-trnfM*) that had at least 30 species. Cluster FASTA files were aligned using MUSCLE in Geneious v.10.1 (http://www.geneious.com, Kearse et al. 2012). We added three outgroups that were used in previous Araliaceae-wide phylogenies (Plunkett et a. 2004): *Apiopetalum velutinum* (Apiaceae), *Delarbrea paradoxa* (Myodocarpaceae) and *Myodocarpus*

fraxinifolius (Myodocarpaceae). Species were removed from the concatenated chloroplast alignment if they had > 75% missing data compared to the ungapped alignment consensus sequence.

Chronogram calibration: One of the most recent evaluations of fossils in Araliaceae can be found in the Nicolas and Plunkett (2014) evaluation of Apiales diversification and biogeography. They use representative taxon sampling (223 tips) with rpl16 and trnD-trnY-trnE-trnT markers and more reliable fossils than have been used in previous papers (e.g., Martinez-Millan 2010). They note that, as of yet, crown dates of Apiales have been estimated from somewhat ambiguous fossils, and that within-Apiales divergence times have remained largely unexamined. They use a fruit fossil of *Paleopanax oregonensis*, generally placed at the stem of *Metapanax*, dated at 44+/-3.2 Ma (Manchester 1994), and *Dendropanax eocenensis* leaf fossils (Dilcher and Dolph, 1970), all of which are in the Asian Palmate Group. They argue to conservatively place these fossils as a minimum crown age of 42.9 Ma (37.2 - 48.6 Ma 95% confidence interval) for the Asian Palmate Clade to account for the similarity between *D. eocenensis* and *Oreopanax* described in Dilcher and Dolph (1970). We used this calibration, as well as a secondary crown age of 94.76 Ma (82.56-106.80 Ma 95% confidence interval) for Araliaceae (Nicolas and Plunkett 2014). In addition, we used a 11.2 +/- 1.56 Ma crown age of the *Aralia elata* group (*A. elata*, *A. finlaysoniana*, *A. vietnamensis*) following Li and Wen (2013) and Valcarcel et al. (2014). This date is based on fossil seeds of *Aralia* that are similar to those of *Aralia elata* (Szafer 1961).

Previous nrITS analyses placed *Oplopanax*, diverging with *Fatsia* at ~30Mya, embedded within the Asian Palmate Clade (Varcarcel et al. 2014). The six cpDNA markers from the same study placed *Oplopanax* sister to rest of the Asian Palmate Clade, dated to ~69Mya. Combined

nrITS and cpDNA trees have been estimated (Li and Wen 2013), and we tested this combined approach as well.

Phylogenies: We ran the concatenated cpDNA data set and the nrITS data set separately and combined with partitions using RAxML v.8.2.10 (Stamatakis 2014) and MrBayes v.3.2.6 (Ronquist et al. 2012) on the CIPRES platform (Miller et al. 2010). In RAxML we used the GTR model of nucleotide substitution under a Γ model of rate heterogeneity with alpha estimated, and we used the rapid hill-climbing algorithm. In MrBayes, we specified one outgroup, *Myodocarpus fraxinifolius*, and ran a GTR model with 100 million generations and 25% burn-in. We ran the combined data set with partitions in BEAST v2.4.6 (Bouckaert et al. 2014) on the CIPRES platform. BEAST was run with GTR + Γ, a relaxed lognormal clock model, and a yule birth model. Tree likelihood calculation was set to treat ambiguities as equally likely. Fossil dates and secondary dates were given a lognormal and normal distribution, respectively. Outgroups, Araliaceae, the Asian Palmate Clade, and the *Aralia elata* group were each constrained to be monophyletic. We constrained the crown ages of Araliaceae, the Asian Palmate Clade, and the *Aralia elata* group to be 94.76 Ma, 42.9 Ma, and 11.2 Ma, respectively. The secondary Araliaceae calibration was given a normal distribution and the minimum fossil ages of the Asian Palmate Clade and the *Aralia elata* group were given a lognormal distribution. We evaluated the posterior distributions of the priors in Tracer v.1.6.0 (Rambaut et al. 2018) and calculated the maximum clade credibility tree using TreeAnnotator v.2.4.6 with 25% burn-in.

Genotyping-by-sequencing: We selected 4-9 individuals per population for genotyping-by-sequencing and included two samples of *O. elatus* and one sample of *O. japonicus*. Sample library prep and sequencing were conducted at the UW-Madison Biotechnology Center using the

ApeKI restriction digest enzyme and pooling samples 48-plex on 3 lanes of the Illumina Hi-Seq 2500 platform.

Bioinformatics: We ran cutadapt v.1.13 (Martin 2011) to trim off common adapters at the 3' end and retain only full-length reads. We ran ipyrad v.0.7.13 (Eaton 2014) steps 1 and 2 to demultiplex samples allowing 0 mismatches in barcodes, remove barcodes, then filter out reads with any low-quality bases (phred Q score < 20). We ran cutadapt again to trim off the 4 base pair overhang at the 5' end, resulting in 89bp read length. We removed populations with very low mean read coverage, including our second outgroup species, *O. japonicus*, and 14 populations remained in the final dataset, including our *O. elatus* outgroup samples (Table 1). We assembled reads using ipyrad and Stacks (Catchen et al. 2013) and ran them on the Condor computing cluster at the Center for High Throughput Computing facility at UW-Madison. We ran ipyrad in a Docker container (Merkel 2014) to ensure access to required dependencies.

Parameter testing: We tested three influential clustering parameters as described in Paris et al. (2017) and Rochette and Catchen (2017) (Stacks), and in Anderson et al. (2017) (PyRAD): minimum depth to make a statistical base call (m), maximum number of mismatches allowed to form a locus within an individual (M), and maximum number of mismatches allowed to align loci between individuals (n). We tested values 1-10 for the minimum depth to make a statistical base call. Results from this revealed that our data were at lower than 20X coverage at the default value (m = 2) (Appendix 4) and as expected, coverage was higher when requiring a higher minimum depth to retain a locus. Based on suggestions from the Stacks developer (pers. comm, Stacks Googlegroup), data with lower coverage should not be forced to a higher depth, and m should be kept at 2 or 3 to allow for statistical calls. We used m = 3, and tested M = n for values 1-9. We chose M = n = 4. The m, M, and n parameter values were set to maximize the number of

loci formed, stabilize the proportion of loci recovered for n-SNPs/locus, and minimize the proportion of heterozygotes at each base pair (Appendix 4).

Phylogeography: To resolve the relationships between populations of *O. horridus* and uncover which western North American population(s) were the source of the GLR populations, we estimated phylogenies using concatenated SNP data with Maximum Likelihood (ML) in RAxML and a multi-species coalescent approach in SVDQuartets v.1.0 (Chifman and Kubatko 2014) in PAUP v.4.0a150 (Swofford 2002). To eliminate the issue of allele phasing across concatenated loci, we used ambiguity codes to summarize the two alleles into one sequence per individual (INDIV), and we took a majority-rule consensus approach (CONS), which allows the use of heterozygous sites, but calls only one allele at each site. The majority-rule approach reduces ambiguity.

RAxML: To reduce the likelihood of overestimating branch lengths (Leaché et al. 2015), we implemented the Lewis ascertainment bias correction in RAxML to account for unobserved invariant sites. For the INDIV data set, we ran the populations module in Stacks on all samples to get an output of loci present in at least 75% of individuals, concatenating only one SNP per locus. We removed any samples that had >50% missing data and any samples that were the only representative from a population. To implement the ascertainment bias correction in RAxML we needed to remove invariant sites. Because the populations module includes Ns as variant sites, we changed Ns to gaps and masked invariant sites in Geneious. We changed the gaps back to Ns and ran RAxML on the CIPRES platform with the Lewis ascertainment bias correction, the GTR model of nucleotide substitution under a Γ model of rate heterogeneity with alpha estimated, and the rapid hill-climbing algorithm. We conducted 100 GTRCAT rapid bootstrap replicates. For

the CONS data set, we extracted the majority-rule consensus sequence for each population from the INDIV data set. We combined the population sequences without realigning since they were all uniform length. We replaced Ns with gaps to mask invariant sites, and because some ambiguous sites were being called as invariant, we also masked ambiguous sites. This resulted in a very short alignment, so we also ran the alignment without masking and without the ascertainment bias correction. We replaced gaps with Ns and ran both data sets in RAxML with the same settings.

SVDQuartets: To account for potential incomplete lineage sorting (ILS), we used SVDQuartets in Paup to estimate the history of each locus individually and obtain a multi-species coalescent tree. This program handles ambiguities by proportionally distributing the probability of allelic variants in the Q-matrix of each quartet, across all sites. This allows us to include all true allelic ambiguity. We ran populations with the outgroup to get an output of loci present in at least 75% of individuals, retaining one SNP per locus. We ran SVDQuartets with multiple samples assigned to a locality population, which samples only once from a population for each quartet. We used the multi-species coalescent tree model to evaluate all possible quartets and used the QFM quartet assembly to summarize quartets and obtain the bootstrap majority rule tree (Reaz et al. 2014). We distributed heterozygote ambiguity instead of masking it and we ran 100 bootstrap replicates on subsets of 100,000 quartets.

Population structure and genetic diversity: To estimate population structure and identify potential admixture between populations, we used Structure (Pritchard et al. 2000). This program uses and uses Hardy-Weinberg F_{ST} to construct probabilities of population assignment for each individual given a number of populations (K). We ran populations in Stacks without the

outgroup and retained only loci that were present in 100% of the populations and at least 75% of individuals within each population. We retained one SNP per locus to maximize independence between SNP loci. We ran fastStructure with the logistic setting to detect any subtle structure (Raj et al. 2014). We tested K values 2-10 and ran 5 iterations of each, with 10^6 generations. Following Barker et al. (2017) we tested consistency between runs of the same K value using CLUMPP v 1.1.2 (Jackobsson and Rosenberg 2007) to estimate the highest value of pairwise similarity (H' value). In CLUMPP we used the Greedy algorithm with 1000 random input orders for K 2-7 and the LargeKGreedy algorithm for K 8-10. For each of the five iterations, we ran the chooseK module in fastStructure, which reports the K values that give the highest marginal likelihood score and the number of populations used to explain the structure in the data. We selected the most common best K value reported from the five iterations and plotted the results of the mean population assignment determined in CLUMPP for that K value. We plotted the results using distruct2.2 (courtesy of Chhatre, 2016).

Because of the highly restricted range of *O. horridus* in the GLR, we wanted to compare the genetic diversity of these populations to the western North American populations to assess potential threats to future persistence in the GLR. We were also interested in these metrics for identifying a potential genetic bottleneck from a long-distance dispersal founder event or a widespread distribution that was restricted to the Driftless Area during the Pleistocene. We looked at nucleotide diversity (π), expected and observed heterozygosity (H_e and H_o), the inbreeding coefficient (F_{IS}), the number of private alleles, and the percent of polymorphic loci for each population. Because the Isle Royale populations that we collected in the GLR might be considered one population, we calculated these genetic diversity statistics for locality populations in Isle Royale individually, combined, and using only the largest representative population on

Passage Island. We ran the populations module in Stacks, grouping samples by their collection locality and retaining only loci that were present in 100% of the populations and at least 75% of individuals within a population. We ran these same filters coding the Isle Royale populations as one population. We ran diversity statistics in the populations module using kernel-smoothing and 100 bootstrap replicates. The F_{IS} statistic was calculated with the Bonferroni correction using a sliding window with the base p-value of 0.05. We performed unpaired, two-sample t-tests to compare mean values between the GLR and western populations.

Demographic modeling with DIYABC: To explicitly test between migration scenarios we ran Approximate Bayesian Computational (ABC) models using DIYABC v.2.1.0 (Cornuet et al. 2014). This version of DIYABC can use SNPs with missing data and allows users to set a minimum allele frequency (MAF) filter for the data set. DIYABC follows the standard ABC approach (Excoffier et al. 2005) of simulating data sets based on coalescent theory, selecting the simulated data sets that are closest to the observed data set measured by Euclidean distance, and estimating posterior distributions of parameters using local linear regression. When testing between scenarios, posterior probabilities are measured in two ways. The first is by the relative proportion of each scenario in the group of simulated data sets that are closest to the observed data set. The second is obtained by a logistic regression of each scenario's probability on the deviations between simulated and observed summary statistics (Fagundes et al. 2007, Beaumont 2002).

To prepare the SNP haplotypes input file, we ran the populations module in Stacks without outgroup samples and retained loci that were found in 75% of individuals. We extracted the haplotypes VCF file from Stacks. DIYABC requires there to be only polymorphic loci, so we

ran VCFtools v.0.1.13 (Danecek et al. 2011) to first remove outgroup samples and then remove monomorphic loci using the minimum allele frequency "--maf" flag set to 0.05, which is suggested in the DIYABC manual. Given the low mean coverage of our data this setting is likely filtering out monomorphic loci and not filtering out important minor alleles from the data set. We also filtered out sites for which all individuals in a population had missing data. We ran a python script courtesy of Etienne Loire (https://github.com/loire/vcf2DIYABC.py) to convert the VCF file into DIYABC SNP format.

To simplify the analyses, we grouped the major regions and ran DIYABC to test between three migration scenarios for the GLR populations: 1. Outward migration from the Pacific Northwest (PNW) to the North Northwest (NNW) followed by migration from the PNW to the GLR, 2. Migration from the PNW to the GLR followed by migration from the PNW to the NNW, and 3. Migration from the PNW to the NNW followed by migration from the NNW to the GLR (Fig. 1). Our model parameters for the first analysis were current population sizes of GLR, PNW, and NNW, the bottlenecked population sizes of PNW and NNW post divergence, the timing of the two divergence events, and the timing of the two bottleneck events. Constraints placed on any of the parameters had to be true across all models. Overall, we set 6 constraints: divergence 1 was older than divergence 2, the time of population bottleneck 1 <= divergence 1, the time of population bottleneck 2 <= divergence 2, and population size was smaller than the ancestral population size at bottlenecks 1, 2, and 3. The third bottleneck parameter was required for scenario 3 where one of the ancestral populations was different. The model parameters followed the same structure for the second analysis. We set log-uniform priors for our parameters. Mutation model selection is standardized for SNP data, based on bi-allelic data of ancestral and derived states in a coalescence model (Cornuet et al. 2014).

Our data filtering kept 6561 SNPs and we ran 100,000 simulations as recommended by the program based on our parameter complexity, distributed equally among scenarios (33,333 simulations per scenario). The mean values of the parameters from these simulated datasets were not significantly different from our observed data, so we analyzed the posterior probabilities of the three scenarios with direct comparison and the logistic regression to determine the best scenario.

RESULTS

***Araliaceae phylogeny and chronogram*:** The nrITS and cpDNA ML trees each showed the same topological incongruence as previously published trees. The nrITS tree placed *Oplopanax* sister to *Fatsia* embedded within the Asian Palmate Clade, and the cpDNA tree and cpDNA +nr ITS tree placed it sister to the rest of the Asian Palmate Clade (not shown). The support across the backbone of Araliaceae is negligible for all three data sets. Our Bayesian consensus trees had the same topological incongruence based on which data set was used, and the posterior probabilities across the backbone for combined analysis were higher (Appendix 5). The BEAST chronogram from the combined data set was poorly supported along the backbone of the Asian Palmate Clade, which is consistent with previous studies that have suggested an early radiation (Fig. 2). Our chronogram placed two *Schefflera* species as sister to the rest of the Asian Palmate Clade with high support. Although *Schefflera* is comprised of two non-monophyletic clades, the placement of these two species fell outside of both clades. *Oplopanax* diverged from one species of *Oreopanax* (~21 Ma), which is also falling outside of its monophyletic group. This age is younger than the ages previously reported from nrITS, cpDNA, or combined analyses. These are sister to the rest of the Asian Palmate Clade. The crown age of *O. horridus* and *O. elatus*, which

were the two that remained in our dataset post-filtering, is ~4 Ma (0-7.5 Ma 95% HPD), which is most consistent to the 7.37 Ma age for *Oplopanax* reported by Wen et al. (2010). This suggests a pre-Pleistocene origin of *O. horridus*, but is not conclusive given the absence of *O. japonicus* from the analysis.

***Phylogeography of* Oplopanax horridus:** *RAxML*: Our INDIV data set was 24,240 concatenated SNPs. Our CONS data sets with masking and without masking were 5,636 SNPs and 26,128 SNPs respectively. The RAxML tree with one individual per tip resulted in near monophyly of each locality population (Fig. 3). It shows high support for the most southern western population, Willamette, sister to the rest of the populations. It also shows high support for a clade of populations from Alaska, British Columbia and the GLR. The CONS tree without invariant and ambiguous sites and estimated with the ascertainment bias correction shows a similar pattern, with lower support for Willamette as sister to the rest, but indicating a grade of Pacific Northwest populations (Willamette, Stevens, and Cascade) and a clade of the GLR populations sister to Alaskan and British Columbia populations with high support (Fig. 4a). The CONS tree with more data and without ascertainment bias correction had the same topology and comparable branch lengths as the first CONS tree with slightly lower support but recovering the full support for Willamette sister to the rest of the populations (Fig. 4b). In addition, these CONS trees show the Alaskan populations embedded within a well-supported grade of British Columbia populations. These results support a hypothesis of post-glacial northward dispersal from the southern-most western North American populations, but this topology does not fully resolve the ancestral range of the GLR populations.

SVDQuartets: Our data set, which included two outgroup samples of *O. elatus*, was 26,128 SNPs. The outgroups had comparable missing data to the rest of the samples (See Chapter 2), and there is no known hybridization between *O. elatus* and *O. horridus*, so we included them in the analysis. The SVDQuartet tree recovers a slightly different topology than the RAxML tree (Fig. 5). Instead of a clear grade of western North American populations from Oregon up through Alaska, the topology opens up the possibility that the ancestral range of the British Columbia populations was in Alaska, perhaps in a glacial refugium. In this analysis, the GLR populations clearly originate from the NNW populations, which was suggested, but not confirmed by the RAxML analysis. If we assume that the NNW populations, specifically those in British Columbia, established post-ice melt, this indicates a post-glacial migration route into the GLR. However, this should be confirmed with additional dating and demographic analysis.

***Population structure*:** Assignment probabilities for all K values except K = 2 were inconsistent among the 5 iterations testing K = 2-10 (Appendix 6). Despite this, the chooseK module in fastStructure chose K = 3 and K = 4 for three and two of the iterations, respectively, indicating that although fastStructure had difficulty consistently assigning individuals to populations for most values of K, 3 or 4 populations was the best model complexity to explain the data. Using the mean population assignment from CLUMPP helped to smooth the inconsistency, and we plotted results for K = 3 and K = 4 (Fig. 6a-b). For both K = 3 and K = 4, the major population groupings were: western North America, south of the glacier line in Oregon and Washington; western North America, north of the glacier line in British Columbia and Alaska; and the GLR Isle Royale populations. For K = 4 the additional population assignment refined the admixture between the two western groups, but it did not add a stand-alone population. There is admixture

at the glacial boundary between the Washington and British Columbia populations, which is more noticeable from the additional grouping in the K = 4 plot. In addition, both plots show signal of fixation of rare genotypic combinations in the Isle Royale populations that are also found in the two western regions.

***Genetic diversity*:** When coding each locality from Isle Royale as a separate population, mean values for all of the genetic diversity metrics (π, H_e, H_o, F_{IS}, private alleles, and percent polymorphic loci) were significantly lower than those from western North America, except for the inbreeding coefficient (F_{IS}) (Table 2b). When coding all localities as one population, or when using the largest and most isolated population, Passage Island, as the representative population, the differences in the metrics are no longer significant (Table 2c). This indicates that the small clusters of individuals found on small islands are bringing down the mean value of these metrics for Isle Royale as a whole. The low genetic diversity of these individual islands can be explained by localized founder events with minimal subsequent gene flow due to isolation and the prolonged time to establishment characteristic of *O. horridus*. In contrast, however, the larger, more isolated Passage Island, which is northeast of the smaller populations, does not show any significant difference in genetic diversity metrics compared to the west. This pattern of genetic diversity within Isle Royale points to Passage Island as the ancestral population of the smaller islands and suggests the absence of a historical genetic bottleneck in the GLR.

***Demographic modeling*:** DIYABC used 3317 of the 6561 sites, removing sites that contained more than 2 alleles. Of the three scenarios tested, the scenario with the highest support was dispersal from the PNW to the NNW, followed by dispersal from the NNW to the GLR.

DISCUSSION

***Effects of an early radiation on the dating of* Oplopanax horridus:** The nrITS topology placing *O. horridus* sister to *Fatsia* in the RAxML and MrBayes trees is not well supported. This is apparent when the signal is lost in the combined nuclear and cpDNA analysis, placing *O. horridus* sister to the rest of the Asian Palmate Clade. nrITS is used for phylogenetic reconstruction because it evolves rapidly enough to provide synapomorphies, and it is conserved among taxa. However, nrDNA is also prone to duplication, which can result in multiple copies that are released from constraint with subsequent multiple-hit mutations that can blur the ancestral history of the species. Because of the rapid radiation along the backbone of the Asian Palmate Clade and this potential loss of signal in ITS, the evolutionary history of the Asian Palmate Clade will need further analysis with targeted sequence capture and Bayesian concordance analysis. Despite the incongruence that we found between the cpDNA and the nrITS data in our RAxML and MrBayes trees, our age estimate of *O. horridus* divergence with *O. elatus* from the combined data set is congruent with the age of *Oplopanax* previously reported (Wen et al. 2010). We believe this dating analysis confidently places the age of *O. horridus* in a pre-Pleistocene context, but further work should be done to tease apart genomic incongruence and add *O. japonicus* to the data set.

***Post-glacial migration*:** Our population phylogenetic analysis reveals an ancestral range in western North America south of the glacier line, with subsequent dispersal northward. Major areas in Alaska remained ice-free during the last few glacial cycles (Brigham-Grette et al. 2003), and Beringia was proposed as a refugium for arctic plants (Hultén 1937, Nimis et al. 1998). Our phylogenetic and structure analyses cannot distinguish between a stepwise northward dispersal

and the possibility of dispersal from the southern ancestral range to Alaska with subsequent re-colonization into British Columbia. To test between these alternative migration scenarios, further demographic work that tests for dispersal events and bottleneck events should be done in addition to population divergence dating. The SVDQuartet analysis supported dispersal from British Columbia to the GLR, but the RAxML analysis was not as conclusive. Our demographic analysis, however, strongly supports the SVDQuartet results, and reveals that the GLR dispersed from the northwest American group of British Columbia + Alaska populations.

There is one caveat for our treatment of the geographic groups in our demographic models that needs to be considered. Our analyses were simplified for computational feasibility. To do this, we grouped populations based on their geographic clustering and population structure, but ideally this grouping would also be based on monophyly. The NNW group was a grade embedded within a grade of PNW populations according to the RAxML (in part) and SVDQuartets analyses, and there is potential concern that this would bias the model performance against dispersal from the PNW to the GLR in the first analysis. This can be addressed in the future by randomly sampling single representatives from the PNW and the NNW to run the demographic models.

The phylogenetic, population structure, and demographic analyses illuminated where the GLR populations came from, and that they arrived there post-glacially. However, our genetic diversity metrics provided the most insight into how they got there. Our t-tests rejected any significant difference in the genetic diversity metrics between the GLR and western North America populations, when combining the Isle Royale populations or when looking at Passage Island alone. This is remarkable given the current isolation of these populations that likely results in low levels of gene flow. This level of genetic diversity points to a lack of a bottleneck event in

the GLR, rejecting hypotheses of long-distance dispersal or population restriction in a local glacial refugium. However, this should be interpreted with caution since we did not directly test for a bottleneck event with the allele frequency spectrum or its one-dimensional summary statistic, Tajima's D.

Each of the genetic diversity statistics should be interpreted differently with regards to a bottleneck event. Private alleles tend to be removed during a dispersal event, and we would expect a low number of private alleles in a recently established population, especially from a bottleneck event. The high number of private alleles in Passage Island is not necessarily evidence against a bottleneck, however, because it could be due to mutation accumulation over time, even after a bottleneck event. Likewise, a high, or comparable nucleotide diversity could be high due to a longer period since dispersal. In contrast, the measures of heterozygosity and the inbreeding coefficient are those that most support the absence of a bottleneck event. These metrics would be low and difficult to recover from after a severe reduction in population size without additional gene flow into the population. Therefore, these two metrics point to a route other than long-distance dispersal or a midwestern glacial refugium in the GLR.

There are two main competing hypotheses for the alternative route to the GLR. First, the medicinal properties of *O. horridus* have been utilized by indigenous groups for many years (Lantz et al. 2004). The earliest ethnographic record dates to 1842 (Blaschke 1842, Lantz et al. 2004), but we do not know how far into the past its traditional use extends. The root stock is regularly harvested from Isle Royale, and it is possible that this plant was historically transported by people to this region. The north-south movement of plants over long distances along the east coast of North America during the Holocene was studied and recently reviewed (MacDougall 2003). Although they did not find that movement of native plants was widespread and frequent,

they suggest that the amount of missing data in their study leaves this hypothesis open for investigation. They suggest that if plants were transported by Native Americans, this would have occurred between 6000-3000 BP, when people began to disseminate into glaciated regions in northeast North America. Movement of *O. horridus* from multiple plant stocks might have resulted in the higher than expected heterozygosity and inbreeding coefficient in the GLR. This hypothesis will need to be explored with ethnobotanical work and more precise dating methods. The alternative competing hypothesis is a peri-glacial route from British Columbia. In this scenario, the lack of intermediate populations might be explained by early peri-glacial dispersal, followed by extirpation via the formation of Lake Agassiz. This glacial lake was at its greatest extent ~13,000 years ago and covered parts of Manitoba, Ontario, Minnesota, the Dakotas, and Saskatchewan (Teller et al. 1983). In the future, this could be tested with a robust dating analysis of population divergence that uses full GBS loci and incorporates absolute time with secondary dates outside of *Oplopanax* (Eaton et al. 2016). However, the lack of geographically intermediate populations between western North America and the GLR may continue to impede the unravelling of the historical phylogeography of Devil's-club.

CONCLUSION

Although this particular case study of the western North America-GLR disjuncts posed a perplexing analytical challenge because of its restricted range, we were able to uncover major elements of the historical phylogeography of *O. horridus*. Our study provides evidence of a pre-Pleistocene origin of *O. horridus* and an ancestral range in western North America south of the glacial line that re-colonized northward into previously glaciated territory. We propose that further collaborative work needs to be done to assess the divergence of *O. horridus* from its

sister taxa to fully understand the complex evolutionary context of the Asian Palmate Clade. In addition, peri-glacial dispersal or human-assisted migration from British Columbia to the GLR still needs to be rigorously tested with robust population dating and ethnobotanical analysis.

We have shown that the genetic diversity of the populations in the GLR is comparable to that of the western populations. This remarkable feature might mitigate imminent threats of climate change, but the low number of populations as well as their restricted range in Lake Superior pose additional challenges for the survival of this species in the GLR. We propose that its suitable niche space should be investigated using climate, soil, and elevation data, with a focus on the GLR, to assess whether and where there will be pockets of suitable habitat for *O. horridus* in the future.

Table 1: List of populations sampled for genotyping-by-sequencing, their locality, and the number of samples per population in the final, cleaned data set. The population names have the state/province appended to the beginning for easier interpretation of their general location.

Name	Locality	Number of samples in final data set
IR_BlakesPt	48.189762, -88.423502	9
IR_BoysIs	48.180963,-88.433341	8
IR_EdwardIs	48.172375, -88.436428	6
IR_Ngov	48.179279, -88.421539	7
IR_Sgov	48.169138, -88.423289	8
IR_PassageIs	48.230417, -88.358020	11
AK_Riverview	61.283238, -149.491896	11
AK_Kachemak	59.622336, -151.189567	12
BC_McLeod	54.903864, -122.935420	11
BC_MtRob	53.058130, -119.212323	4
WA_Stevens	47.751218, -121.086060	12
WA_Cascade	48.475345, -121.073577	12
OR_Willamette	44.395234, -122.137903	10

Table 2a. The genetic diversity metrics for each population based on all loci that were present in 100% of the populations and at least 75% of individuals within a population: number of private alleles, percent of loci that were polymorphic, observed and expected heterozygosity, nucleotide diversity (Pi), and the inbreeding coefficient (F_{IS}). Diversity statistics were measured with kernel-smoothing and 100 bootstrap replicates. F_{IS} was calculated with the Bonferroni correction with a base p-value of 0.05. **2b.** The comparison between the Isle Royale populations as a whole and western North American populations and **2c.** The comparison between Passage Island alone and western North American populations. P-values are from unpaired, two-sample t-tests.

Table 2a

Pop_ID	# of Sites	# Private Alleles	% Polymorphic Loci	Coded Region
BlakesPoint	1564598	123	0.18842	GLR
BoysIs	1564598	96	0.1835	GLR
CascadePass	1564598	677	0.30864	West
EdwardIs	1564598	56	0.17934	GLR
KachInland	1564598	284	0.23546	West
McLeod	1564598	264	0.25892	West
MtRob	1564598	113	0.229	West
Ngov	1564598	75	0.17519	GLR
PassageIs	1564598	93	0.19366	GLR
Riverview	1564598	350	0.24275	West
Sgov	1564598	140	0.17167	GLR
Stevens	1564598	431	0.31139	West
Willamette	1564598	544	0.30912	West

Table 2a (cont'd)

Pop_ID	Obs_Het	Exp_Het	Pi	F_{IS}
BlakesPoint	0.00159	0.00086	0.00091	-0.00128
BoysIs	0.00154	0.00082	0.00087	-0.00122
CascadePass	0.0018	0.00113	0.00119	-0.00114
EdwardIs	0.00158	0.00084	0.00092	-0.00119
KachInland	0.00169	0.00097	0.00102	-0.00128
McLeod	0.00178	0.00106	0.00112	-0.00125
MtRob	0.00177	0.00103	0.00119	-0.001
Ngov	0.00156	0.00082	0.00088	-0.00125
PassageIs	0.00163	0.00089	0.00094	-0.00131
Riverview	0.0017	0.00098	0.00104	-0.00126
Sgov	0.00152	0.00079	0.00085	-0.00124
Stevens	0.00191	0.00119	0.00124	-0.00126
Willamette	0.00203	0.001275	0.00143	-0.00106

Table 2b

	Mean Isle Royale Value	Mean Western Value	p-value
# Private Alleles	97.16667	380.42857	0.004007**
% Polymorphic Loci	0.1819633	0.2707543	0.0001513**
Obs_Het	0.00157	0.00181143	0.0006854**
Exp_Het	0.00083667	0.00109071	0.0002691**
Pi	0.000895	0.00117571	0.000539**
F_{IS}	-0.0012483	-0.0011786	0.1819

Table 2c

	Mean Passage Island	Mean Western Value	p-value
# Private Alleles	1100	564	0.07338
% Polymorphic Loci	0.24585	0.2919912	0.3275
Obs_Het	0.00154	0.0018525	0.07667
Exp_Het	0.00089	0.00115875	0.1019
Pi	0.0009	0.00125375	0.09334
F_{IS}	-0.00121	-0.0011	0.3908

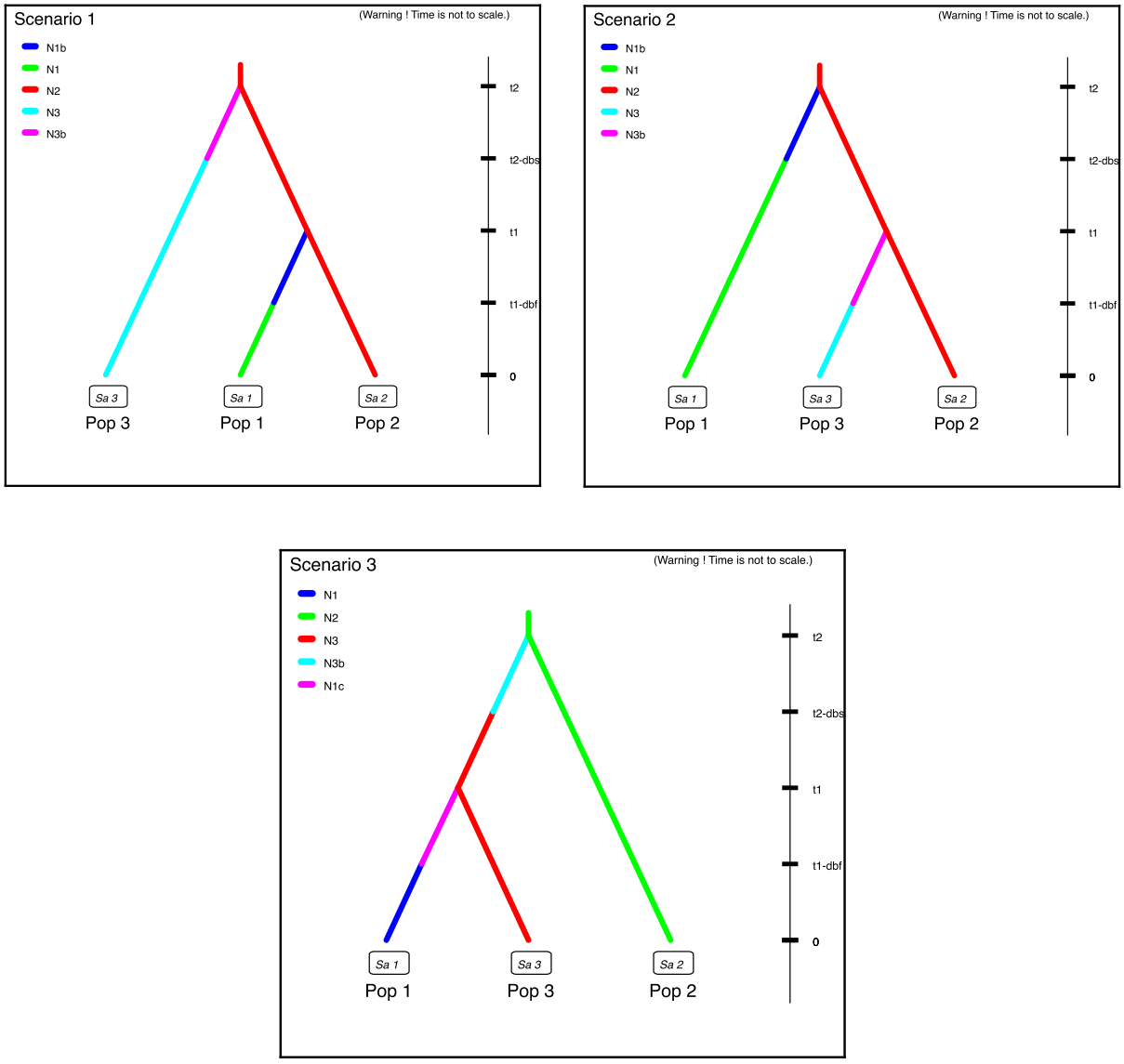
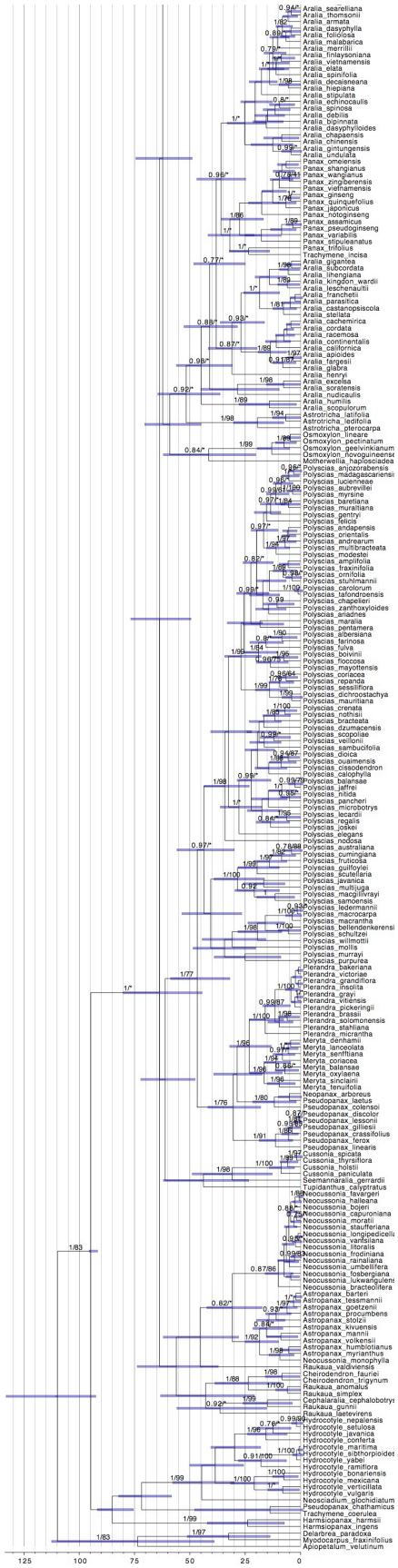
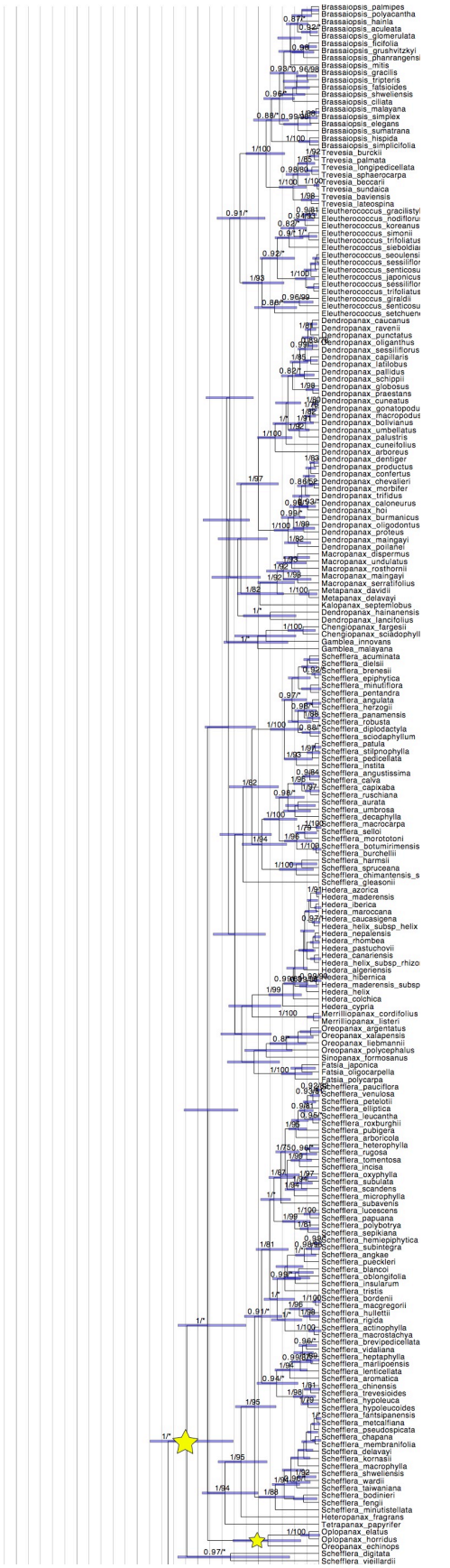


Fig. 1 DIYABC scenarios testing dispersal to the GLR. Pop 1 = GLR, Pop 2 = PNW, Pop 3 = NNW. Scenario 1: dispersal from Pop 2 to Pop 3, then dispersal from Pop 2 to Pop 1. Scenario 2: dispersal from Pop 2 to Pop 1, then dispersal from Pop 2 to Pop 3. Scenario 3: dispersal from Pop 2 to Pop 3, then dispersal from Pop 3 to Pop 1.



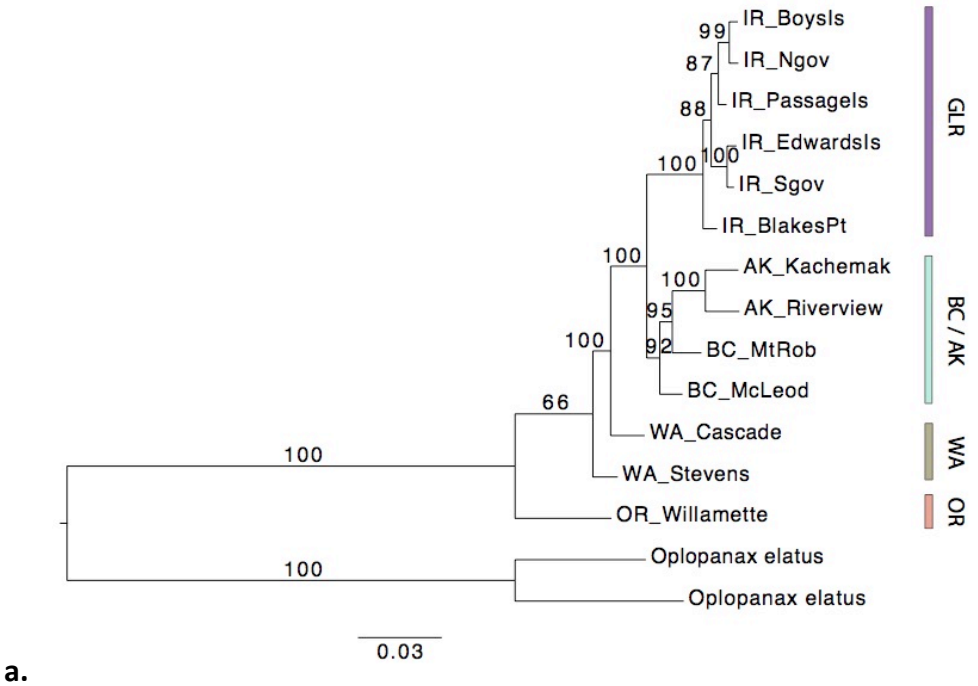
a)



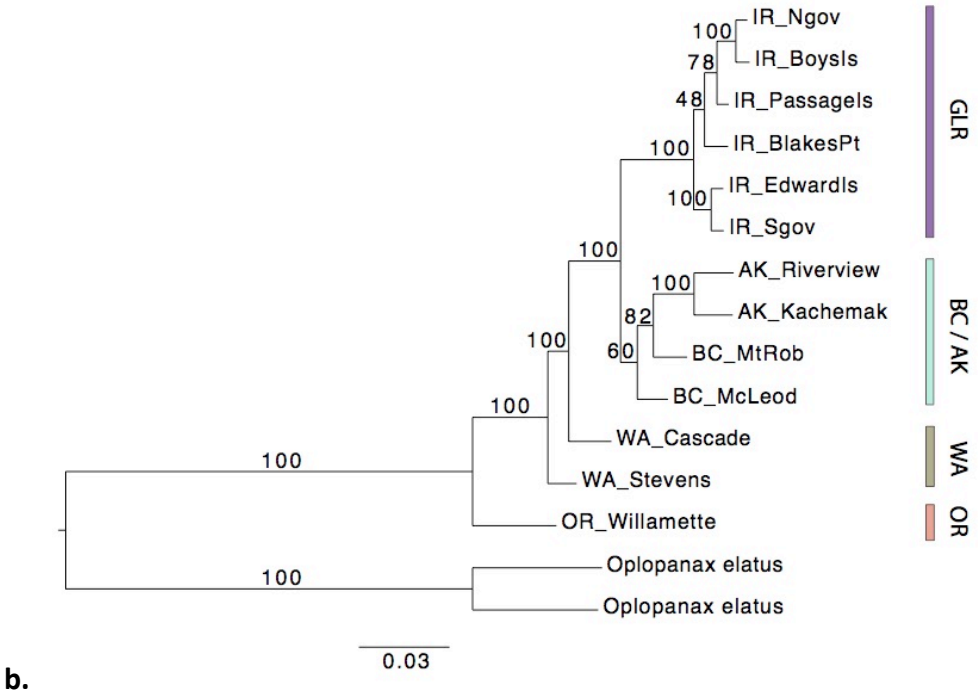
b)

Fig. 2 (Above) BEAST chronogram based on nrITS + 12 cpDNA makers and 100 million MCMC generations with 25% burn-in. We placed secondary calibration (94.76 Ma)for the crown of Araliaceae using a normal distribution, a fossil constraint (42.9 Ma) on the crown of the Asian Palmate Clade using a lognormal distribution (indicated by large star), and a fossil constraint (11.2 Ma) on the crown of the *Aralia elata* group. Smaller star indicates the stem age of *Oplopanax*.

Fig. 3 RAxML tree estimated with GTR + Γ using the INDIV data set of 24,240 concatenated SNPs and implementing the Lewis ascertainment bias correction. Bootstrap values above 70 are shown above the branches. Major clades are: Alaska + British Columbia + Isle Royale, and OR_Willamette sister to all other populations. Hash marks indicate a truncated long branch.



a.



b.

Fig. 4a RAxML tree using the CONS data set of 5,636 concatenated, majority rule SNPs for each site within a population, and implementing the Lewis ascertainment bias correction. **4b** RAxML tree using the CONS unmasked data set of 26,128 concatenated, majority rule SNPs for each site within a population, without ascertainment bias correction. Both trees were estimated with GTR + Γ and bootstrap values are reported above the branches. Salmon = Oregon, green = Washington, blue = AK embedded within BC, purple = GLR.

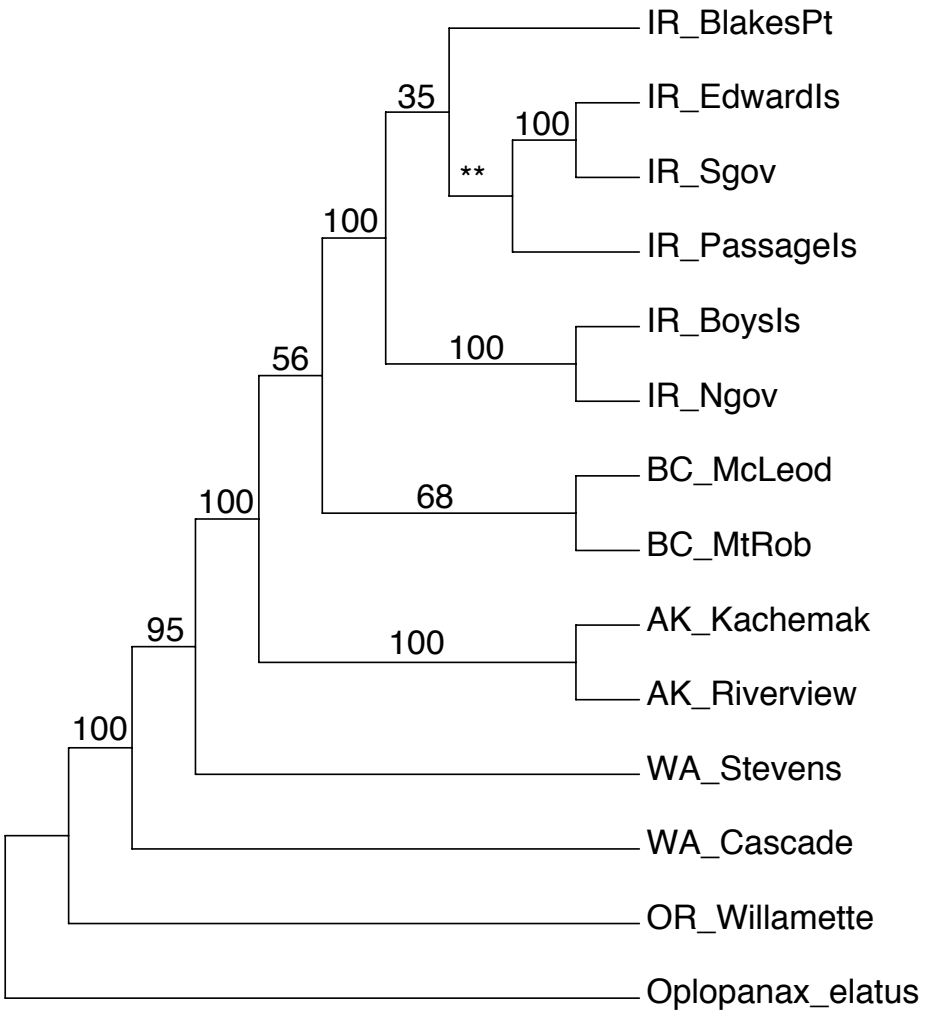


Fig. 5 Paup SVDQuartet coalescence tree using the CONS, unmasked data set of 26,128 SNPs with the two *O. elatus* outgroup samples. Bootstrap values are from 100 bootstrap replicates on 100,000 sub-sampled quartets. Low support for some of the clades indicates conflict between SNPs. Asterisks indicate a different topology in the majority bootstrapped tree than the tree derived from summarizing all possible quartets. Topology points to BC as the ancestral range of the GLR populations.

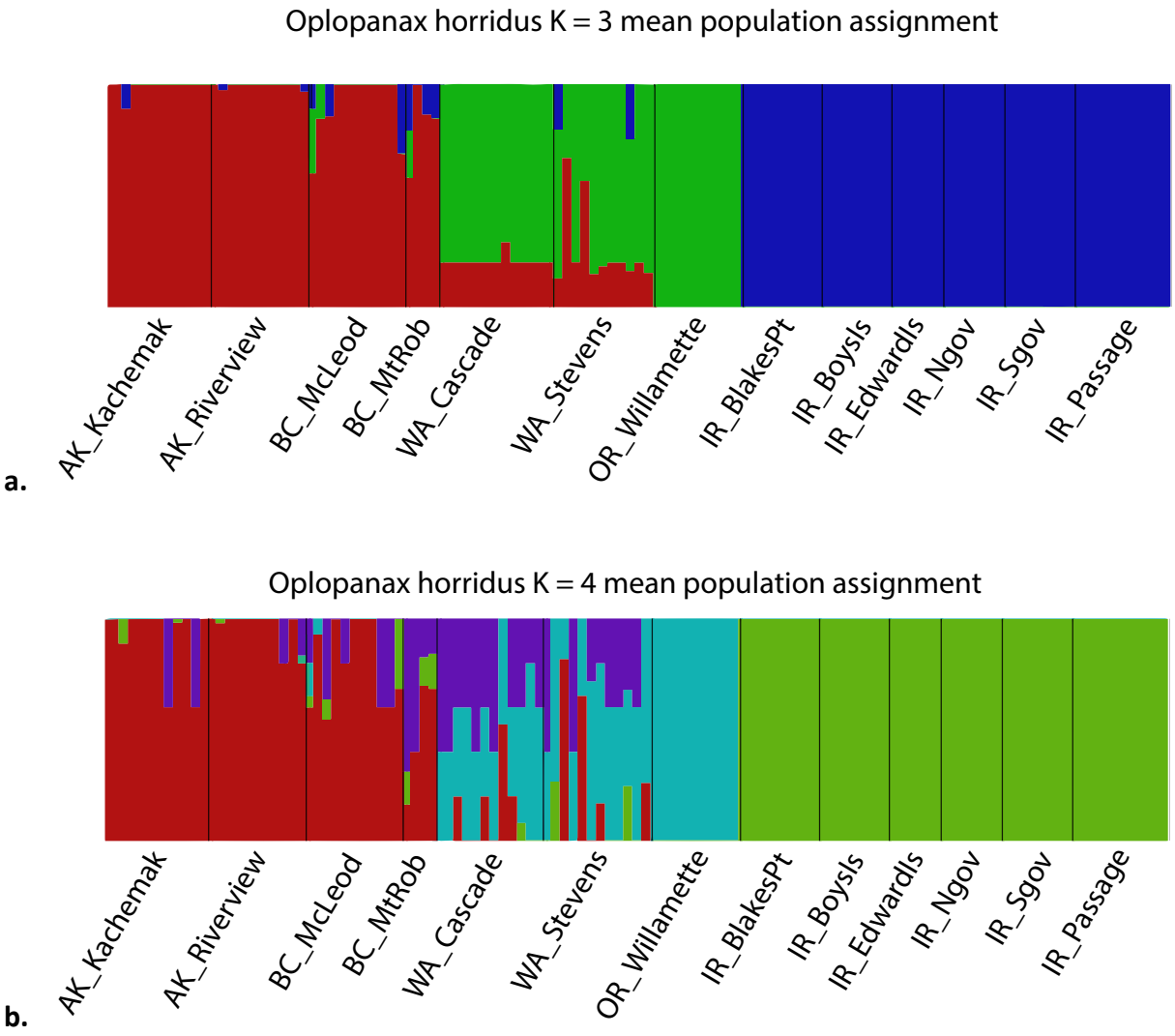


Fig. 6 Results from fastStructure with K = 3 **(a)** and K = 4 **(b)**. Mean population clusters were generated in CLUMPP, averaging across five iterations of fastStructure. Populations in the plot are generally ordered West to East and are in the same order for both plots, although different colors are used. Plots for both K values show three distinct populations and the purple color in K = 4 reveals additional admixture.

REFERENCES

Anderson, B.M. et al., 2017. Genotyping-by-Sequencing in a Species Complex of Australian Hummock Grasses (Triodia): Methodological Insights and Phylogenetic Resolution G. Sun, ed. *PLoS ONE*, 12(1), pp.e0171053–34.

Ash, J.D., Givnish, T.J. & Waller, D.M., 2016. Tracking lags in historical plant species' shifts in relation to regional climate change. *Global Change Biology*, 23(3), pp.1305–1315.

Association, D.S.S. et al., *PAUP* Phylogeny Analysis Using Parsimony (* and other methods), version 40b10*.

Barker, B.S. et al., 2017. Population genomic analyses reveal a history of range expansion and trait evolution across the native and invaded range of yellow starthistle (Centaurea solstitialis). *Molecular Ecology*, 26(4), pp.1131–1147.

Beaumont, M.A., Zhang, W. & Balding, D.J., 2002. Approximate Bayesian computation in population genetics. *Genetics*, 162(4), pp.2025–2035.

Blaschke E. 1842. *Topographia Medica Portus Novi-Archangelscensis*. Wiehoberi, St. Petersburg.

Bouckaert, R. et al., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. A. Prlic, ed. *PLoS computational biology*, 10(4), p.e1003537.

Brigham-Grette, J. & Gualtieri, L.M., 2003. Chlorine-36 and 14C chronology support a limited last glacial maximum across central Chukotka, northeastern Siberia, and no Beringian ice sheet. *Quaternary*, 59, pp.386–398.

Catchen, J. et al., 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22(11), pp.3124–3140.

Chhatre, 2016, http://www.crypticlineage.net/pages/distruct.html

Chifman, J. & Kubatko, L., 2014. Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, 30(23), pp.3317–3324.

Cornuet, J.-M. et al., 2014. DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, 30(8), pp.1187–1189.

Danecek, P. et al., 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15), pp.2156–2158.

Deevey Jr, E.S., 1949. Biogeography of the Pleistocene: Part I: Europe and North America. *Geological Society of America Bulletin*, 60(9), pp. 1315-1416.

Dilcher, D.L. & Dolph, G.E., 1970. Fossil leaves of Dendropanax from Eocene Sediments of Southeastern North America. 57(2), pp.153–160.

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

Eaton, D.A.R. et al., 2016. Misconceptions on Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants. *Systematic Biology*, 56, p.syw092.

Ehlers, J. & Gibbard, P.L., 2007. The extent and chronology of Cenozoic Global Glaciation. *Quaternary International*, 164-165, pp.6–20.

Etienne Loire. vcf2DIYABC.py open source script. (https://github.com/loire/vcf2DIYABC.py)

Excoffier, L., Estoup, A. & Cornuet, J.-M., 2005. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, 169(3), pp.1727–1738.

Fagundes, N.J.R. et al., 2007. Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences*, 104(45), pp.17614–17619.

Gottesfeld, L.M.J., 1992. La importáncia de los productos de corteza en las economías indígenas del Noroeste de la Columbia Británica en Canada. *Economic Botany*, 46(2), pp.148–157.

Hopkins, David Moody, ed. *The Bering land bridge*. Vol. 3. Stanford University Press, 1967.

J, M.R. & G, V.E., 1981. Distributions of some western North American plants disjunct in the Great Lakes region. *Michigan Botanist*.

Jakobsson, M. & Rosenberg, N.A., 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23(14), pp.1801–1806.

Jansen, R.K. et al., 1998. Phylogeny and biogeography of Aralia Sect. Aralia (Araliaceace). 85(6), pp.866–875.

Kearse, M. et al., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), pp.1647–1649.

Lantz, T., Swerhun, K, & Turner, N.J. 2004. Devil's club (Oplopanax horridus): An ethnobotanical review. *Herbalgram*, 62, pp33-48.

Leaché, A.D. et al., 2015. Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. *Systematic Biology*, 64(6), pp.1032–1047.

Li, R. & Wen, J., 2013. Phylogeny and Biogeography of Dendropanax (Araliaceae), an Amphi-Pacific Disjunct Genus Between Tropical/Subtropical Asia and the Neotropics. *Systematic Botany*.

MacDougall, A., 2003. Did Native Americans influence the northward migration of plants during the Holocene? *Journal of Biogeography*, 30(5), pp.633–647.

Manchester, Steven R. *Fruits and seeds of the middle Eocene nut beds flora, Clarno Formation, Oregon*. Paleontological Research Institution, 1994.

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

Martínez-Millán, M., 2010. Fossil Record and Age of the Asteridae. *The Botanical Review*, 76(1), pp.83–135.

Martinson, D., N. Pisias & J. Hays., 1987. Age Dating and the Orbital Theory of the Ice Ages: Development High-Resolution 0 to 300,000-Year Chronostratigraphy. Quaternary Research 29:1–29.

Merkel, D., 2014. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014(239), p2.

Miller, M.A., Pfeiffer, W. & Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In 2010 Gateway Computing Environments Workshop (GCE). IEEE, pp. 1–8.

Mitchell, A. & Wen, J., 2004. Phylogenetic Utility and Evidence for Multiple Copies of Granule-Bound Starch Synthase I (GBSSI) in Araliaceae. *Taxon*, 53(1), p.29.

Nimis, P.L. et al., 1998. A multivariate phytogeographic analysis of plant diversity in the Putorana Plateau (N. Siberia). *Opera Botanica*, 136.

Paris, J.R., Stevens, J.R. & Catchen, J.M., 2017. Lost in parameter space: a road map for stacks S. Johnston, ed. *Methods in Ecology and Evolution*, 8(10), pp.1360–1373.

Plunkett, G.M. & Nicolas, A.N., 2014. Diversification Times and Biogeographic Patterns in Apiales. *The Botanical Review*, 80(1), pp.30–58.

Plunkett, G.M., Wen, J. & Lowry, P.P., II, 2004. Infrafamilial classifications and characters in Araliaceae: Insights from the phylogenetic analysis of nuclear (ITS) and plastid (trnL-trnF) sequence data. *Plant Systematics and Evolution*, 245(1-2), pp.1–40.

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

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

Rambaut, A. et al., 2018. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. E. Susko, ed. *Systematic Biology*, 22, p.341.

Reaz, R., Bayzid, M.S. & Rahman, M.S., 2014. Accurate Phylogenetic Tree Reconstruction from Quartets: A Heuristic Approach R. Wu, ed. *PLoS ONE*, 9(8), pp.e104008–13.

Rochette, N.C. & Catchen, J.M., 2017. Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Publishing Group*, 12(12), pp.2640–2659.

Ronquist, F. et al., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3), pp.539–542.

Smith, S.A. & Brown, J.W., 2018. Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany*, 105(3), pp.302–314.

Spalink, D. et al. Spatial phylogenetics reveals evolutionary constraints on the assembly of a large regional flora. *In press*.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), pp.1312–1313.

Szafer, W. 1961. Miocene flora from Stare Gliwice in upper Silesia. *Institut Geologiczny Prace* 33: 19–20.

Teller, J.T. et al.,1983. Maximum extent and major features of Lake Agassiz. *Geological Association of Canada, Special Paper*, 26, pp. 43-45.

Thule, E.H.S.B.-L.A.1937, *Outline of the history of Boreal and Arctic biota during the Quaternary Period*,

Turner, N.J. 1982. Traditional use of devil's-club (Oplopanax horridus; Araliaceae) by native peoples in western North America. *J Ethnobiol*, 3, pp.17-38.

Valcárcel, V., Fiz-Palacios, O. & Wen, J., 2014. The origin of the early differentiation of ivies (Hedera L.) and the radiation of the Asian Palmate group (Araliaceae). *Molecular Phylogenetics and Evolution*, 70, pp.492–503.

Wen, J., 1999. Evolution of eastern Asian and eastern North American disjunct distributions in flowering plants. 30(1999), pp.421–455.

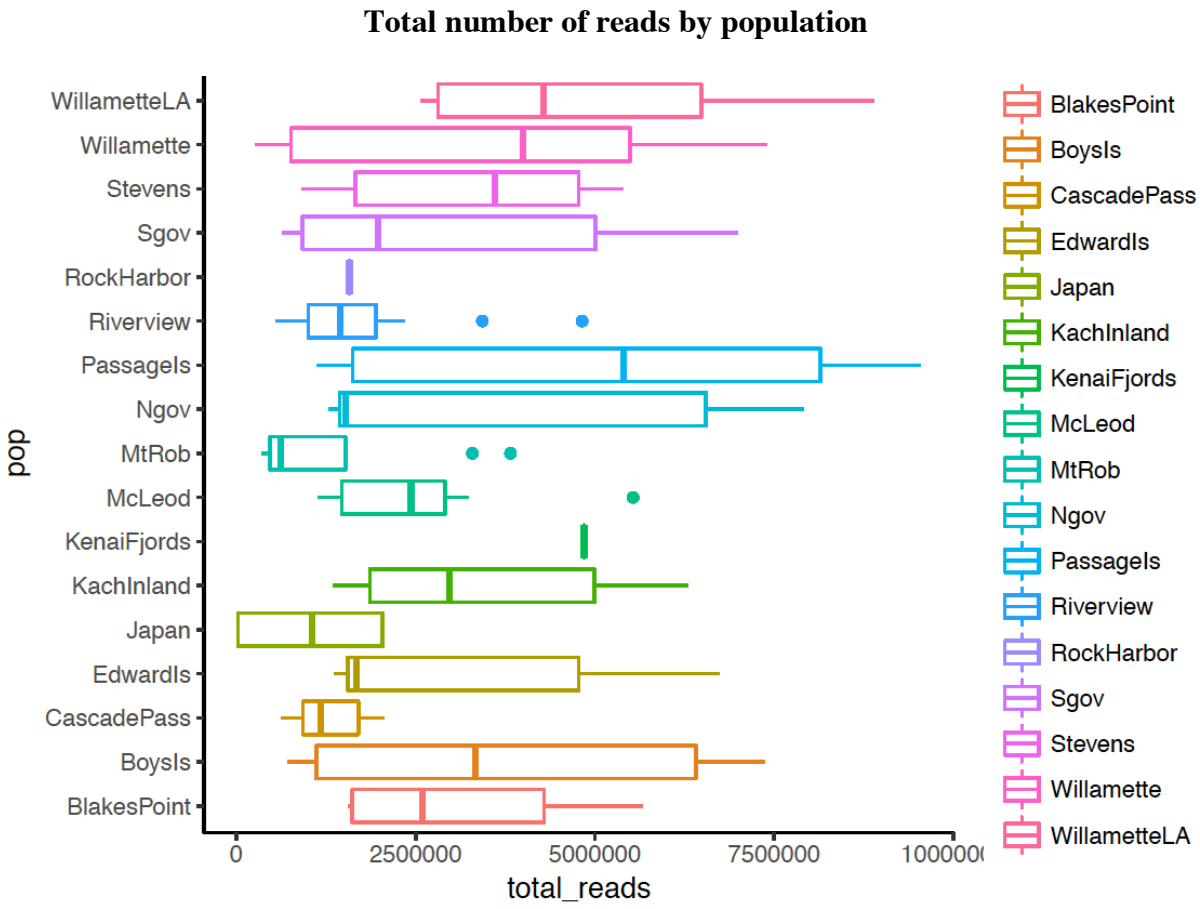
Wen, J. & Zimmer, E.A., 1996. Phylogeny and Biogeography of Panax L . (the Ginseng Genus, Araliaceae): Inferences from ITS Sequences of Nuclear Ribosomal DNA. 6(2), pp.167–177.

Wen, J. et al., 2001. The Evolution of Araliaceae: A Phylogenetic Analysis Based on ITS Sequences of Nuclear Ribosomal DNA. 26(1), pp.144–167.

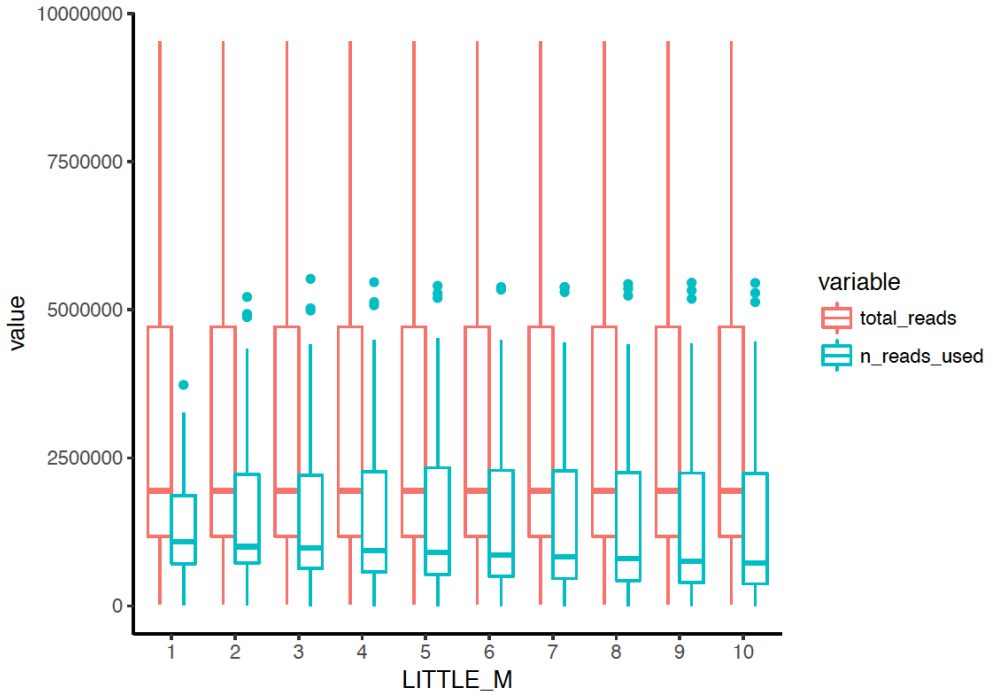
Wen, J. et al., 2010. Timing and modes of evolution of eastern Asian-North American biogeographic disjunctions in seed plants. *Darwin's Heritage Today: Proceedings of the Darwin 2010 Beijing International Conference*, pp.252-269. Higher Education Press.

Yokoyama, Y., Lambeck, K. & De Deckker, P., 2000. Timing of the Last Glacial Maximum from observed sea-level minima. 406(August), pp.713–717.

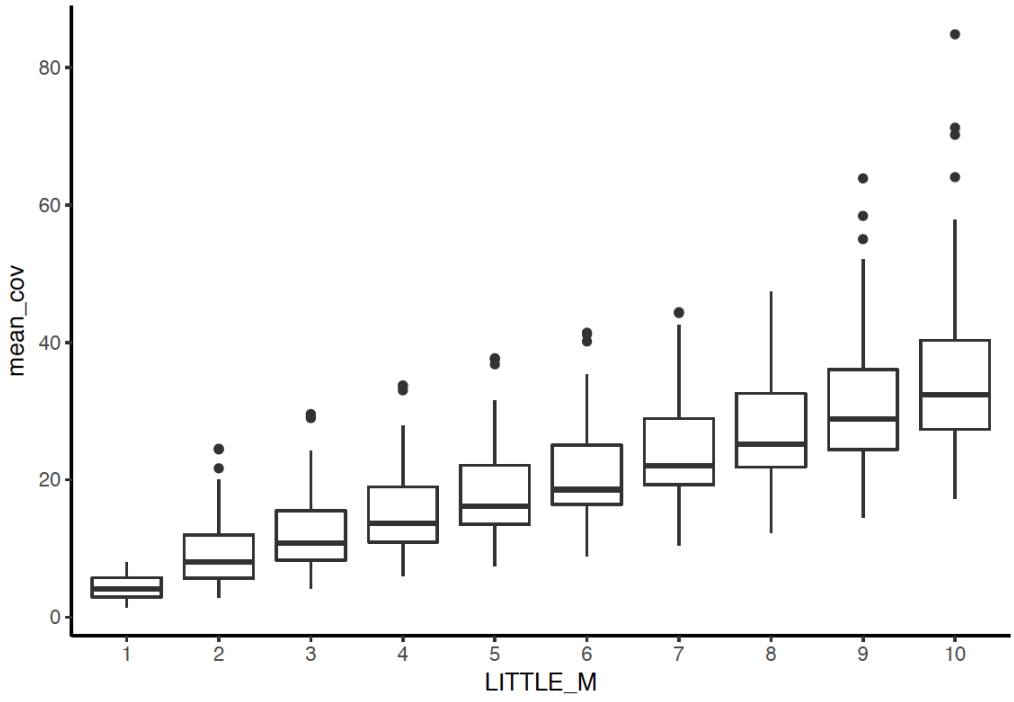
Appendix 4: Results from Stacks assembly optimization.



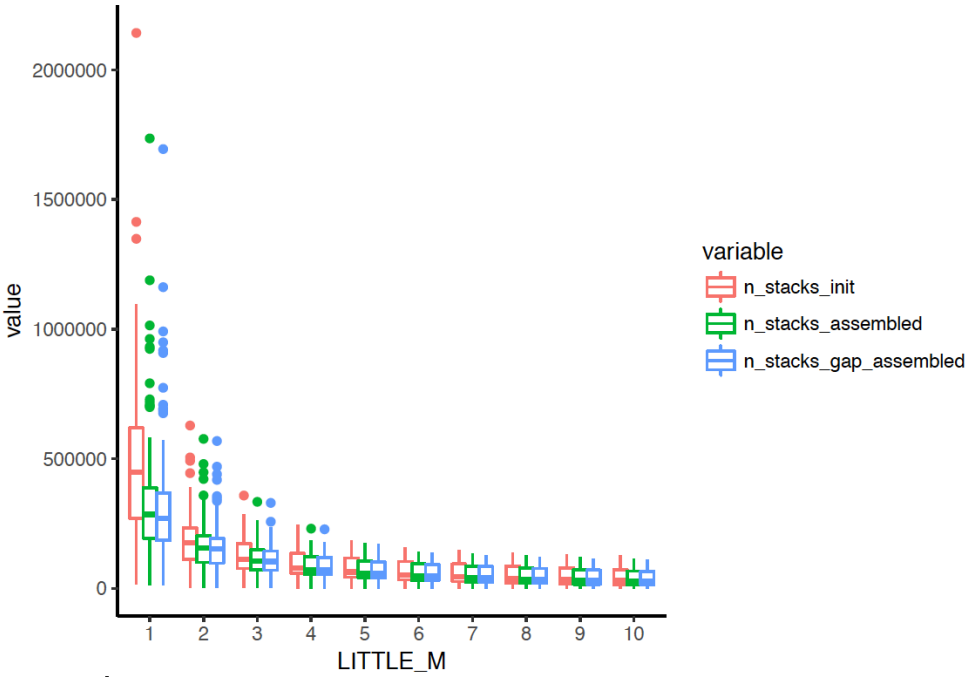
Number of reads used based on 'm' assembly parameter



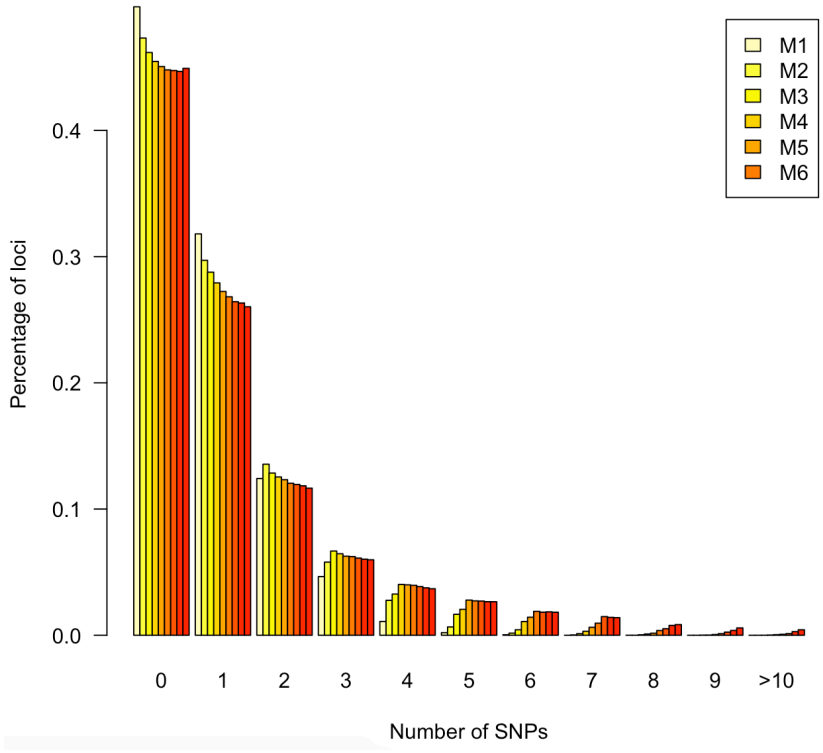
Mean coverage per locus based on 'm' assembly parameter



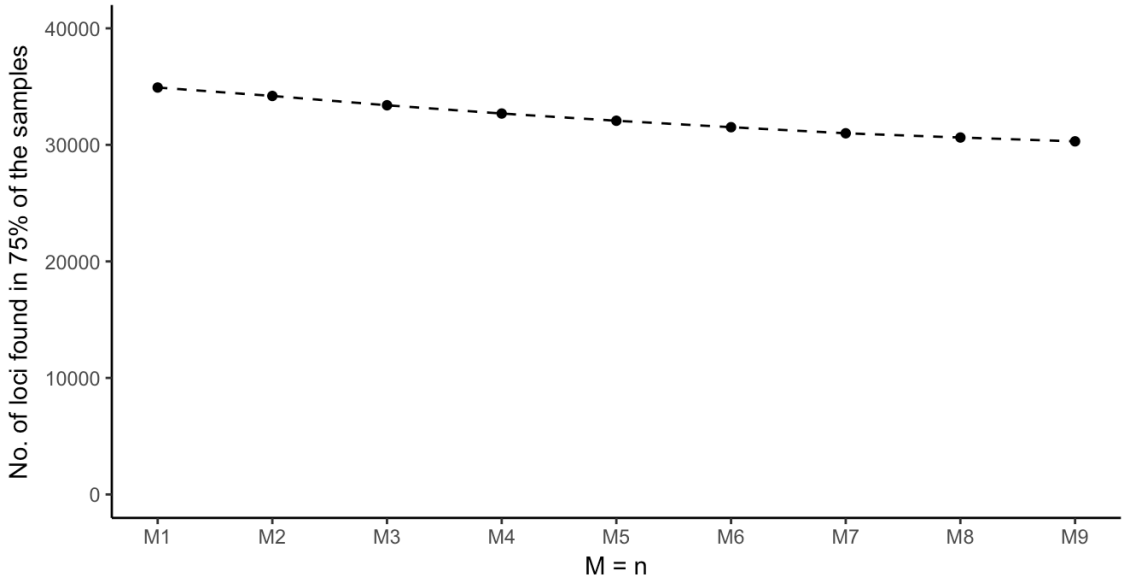
Number of loci (stacks) formed based on 'm' assembly parameter



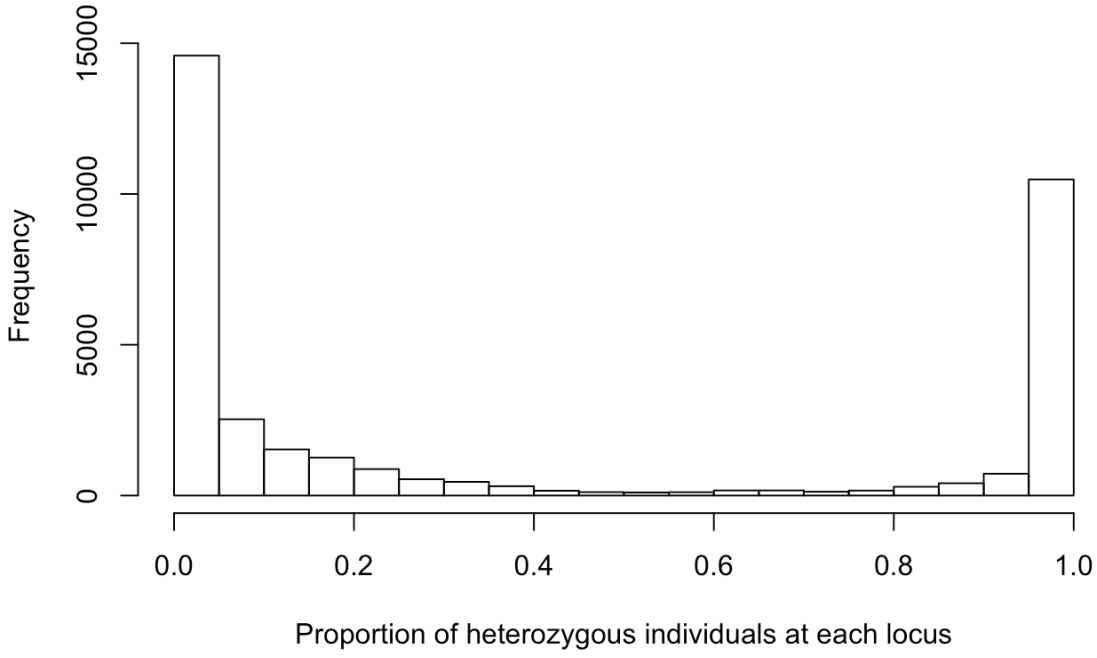
Proportion of loci with n-SNPs based on "M" assembly parameter



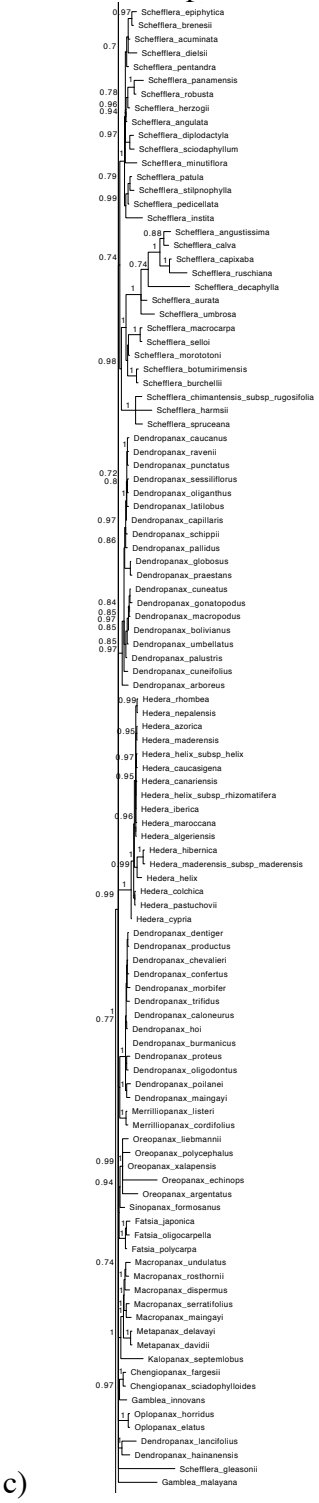
Number of loci in 75% of samples for each value of M == n



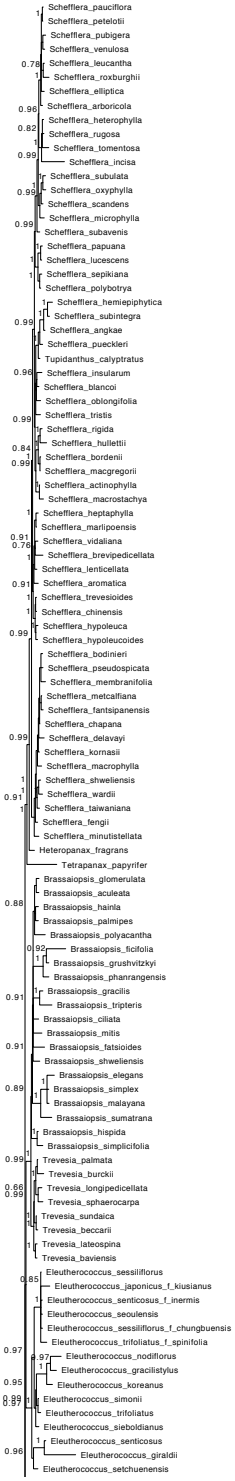
Proportion of heterozygotes at each locus (SNP) for M = n = 4



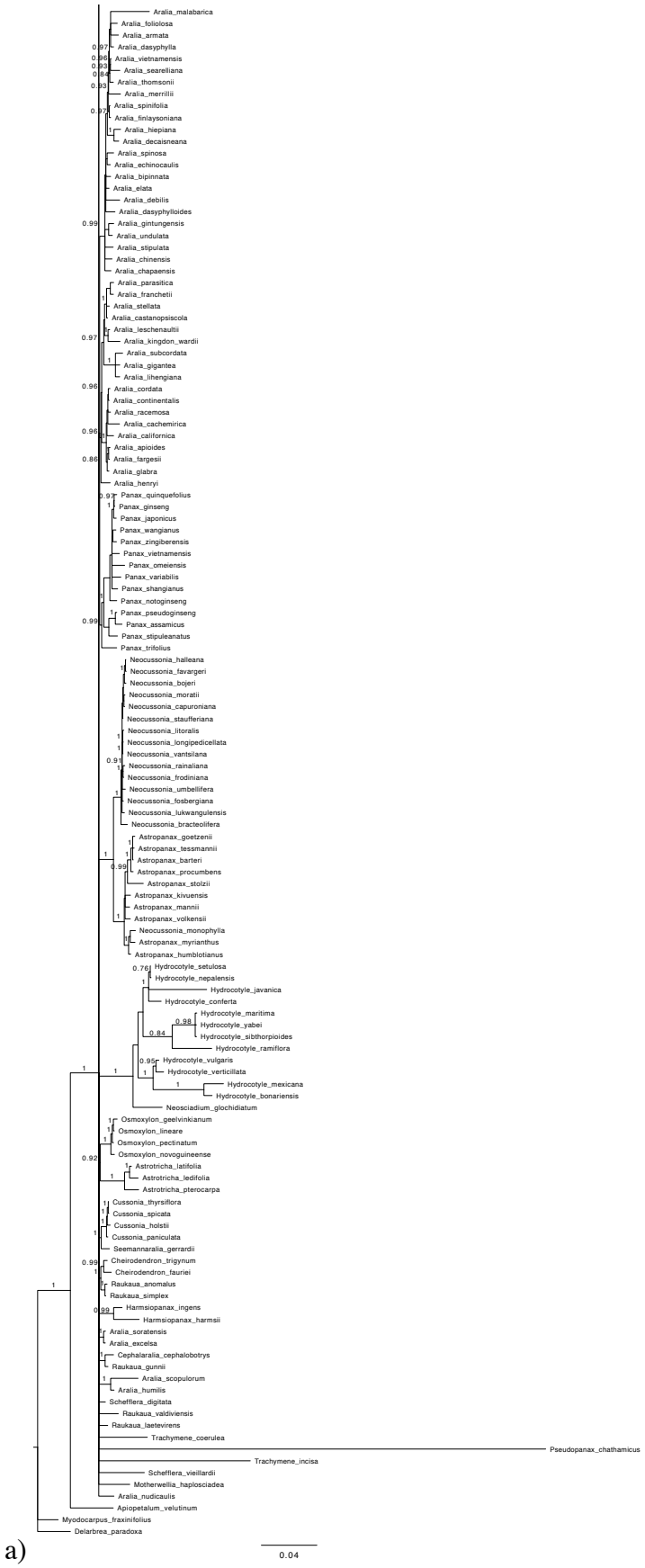
Appendix 5 MrBayes consensus tree based on nrITS + 12 cpDNA marker data set.100 million MCMC generations with 25% burn-in. Posterior probabilities >75% were retained and listed above the branches. Branches along the tree are very short, so support values were often listed to the left side of the backbone, but are in-line with the corresponding branch. Tree is split into four sections to help with visibility (a-d), from the earliest divergence to the most recent divergence.



c)



d)


Aralia_malabarica
Aralia_foliolosa
Aralia_armata
Aralia_dasyphylla
Aralia_vietnamensis
Aralia_searelliana
Aralia_thomsonii
Aralia_merrillii
Aralia_spinifolia
Aralia_finlaysoniana
Aralia_hiepiana
Aralia_decaisneana
Aralia_spinosa
Aralia_echinocaulis
Aralia_bipinnata
Aralia_elata
Aralia_debilis
Aralia_dasyphylloides
Aralia_gintungensis
Aralia_undulata
Aralia_stipulata
Aralia_chinensis
Aralia_chapaensis
Aralia_parasitica
Aralia_franchetii
Aralia_stellata
Aralia_castanopsiscola
Aralia_leschenaultii
Aralia_kingdon_wardii
Aralia_subcordata
Aralia_gigantea
Aralia_lihengiana
Aralia_cordata
Aralia_continentalis
Aralia_racemosa
Aralia_cachemirica
Aralia_californica
Aralia_apioides
Aralia_fargesii
Aralia_glabra
Aralia_henryi
Panax_quinquefolius
Panax_ginseng
Panax_japonicus
Panax_wangianus
Panax_zingiberensis
Panax_vietnamensis
Panax_omeiensis
Panax_variabilis
Panax_shangianus
Panax_notoginseng
Panax_pseudoginseng
Panax_assamicus
Panax_stipuleanatus
Panax_trifolius
Neocussonia_halleana
Neocussonia_favargeri
Neocussonia_bojeri
Neocussonia_moratii
Neocussonia_capuroniana
Neocussonia_staufferiana
Neocussonia_litoralis
Neocussonia_longipedicellata
Neocussonia_vantsilana
Neocussonia_rainaliana
Neocussonia_frodiniana
Neocussonia_umbellifera
Neocussonia_fosbergiana
Neocussonia_lukwangulensis
Neocussonia_bracteolifera
Astropanax_goetzenii
Astropanax_tessmannii
Astropanax_barteri
Astropanax_procumbens
Astropanax_stolzii
Astropanax_kivuensis
Astropanax_mannii
Astropanax_volkensii
Neocussonia_monophylla
Astropanax_myrianthus
Astropanax_humblotianus
Hydrocotyle_setulosa
Hydrocotyle_nepalensis
Hydrocotyle_javanica
Hydrocotyle_conferta
Hydrocotyle_maritima
Hydrocotyle_yabei
Hydrocotyle_sibthorpioides
Hydrocotyle_ramiflora
Hydrocotyle_vulgaris
Hydrocotyle_verticillata
Hydrocotyle_mexicana
Hydrocotyle_bonariensis
Neosciadium_glochidiatum
Osmoxylon_geelvinkianum
Osmoxylon_lineare
Osmoxylon_pectinatum
Osmoxylon_novoguineense
Astrotricha_latifolia
Astrotricha_ledifolia
Astrotricha_pterocarpa
Cussonia_thyrsiflora
Cussonia_spicata
Cussonia_holstii
Cussonia_paniculata
Seemannaralia_gerrardii
Cheirodendron_trigynum
Cheirodendron_fauriei
Raukaua_anomalus
Raukaua_simplex
Harmsiopanax_ingens
Harmsiopanax_harmsii
Aralia_soratensis
Aralia_excelsa
Cephalaralia_cephalobotrys
Raukaua_gunnii
Aralia_scopulorum
Aralia_humilis
Schefflera_digitata
Raukaua_valdiviensis
Raukaua_laetevirens
Trachymene_coerulea
Pseudopanax_chathamicus
Trachymene_incisa
Schefflera_vieillardii
Motherwellia_haplosciadea
Aralia_nudicaulis
Apiopetalum_velutinum
Myodocarpus_fraxinifolius
Delarbrea_paradoxa
0.04
a)
Polyscias_anjozorabensis
Polyscias_madagascariensis
Polyscias_lucienneae
Polyscias_aubrevillei
Polyscias_myrsine
Polyscias_baretiana
Polyscias_muraltiana
Polyscias_gentryi
Polyscias_orientalis
Polyscias_andapensis
Polyscias_andrearum
Polyscias_multibracteata
Polyscias_modestei
Polyscias_felicis
Polyscias_ornifolia
Polyscias_stuhlmannii
Polyscias_amplifolia
Polyscias_fraxinifolia
Polyscias_chapelieri
Polyscias_zanthoxyloides
Polyscias_carolorum
Polyscias_tafondroensis
Polyscias_ariadnes
Polyscias_pentamera
Polyscias_maralia
Polyscias_farinosa
Polyscias_albersiana
Polyscias_fulva
Polyscias_boivinii
Polyscias_floccosa
Polyscias_mayottensis
Polyscias_repanda
Polyscias_coriacea
Polyscias_sessiliflora
Polyscias_dichroostachya
Polyscias_mauritiana
Polyscias_crenata
Polyscias_nothisii
Polyscias_bracteata
Polyscias_dioica
Polyscias_ouaimensis
Polyscias_cissodendron
Polyscias_scopoliae
Polyscias_veillonii
Polyscias_sambucifolia
Polyscias_calophylla
Polyscias_dzumacensis
Polyscias_pancheri
Polyscias_nitida
Polyscias_balansae
Polyscias_jaffrei
Polyscias_lecardii
Polyscias_regalis
Polyscias_joskei
Polyscias_microbotrys
Polyscias_elegans
Polyscias_nodosa
Polyscias_macrocarpa
Polyscias_ledermannii
Polyscias_macrantha
Polyscias_schultzei
Polyscias_bellendenkerensis
Polyscias_willmottii
Polyscias_mollis
Polyscias_australiana
Polyscias_cumingiana
Polyscias_fruticosa
Polyscias_guilfoylei
Polyscias_scutellaria
Polyscias_multijuga
Polyscias_javanica
Polyscias_macgillivrayi
Polyscias_samoensis
Polyscias_murrayi
Polyscias_purpurea
Plerandra_grayi
Plerandra_vitiensis
Plerandra_grandiflora
Plerandra_bakeriana
Plerandra_victoriae
Plerandra_pickeringii
Plerandra_insolita
Plerandra_brassii
Plerandra_solomonensis
Plerandra_stahliana
Plerandra_micrantha
Meryta_senfftiana
Meryta_lanceolata
Meryta_denhamii
Meryta_balansae
Meryta_oxylaena
Meryta_coriacea
Meryta_sinclairii
Meryta_tenuifolia
Pseudopanax_lessonii
Pseudopanax_discolor
Pseudopanax_gilliesii
Pseudopanax_crassifolius
Pseudopanax_ferox
Pseudopanax_linearis
Pseudopanax_laetus
Neopanax_arboreus
Pseudopanax_colensoi
b)

Appendix 6: H' consistency values from CLUMPP, comparing population assignments across five iterations of each K value. We tested K = 2-10. Higher H' values indicate higher consistency. Only K2 was consistent, using a 0.8 cut-off.

	H'	**algorithm**
K2	0.999	Greedy
K3	0.796	Greedy
K4	0.616	Greedy
K5	0.64	Greedy
K6	0.688	Greedy
K7	0.563	Greedy
K8	0.601	LargeKGreedy
K9	0.424	LargeKGreedy
K10	0.59	LargeKGreedy

CHAPTER 4: Phylogeographic analysis of *Aconitum columbianum* (Ranunculaceae) reveals parallel dispersals to eastern and western North America from a central ancestral range: a contrast to additional Western North America-Great Lakes Region disjuncts

Co-authors: Margaret A. Kuchenreuther, Christopher Setzke

ABSTRACT

Aconitum columbianum Nutt. (Columbian monkshood, Ranunculaceae, subg. *Aconitum*) is an arctic alpine species native to North America. It is distributed broadly west of the Rocky Mountains with disjunct populations in Ohio, New York, the Black Hills of South Dakota, and the algific talus slopes in the Driftless Area of the Midwest. The populations in the Midwest once held their own species rank, *A. noveboracense* A. Gray ex Coville, and its restricted range made it a federally threatened species. Attempts have been made to distinguish morphological types, but morphological variation in this species is not geographically structured, and even individuals within populations exhibit confounding variation. Here we use genotyping-by-sequencing to reconstruct the population history of *A. columbianum* to determine whether there is a clear dispersal pattern that can explain its disjunction. We test its population structure to detect admixture among these morphologically intergrading populations. Finally, we compare the genetic diversity between the Great Lakes Region (GLR) and western populations to determine whether there is evidence of historical bottlenecks that might increase the vulnerability of these threatened populations in the future. We show that there are parallel eastward and westward dispersals from an ancestral range in the southern Rocky Mountains. Population structure is admixed between distant populations, perhaps contributing to the morphological complexity of this group and indicating high rates of dispersal. Finally, the populations in the GLR and NY have high levels of genetic diversity, similar to results found in previous studies, indicating the

absence of founder events to these areas and suggesting that its adaptive potential may not be the limiting factor that determines its persistence in this region.

INTRODUCTION

There are over 100 vascular plants that are disjunct between western North America and the Great Lakes Region (GLR) (See Chapter 1). Taxa of different ranks exhibit this distribution, indicating either variation in length of divergence time since disjunction or inconsistency in nomenclatural assignment between plant groups, perhaps driven by blurry species boundaries. One exemplar of this nomenclatural uncertainty is *Aconitum columbianum* Nutt. (Columbian monkshood, Ranunculaceae, subg. *Aconitum*). It is found in western North America from southern British Columbia through Arizona in moist to wet, subalpine to alpine meadows. In western North America it is often found in openings of coniferous forest and in riparian areas, in sandy, loamy, and clay soils. In the Driftless Area of the Midwest it is restricted to algific talus slopes, and in Ohio and New York it is found in microhabitats that mimic higher elevation (Kuchenreuther 1996). The eastern North American populations once held their own species rank, *A. noveboracense* A. Gray ex Coville, but were subsumed into *A. columbianum* due to morphological and genetic similarity (Kuchenreuther 1991, Brink and Woods 1997, Cole and Kuchenreuther 2001). In addition, the spectrum and complexity of floral and vegetative variation among populations of *A. columbianum* has led to its classification as one intergrading species with only two subspecies (*A. columbianum* ssp. *columbianum*, *A. columbianum* ssp. *viviparum* (Greene) D.E. Brink) defined by the absence or presence, respectively, of reproductive bulbils in the leaf axils (Brink 1980, Brink and deWet 1980). The recognized bulbil-bearing subspecies is geographically distinct in western North America. However, collections from the GLR show that

this trait varies even within populations, making the presence or absence of bulbils a poor taxonomic character (Kuchenreuther 1991). As a result, here we treat all *A. columbianum* populations as one taxonomic entity.

Accompanying the complex morphological variation within this species, there is a striking disjunct distribution that is not mirrored as strongly by morphology as in other species sharing this western North America-GLR disjunct distribution, such as *Rubus parviflorus* (See Chapter 2). The disjunct distribution of *A. columbianum* is also not easily explained by long-distance dispersal, since the species produces follicle fruits and its seeds fall close to the parent. Local groups of populations tend to be distinct and widely separated, and regional distributions of populations are isolated on montane high elevation "islands." The greater disjunction between western and eastern regions of North America is therefore not wholly uncharacteristic of the species (Brink 1982), but the long-distance disjunction is unusual and raises important questions. The disjunct populations in eastern North America have been hypothesized relicts of a historically contiguous spread of isolated populations along the cool, moist glacial margin, with the eastern populations now restricted to glaciated regions with the only remaining suitable, cool habitat (Iltis 1965, Brink 1982). Here we test whether these eastern North American populations arose via a stepping-stone model of dispersal from the West and test the intergradation between populations indicative of a historically widespread distribution. We tease apart the evolutionary history and geographic dispersal history of *A. columbianum* using phylogenetics, and we test whether there is significant admixture between these supposed intergrading populations by looking at population structure analyses.

The geographic distributions and relationships among species of *Aconitum* in North America are also key components in understanding the history of the *A. columbianum* complex.

Aconitum, with about 250 species in subg. *Aconitum* and 50 species in subg. and 50 species in subg. *Lycoctonum*, is circumboreal and most concentrated in arctic alpine habitats in Asia, but it extends south along the mountain ranges in North America where there are mesic, high elevation conditions (Brink 1982). The four additional *Aconitum* species in North America are *A. delphiniifolium*, *A. maximum*, *A. reclinatum*, and *A. uncinatum*, and only *A. reclinatum* belongs to subg. *Lycoctonum* based on its elongate roots. This list does not include the cultivated aconites that are introduced in North America. *Aconitum columbianum* was shown to have dispersed from Asia no earlier than 3.3 Ma, pre-Pleistocene (Jabbour and Renner 2012), although this analysis was based on an incompletely sampled taxon dataset (*A. columbianum* was the only North American species sampled). If Pliocene in origin, the likely route across the Bering Strait was likely submerged (Brigham-Grette 2001). Both *A. delphiniifolium* and *A. maximum* are found in northeastern Asia and extend into Alaska. *Aconitum delphiniifolium* additionally extends into British Columbia where it is allopatric with *A. columbianum* and possibly intergrades with *A. columbianum* (Brink 1982). *Aconitum reclinatum* and *A. uncinatum*, however, are restricted to the unglaciated regions of eastern North America, and *A. uncinatum* is hypothesized to have historically hybridized with *A. columbianum* in the West Virginia region of the Appalachian Mountains (Hardin 1964). To date, phylogenies of *Aconitum* have shown that *A. columbianum* and *A. delphiniifolium* are sister taxa (Luo et al. 2005, Jabbour and Renner 2012), but none have included other North American *Aconitum* species. Here we add *A. reclinatum* into our phylogenetic data set to test its placement in subg. *Lycoctonum*, and estimate the age of *A. columbianum* divergence with *A. delphiniifolium* to test its pre-Pleistocene origin.

In addition to the history of *A. columbianum*, we also look to the future regarding the eastern GLR populations. The populations in eastern North America have a highly restricted

range and were given federally threatened status in 1978. This status was the result of their unique habitat, the rarity of that habitat in this region, and the generally isolated geographic structure of *A. columbianum* (Federal review, 2008). This status continues today under the previous species name *A. noveboracense* (Federal review, 2008). Genetic diversity was tested for the eastern North American populations based on one nuclear ribosomal marker (Kuchenreuther 1991). We expand this analysis using genomic data to look at genetic diversity and the inbreeding coefficients of populations in eastern and western North America to assess the adaptive potential and sustainability of the eastern North American populations in the face of rapid climate change.

METHODS

Sampling: We collected leaf samples from 12 populations across the North American range of *Aconitum columbianum*. Sampling from the Rocky Mountains, the Great Lakes Region (GLR), and New York was completed between 1989-1991, and leaf samples were collected from 20-30 individuals per population from stems well separated from each other along a linear transect. These leaf samples were bagged separately and frozen at -80°C. Sampling from western North America was completed in five field seasons between 2013-2017, and leaf samples were collected from 15-20 individuals per population at 5-meter intervals along a linear transect. These leaf samples were individually dried in silica gel. In both of these sampling schemes, individual plant collections were spaced at 5-meter intervals in order to avoid sampling potential clones, or genets, that are produced via rhizomes. One herbarium voucher specimen was collected for each population in 2013-2017, and they currently reside in the Wisconsin State

Herbarium. We collected leaf tissue from a population of *A. delphiniifolium* in Alaska to serve as the outgroup.

DNA Extraction: All silica-dried samples were disrupted with a TissueLyser (Qiagen, Valencia, CA, USA) using tungsten-carbide beads, and all fresh-frozen samples were disrupted using a mortar and pestle with liquid nitrogen. DNA was extracted using the Qiagen DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA) following the manufacturer's protocol. DNA integrity was confirmed by presence of high-mass bands on a 0.8% agarose gel, and DNA quantification was conducted with the Qubit dsDNA High-Sensitivity Assay Kit (Invitrogen, USA).

Aconitum *chronogram*:

Multi-gene alignments: We constructed a dated phylogeny of *Aconitum*, which is embedded within *Delphinium* (Jabbour and Renner 2012), to estimate the stem age of *A. columbianum*. We took a super-matrix approach and used PyPHLAWD (Smith and Brown 2018) to extract all available sequences from the Plant and Fungal NCBI database and build clusters of loci across *Aconitum*, retaining one accession per species. Our clustering specifications were kept at the default settings, including sequence length >= 600, clustering length limit = 0.65, and percent identity = 20. We retained clusters of nuclear ribosomal (nrDNA) (18S + ITS1 + 5.8S + ITS2 + 28S; ETS), and seven chloroplast DNA (cpDNA) markers (*psbA+psbA-trnH+trnH*, *ndhF-trnL*, *psbD-trnT*, *matK*, *trnL*, *trnT-trnL*, and *rbcL*) that had at least 30 species. Cluster FASTA files were aligned using MUSCLE in Geneious v.10.1 (http://www.geneious.com, Kearse et al. 2012). We conducted our own Sanger sequencing of *trnS-trnG* at the UW-Madison Biotechnology Center on an ABI 3730x1 DNA Analyzer for populations across the geographic range of *A.*

columbianum, so we retained the *trnS-trnG* PyPHLAWD cluster even though it had fewer taxa. We built assemblies, edited sequences, and re-aligned our sequences to the PyPHLAWD *trnS-trnG* alignment in Geneious. We added five outgroups from within Delphinieae for which sequences were available: *Consolida ajacis*, *Delphinium venulosum*, *Delphinium virgatum*, *Gymnaconitum gymnandrum*, and *Staphisagria requienii*. *Consolida ajacis, D. virgayum*, and *D. venulosum* are part of *Delphinium* sensu Jabbour and Renner (2012) and they form a monophyletic clade with *Gymnaconitum*, sister to *Aconitum* (Wang et al. 2009, Wang et al. 2016). Species were removed from the concatenated cpDNA alignment if they had > 75% missing data compared to the ungapped alignment consensus sequence. ETS, ITS, and concatenated cpDNA markers were combined for downstream partitioned phylogenetic analysis.

Chronogram calibration: Divergence between *A. columbianum* and a closely related Asian aconite, *A. yamazakii*, has been dated to ~ 3.3 Ma (Jabbour and Renner, 2012). However, this chronogram was based on a secondary crown age of Ranunculaceae (73 +/- 3 Ma, Britton et al. 2007) and they calibrated internal nodes based on secondary dates from Bell et al. (2010), which contains problematic divergence estimates due to their placement of fossil priors. Here we construct a chronogram of *Aconitum* using a secondary age estimate at the stem, based on the most recent Ranunculaceae-wide chronogram (Wang et al. 2016).

Recent diversification analyses of lineages within Ranunculaceae have used the Anderson et al. (2005) estimated date of 87-73 Ma based on *rbcL* as a secondary calibration for the family (e.g. Emadzade and Horandl 2011, Fior et al. 2013). However, the most recent chronogram of Ranunculaceae provides more robust internal age estimates and used five cpDNA markers and the 28S nuclear ribosomal marker with three fossil constraints and a 125 Ma constraint on the root (Wang et al. 2016). This chronogram dated Ranunculaceae to 108.71 Ma (101.57-114.57

95% HPD), which is older than the ages reported in Anderson et al. (2005) and Bell et al. (2010). In addition, it estimated a stem age of *Aconitum* diverging with *Gymnaconitum* + *Delphninium* (*sensu* Jabbour and Renner 2012) at 26.02 Ma (18.45-34.31 95% HPD), which falls within the age estimate between *Aconitum* and *Gymnaconitum* based on a recent nrITS chronogram (Wang et al. 2016). We use 26.02 Ma as a secondary stem calibration for the MRCA of *Aconitum* + *Gymnaconitum* + *Delphinium* (*sensu* Jabbour and Renner 2012) in our *Aconitum* chronogram.

Phylogenies: We ran the concatenated cpDNA, nrITS, and nrETS data sets separately and combined with partitions using RAxML v.8.2.10 (Stamatakis 2014) and MrBayes v.3.2.6 (Ronquist et al. 2012) on the CIPRES platform (Miller et al. 2010). In RAxML we used the GTR model of nucleotide substitution under a Γ model of rate heterogeneity with alpha estimated, and we used the rapid hill-climbing algorithm. In MrBayes, we were specified one outgroup, *Staphisagria requienii*, and ran a GTR model with 10 million generations and 25% burn-in. Our sanger sequence data placed our populations in a distant clade compared with previously published analyses, so we removed our sequences and re-ran all analyses without *trnS-trnG*. We ran the combined data set with partitions in BEAST v2.4.6 (Bouckaert et al. 2014) on the CIPRES platform. BEAST was run with GTR + Γ, a relaxed lognormal clock model, and a yule birth model. Tree likelihood calculation was set to treat ambiguities as equally likely. We constrained *Aconitum* to be monophyletic and placed the secondary stem age with a normal distribution at the MRCA of *Aconitum*, *Consolida ajacis*, and the two *Delphinium* species. We chose this MRCA node because these are the outgroups that form a monophyletic clade with *Aconitum*. We ran 10 million generations, evaluated the posterior distributions of the priors in Tracer v.1.6.0 (Rambaut et al. 2018), and calculated the maximum clade credibility tree using TreeAnnotator v.2.4.6 (Bouckaert et al. 2014) with 25% burn-in.

Genotyping-by-sequencing: We selected 4-10 individuals per population for genotyping-by-sequencing (GBS) and included four samples of *A. delphiniifolium*. Sample library prep and sequencing was conducted at the UW-Madison Biotechnology Center using the *ApeKI* restriction digest enzyme and pooling samples 50-plex on 2 lanes of the Illumina Hi-Seq 2500 platform.

Bioinformatics: We ran cutadapt v. 1.13 (Martin 2011) to trim off common adapters at the 3' end and retain only full-length reads. We ran ipyrad v. 0.7.13 (Eaton 2014) steps 1 and 2 to demultiplex samples allowing 0 mismatches in barcodes, remove barcodes, then filter out reads with any low-quality bases (phred Q score < 20). We ran cutadapt again to trim off the 4-base pair overhang at the 5' end, resulting in a total length of 89bp. We removed populations with very low mean read coverage, and 12 populations plus the outgroup remained in the final dataset (Table 1). We assembled reads by running ipyrad and Stacks (Catchen et al. 2013) on the Condor computing cluster at the Center for High Throughput Computing facility at UW-Madison. We ran ipyrad in a Docker container (Merkel 2014) to ensure access to required dependencies.

Parameter testing: We tested three influential clustering parameters as described in Paris et al. (2017) and Rochette and Catchen (2017) (Stacks), and in Anderson et al. (2017) (PyRAD): minimum depth to make a statistical base call (m), maximum number of mismatches allowed when forming a locus within an individual (M), and maximum number of mismatches allowed when aligning loci between individuals (n). We tested values 1-10 for the minimum depth to make a statistical base call. Results from this revealed that our data were at lower than 20X coverage at the default value (m = 2) (Appendix 7) and as expected, coverage was higher when requiring a higher minimum depth to retain a locus. Based on suggestions from the Stacks

developer (pers. comm, Stacks Googlegroup), data with lower coverage should not be forced to a higher depth, and m should be kept at 2 or 3 to allow for statistical calls. We used m = 3, and tested M = n for values 2-6. We chose M = n = 4. The m, M, and n parameter values were set to maximize the number of loci formed, stabilize the proportion of loci recovered for n-SNPs/locus, and minimize the proportion of heterozygotes at each base pair (Appendix 7).

Phylogeography: To estimate ancestral migration history of *A. columbianum* using population divergence, we estimated phylogenies using concatenated SNP data with Maximum Likelihood (ML) in RAxML and a multi-species coalescent approach in SVDQuartets v.1.0 (Chifman and Kubatko 2014) in PAUP v. 4.0a150 (Swofford 2002).

To eliminate the issue of allele phasing across concatenated loci, we used ambiguity codes for each individual to represent biallelic data (INDIV), and we took a majority-rule consensus approach (CONS), which includes heterozygous sites, but calls only one allele at each site. This majority-rule approach is useful because it reduces ambiguity.

RAxML: To reduce the likelihood of overestimating branch lengths (Leaché et al. 2015), we implemented the Lewis ascertainment bias correction in RAxML, which accounts for unobserved invariant sites. For the INDIV data set, we ran the populations module in Stacks on all samples to identify loci present in at least 75% of individuals, concatenating only one SNP per locus. We removed any samples that had >50% missing data and any samples that were the only representative from a population. Because the populations module includes Ns as variant sites, we changed Ns to gaps and masked invariant sites in Geneious. In addition, the ascertainment bias correction interprets ambiguity codes as either of the base calls, and some of the ambiguities in our data set resulted in monomorphic loci. To circumvent this issue, we

masked SNPs that had ambiguity codes. We changed the gaps back to Ns and ran RAxML on the CIPRES platform with the Lewis ascertainment bias correction, the GTR model of nucleotide substitution under a Γ model of rate heterogeneity with alpha estimated, and the rapid hill-climbing algorithm. We conducted 100 GTRCAT rapid bootstrap replicates. For the CONS data set, we modified the INDIV data set and extracted the majority rule consensus sequence for each population. We combined the sequences without realigning since they were all of uniform length. We replaced Ns with gaps to mask invariant sites, and because some ambiguous sites were being called as invariant, we also masked ambiguous sites. We replaced gaps with Ns and ran RAxML with the same settings.

SVDQuartets: To account for potential incomplete lineage sorting (ILS), we used SVDQuartets in Paup to estimate the history of each locus individually and obtain a multi-species coalescent tree. Due to the proposed intergradation between *A. delphiniifolium* and *A. columbianum* in the West, we ran populations without the outgroup to identify loci present in at least 75% of individuals, retaining one SNP per locus. We ran SVDQuartets with multiple samples assigned to a locality population. We used the multi-species coalescent tree model to evaluate all possible quartets and used the QFM quartet assembly to summarize quartets (Reaz et al. 2014). We distributed heterozygote ambiguity instead of masking it and we ran 100 bootstrap replicates on subsets of 100,000 quartets. We rooted this tree based on the RAxML topology.

Population structure and genetic diversity: To estimate population structure and identify potential admixture between populations, we used Structure (Pritchard et al. 2000). This program uses a Bayesian model to leverage the assumptions of Hardy-Weinberg F_{ST} to construct probabilities of population assignment for each individual given a value of K (number of

populations). We ran the populations module in Stacks without the outgroup and retained only loci that were present in 100% of the populations and in at least 75% of individuals within each population. We retained one SNP per locus to maximize independence between SNP loci. We ran Structure using the admixture model, estimating the distribution of allele frequencies (Lambda), and allowing alpha to vary. All other settings were kept at their default values. We tested K values 2-10 with 10^4 generations and ran five iterations of each K value. We ran Structure Harvester (Earl and vonHoldt 2012) to select the best K value by evaluating Delta K (Evanno et al. 2005). We tested consistency between the five runs of the best K value using CLUMPP v 1.1.2 (Jackobsson and Rosenberg 2007). In CLUMPP we assessed the H' value using the Greedy algorithm with 1000 random input orders for K 2-7 and the LargeKGreedy algorithm for K 8-10. We plotted the mean population assignment values reported from CLUMPP using distruct1.1 (Rosenberg 2004).

Based on previous research on nuclear ribosomal DNA variation (Kuchenreuther 1991), the populations within the Driftless Area of Iowa and Wisconsin exhibit high amounts of genetic diversity and variation among populations. This study noted a lack of correlation between genetic diversity and geography or population size and attributed this to the random process of founder events and genetic drift. Here, we look at genetic diversity across the genome and across populations from the whole range of *A. columbianum* but with a larger DNA data set and broader population sampling. We quantified nucleotide diversity (π), expected and observed heterozygosity (H_e and H_o), the inbreeding coefficient (F_{IS}), the number of private alleles, and the percent of polymorphic loci for each population. We were particularly interested in whether the populations in the Driftless Area of Wisconsin and Iowa, which are restricted to algific talus slopes, are lower in genetic diversity compared to their counterparts in other regions of North

America. We ran the populations module in Stacks, grouping samples by their collection locality and retaining only loci that were present in 100% of the populations and at least 75% of individuals within a population. We ran diversity statistics in the populations module using kernel-smoothing and 100 bootstrap replicates. The F_{IS} statistic was calculated with the Bonferroni correction using a sliding window with the base p-value of 0.05. We performed unpaired, two-sample t-tests to compare mean values between populations in the GLR and populations in western North America and in New York.

RESULTS

Aconitum *chronogram*: Our RAxML trees for nrITS, nrETS, and the concatenated cpDNA data sets have strong topological congruence of major clades (not shown). The RAxML tree from the combined data set has a similar topology as the trees from separate data sets and shows high support for major clades (Fig. 1). North American *A. reclinatum* is placed into the monophyletic subg. *Lycoctonum*. The only exception to the monophyly of subg. *Lycoctonum* is *A. moschatum*, which falls sister to both subgenera in the combined analysis. However, inclusion of *A. moschatum* in subg. *Lycoctonum* has previously been questioned based on its lack of a nectary blade with a tubular portion, characteristic of all other species in subg. *Lycoctonum* (Kadota 2001, Hong et al. 2017). Our MrBayes trees from separate data sets also had congruent major clades (not shown), and the combined analysis supported the same major clades as the RAxML analysis with lower support along the backbone (Fig. 2). The BEAST analysis estimated a stem divergence age of 4.1 Ma (1.25-7.0 95% HPD) for *A. columbianum* diverging from *A. delphiniifolium*, and the topology was comparable to the RAxML and MrBayes trees with high support (Fig. 3).

Phylogeography: *RAxML:* After filtering samples with >50% missing data and samples that were the only representative of a population, and masking invariant and ambiguous sites, the INDIV data set contained 1,882 concatenated SNPs. After similar filtering and masking of invariant and ambiguous sites, the CONS data set contained 1,240 concatenated SNPs. Our RAxML INDIV tree resulted complete monophyly and high support for the locality populations (Fig. 4). Our RAxML CONS tree revealed the same topology as the INDIV tree, with very high support for each population and most internal branches (Fig. 5). The topology suggests two parallel disjunct distributions, both likely originating in central North America. One clade resolves the Black Hills and the Wyoming populations (Bighorn Mountains) as sister to a clade of the Driftless Area GLR populations (St. Olaf, IA and Hay Valley, WI), which are sister to the New York population (Peekamoose, NY). This clade had short internal branches. Another clade resolves the Ohio population as sister to a clade of parallel dispersals between the southern Rocky Mountains and the Pacific Northwest. Specifically, Colorado and Arizona populations are more closely related to Washington and Oregon populations than they are to each other. The internal branches for this clade were longer than those of the first clade. There is no distinctive west to east signal in relationships between populations. Instead, the results indicate multiple dispersals east and west from a central ancestral range in North America.

SVDQuartets: The SVDQuartet tree shows a slightly different topology than the RAxML trees (Fig. 6). The two westward and two eastward dispersals found in the RAxML tree are also supported with the SVDQuartet topology. However, a more central ancestral range, perhaps in the Colorado region, is not as clear in the SVDQuartet analysis. There is low support for the Black Hills region as sister to the New York populations. If this weakly supported branch is

collapsed, the clade suggests divergence between the Black Hills region and the Driftless Area + New York populations, which is more biologically intuitive. The divergence between the Black Hills region and the Driftless Area + NY is a separate instance from the dispersal to Ohio. Thus, there is evidence of two separate migrations into eastern North America, which is similar to what was found in the RAxML tree.

Population structure and genetic diversity: The best K based on assessment of Delta K was K = 8 (Fig. 7). The Oregon, Washington, Ohio, and New York pre-defined, locality populations each formed their own population (Fig. 7, in blue, yellow, pink, and purple bars respectively). Two of the Colorado populations grouped as a distinct population (Fig. 7, in dark green), the Arizona population grouped with one of the Colorado populations (Fig. 7. in bright green), the Wyoming and South Dakota populations grouped together (Fig. 7, in red), and the Iowa population grouped together with some of the individuals from the Hay Valley population (Fig. 7, in orange). There were three notable signals of admixture. First, there was admixture between the distinct Colorado populations (Fig. 7, in dark green). There was also admixture between the Wyoming + South Dakota group and the Iowa + Wisconsin group (Fig. 7, in red). Finally, there was admixture between the Iowa + Wisconsin group and the New York population (Fig. 7, in orange). There was also minor admixture between the Colorado group and the Washington population (Fig. 7, in dark green), and between the Arizona + Colorado group and the Oregon population (Fig. 7, in bright green). Finally, there were a few individuals in the Wisconsin locality population that showed affinity with the Arizona + Colorado group (Fig. 7, bright green).

As reported in previous studies of genetic diversity in the GLR populations of *Aconitum columbianum* (Kuchenreuther 1991), we found high levels of genetic diversity in the GLR when

looking across the genome. We found no significant differences in any of the genetic diversity metrics (π, H_e, H_o, and F_{IS}, private alleles, and percent polymorphic loci) between the GLR and the western North American populations (Table 2). This is also true when comparing the New York population with western North American populations. Thus, these results provide evidence that bottleneck events were unlikely, and indicate that the adaptive potential of the GLR and eastern North American populations might be a smaller factor in their vulnerability to future climate change.

DISCUSSION

Parallel dispersals in *A. columbianum*: The combined evidence of our RAxML trees, SVDQuartet tree, and population structure analysis provides strong support for an ancestral range of *A. columbianum* mid-continent, near Colorado. Both the RAxML and SVDQuartet analyses reveal evidence of parallel dispersals to eastern North America, and evidence of parallel dispersals to western North America. The two independent westward dispersals are not as clear in the SVDQuartet analysis. However, this pattern was strongly supported in our structure analysis. The structure analysis shows that the CO Arapaho and CO South populations have slight admixture with WA Haney (Fig. 7, in dark green). It also shows that the OR Willamette population has slight admixture with the AZ and CO Gunnison populations (Fig. 7, in bright green). These signals of admixture suggest two separate instances of western gene flow. The support for two dispersals to eastern North America is clear and well supported in both the RAxML and SVDQuartet analyses. Our structure analysis also reveals this pattern of two eastward dispersals. The Wyoming + South Dakota group is admixed with the Iowa + Wisconsin group (Fig. 7, in red). In addition, there are individuals in the Wisconsin population that are

bright green and share that population grouping with the Arizona + Colorado group (Fig. 7, in bright green). The signal of Wisconsin individuals in bright green might be due to fixation of rare alleles, rather than gene flow.

Our combined results suggest, perhaps counterintuitively, that *A. columbianum* has a high rate of dispersal for its condition as follicle-fruited plant. This high rate of dispersal can help to explain the morphological variation that has been shown to lack strong geographic signal (Brink 1980, Brink and deWet 1980). Brink (1982) argued in favor of this dispersal rate, noting that the isolation and distance between local and regional populations would necessitate a higher rate of dispersal than we might expect. The faster rate of divergence within the GLR and NY populations, indicated by shorter branches, suggests that these populations established when there was more contiguous suitable habitat to colonize, perhaps when the glaciers brought cooler, moist climate to lower latitudes. The slower rate of divergence to the Ohio population might be an artifact of our sampling, or local extirpation that would otherwise break up the long branch. The scenario of a more contiguous distribution along the glacial edge is also supported by the comparable genetic diversity in the GLR and the eastern North American populations compared with the western North American populations. In particular, the Ohio population, which has the largest disjunct range from its ancestral range in the southern Rocky Mountains does not show evidence of a founder event. Thus, it was likely a relict population of a previously contiguous distribution of populations along the glacial margin that had a different dispersal history from the parallel clade of eastward dispersal. These comparable levels of genetic diversity also suggest that the vulnerability of the populations in the GLR is more likely to be determined by their specificity of habitat rather than their adaptive potential.

Western disjunction between the southern Rocky Mountains and the Pacific Northwest has been documented in other species (e.g. Mitton et al. 2000, Jaramillo-Correa et al. 2009, DeChaine et al. 2013) and has recently been redefined as a more intricate system of multiple refugia and sometimes refugia within refugia that needs to be carefully teased apart (Shafer et al. 2010). In the case of *A. columbianum*, one key factor that should be investigated in the future is the timing of these population divergences. In order to estimate divergence dates without requiring broad taxonomic sampling and fossil calibration, we would need an estimate of mutation rate and generation time. Since these data are not well known, efforts would need to be made to conduct broader sampling across the genus in order to use fossil calibration with GBS data (Eaton et al. 2016). Timing would allow us to address the hypotheses we have put forward on the differing rates of dispersal eastward.

Divergence of *A. columbianum* in the context of its North American counterparts: *Aconitum delphiniifolium*, which is restricted to Alaska and eastern Asia, was split into three intergrading subspecies based on morphology (Hultén 1941-1950, vol. 4). The most common form of *A. delphiniifolium* has deeply divided leaves and shorter, crescent-shaped hoods. Since there are some *A. columbianum* populations that also exhibit these features instead of the more typical shallow lobes and taller, conical hoods, it has been suggested that intergradation of these two species in British Columbia should be tested further (Brink 1982). We have already conducted additional sampling of *A. delphiniifolium* populations in Alaska and British Columbia to explore this interplay of morphology and genetic divergence in the future. Our chronogram indicates a pre-Pleistocene divergence (~ 4.1 Ma) between *A. columbianum* and *A. delphiniifolium* (Fig. 3). However, future research should include sampling of multiple accessions of *A. delphiniifolium*

and include *A. uncinatum* in order to evaluate the biogeography of these sister taxa in North America. The most parsimonious explanation of the distribution of *A. delphiniifolium* in western North America and *A. uncinatum* in eastern North America would be to invoke a historically contiguous distribution of *A. columbianum*. This scenario would require divergence from *A. delphiniifolium* in western North America, followed by divergence with *A. uncinatum* in eastern North America.

Iltis (1965) suggested that a species pair distribution such as that of *A. columbianum* and *A. uncinatum* indicates western and eastern restriction during the Pleistocene followed by post-glacial migration with range overlap in the glaciated regions of northeastern United States. Based on our RAxML analysis, the clade that includes the Driftless Area and NY populations (Fig. 5) supports this model. The short internal branch lengths suggest a rapid dispersal, which only would have been possible if there was more contiguous suitable habitat. This contiguous suitable habitat might have occurred while glaciers were covering lower latitudes (Brink 1982). The parallel disjunct clade in our RAxML analysis that includes the Ohio populations could also support this model, provided the existence of intermediate populations that were extirpated because they did not survive in a cool refugium once the glacier retreated. In fact, the longer branch leading to the Ohio population in both the RAxML and SVDQuartet trees lends itself to the hypothesis that there were now-extinct, intermediate populations that would have broken up the long branch.

CONCLUSION and FINAL COMPARATIVE REMARKS:

The three species discussed in Chapters 2-4, *Rubus parviflorus*, *Oplopanax horridus,* and *Aconitum columbianum*, are examples of different categories of the Western North America-

GLR disjunct plants (See Chapter 1). We have provided evidence that they do not all share the same phylogeographic history, indicating pseudo-congruence within this disjunct pattern. Both *R. parviflorus* and *O. horridus* were restricted to western North America south of the glacier line during the Pleistocene with subsequent colonization northward. This was most likely followed by peri-glacial dispersal to the previously glaciated regions of the GLR that they now occupy. Our confidence in the timing of this eastward dispersal as peri-glacial and not pre-Pleistocene is influenced by the fact that the GLR populations of both *R. parviflorus* and *O. horridus* have recent common ancestry with populations in British Columbia, an area that was covered with ice. Alternative, earlier timing would require invoking persistence of populations in glacial refugia in British Columbia with dispersal outward from those refugia, and this scenario cannot be ruled out by our analyses. In contrast, the *A. columbianum* populations show a more complex history of dispersal that suggests a central North American ancestral range with parallel instances of eastward and westward dispersal. This history is not inconsistent with the model of west to east peri-glacial dispersal, but it is apparent that what might be considered a restrictive arctic alpine niche likely enabled its survival in high elevation "islands" in the Rocky Mountains during the Pleistocene (Pierce 2003).

The approach we took with genomic data illuminated the patterns of historical movement. We still have not achieved the "how" of this distribution. It will be difficult to answer this question with certainty given the different environmental and biotic context thousands of years in the past. Experimental models of fruit dispersal would be required, and even such experiments would be simplifications of the system. The question of how these plants arrived in the GLR is still open for debate. In particular, future research should include case studies of

wind-dispersed plants to assess the role and frequency of long-distance dispersal in this disjunct system.

Research on this disjunct distribution need not be restricted to species phylogeography. A well-studied example of a genus disjunct is *Rhodiola* (Crassulaceae), which is disjunct in the GLR extending through the St. Lawrence Seaway and eastern coast of North America (Guest 2009, DeChaine et al. 2013, Guest and Allen 2014). The species in western North America and eastern North America are not sister to each other and are hypothesized to have arrived in North America via different routes (Guest and Allen 2014). The western North American populations are disjunct between the southern Rocky Mountains and the Pacific Northwest and are hypothesized to have survived in glacial refugia north and south of the glacier (DeChaine et al. 2013). Research on additional disjunct species pairs would expand this perspective on deeper evolutionary and biogeographic divergence and could be compared to species disjuncts to test whether there are predictive temporal correlations between the age of divergence and biogeographic histories.

New techniques are continually being developed to push the boundaries of genomic data and statistical analysis for phylogeographic research (Harvey et al. 2016, Papadopoulou and Knowles 2016, Satler and Carstens 2016, Blischak et al. 2018, Smith et al. 2018). For example, programs such as SortaDate (Smith et al. 2018) are enabling users to filter sequence data for informative loci with multiple SNPs. This is particularly useful for reduced representation sequencing, such as GBS, because it supports the filtering of large datasets, enabling the use of downstream phylogenetic analyses such as BEAST and BUCKy (Larget et al. 2010, Harvey et al. 2016). In addition, models are being developed to test for hybridization across the genome (HyDe, Blischak et al. 2018), which is important in illuminating phylogeographic population

histories. As new approaches for using these data are being tested, more robust phylogeographic analyses will become possible. Efforts to uncover the migration and demographic histories of the western North American-Great Lakes Region disjuncts and assess their continued persistence in the GLR will continue to foster a greater understanding of the biodiversity of North America and the Great Lakes Region.

Table 1: List of populations sampled for genotyping-by-sequencing, their locality, and the number of samples per population in the final, cleaned data set. The population names have the state/province appended to the beginning (except for Isle Royale populations in Michigan, which use 'IR') for easier interpretation of their general location.

Name	Locality	Number of samples in final data set
WA_Haney	47.321854, -120.526916	6
OR_Willamette	44.397618, -122.135084	8
AZ	34.996795, -111.753797	8
CO_Arapaho	39.865166, -105.749402	4
CO_South	37.913620, -104.990540	8
CO_Gunnison	38.632651, -106.577836	4
WY_Bighorn	44.384099, -107.166453	8
SD_BlackHills	44.123037, -103.759038	10
IA_StOlaf	42.917451, -91.385052	8
WI_HayValley	43.678524, -90.604575	4
OH	39.419446, -82.532335	4
NY_Peekamoose	41.953881, -74.398207	7

Table 2a. The genetic diversity metrics for each population based on all loci that were present in 100% of the populations and at least 75% of individuals within a population: number of private alleles, percent of loci that were polymorphic, observed and expected heterozygosity, nucleotide diversity (Pi), and the inbreeding coefficient (F_{IS}). Diversity statistics were measured with kernel-smoothing and 100 bootstrap replicates. F_{IS} was calculated with the Bonferroni correction with a base p-value of 0.05. **2b.** The comparison between the GLR and western North American populations and **2c.** The comparison between the eastern, NY population and the western North American populations. P-values are from unpaired, two-sample t-tests.

Table 2a

Pop_ID	# of Sites	# Private Alleles	% Polymorphic Loci	Coded Region
CO_Arapaho	595707	130	0.08024	West
AZ	595707	240	0.17559	West
WY_Bighorn	595707	84	0.08427	West
SD_BlackHills	595707	262	0.15763	BH
CO_South	595707	178	0.09619	West
CO_Gunnison	595707	254	0.11885	West
WA_Haney	595707	699	0.17744	West
WI_HayValley	595707	223	0.256	GLR
OH	595707	525	0.116	GLR
NY_Peekamoose	595707	337	0.11348	East
IA_StOlaf	595707	318	0.1682	GLR
OR_Willamette	595707	449	0.10878	West

Table 2a (cont'd)

Pop_ID	Obs Het	Exp Het	Pi	F_{IS}
CO_Arapaho	0.00037	0.00029	0.00033	-0.00006
AZ	0.00056	0.00051	0.00055	0.00001
WY_Bighorn	0.0003	0.00026	0.00028	-0.00002
SD_BlackHills	0.00048	0.00047	0.0005	0.00007
CO_South	0.00033	0.00027	0.00029	-0.00005
CO_Gunnison	0.00047	0.00041	0.00047	0.00002
WA_Haney	0.0006	0.00054	0.0006	0.00003
WI_HayValley	0.0005	0.00084	0.00091	0.00085
OH	0.00049	0.00043	0.0005	0.00005
NY_Peekamoose	0.00042	0.00037	0.0004	-0.00001
IA_StOlaf	0.00052	0.00051	0.00055	0.0001
OR_Willamette	0.00057	0.00037	0.0004	-0.00033

Table 2b

	Mean GLR Value	Mean Western Value	p-value
Private_alleles	355.3333	290.5714	0.6537
Perc_Polymorphic_Loci	0.1800667	0.1201943	0.1205
Obs_Het	0.000503	0.00045714	0.5516
Exp_Het	0.000593	0.00037857	0.06692
Pi	0.000653	0.00041714	0.06065
Fis	-0.00033	-5.71E-05	0.05261

Table 2c

	Mean Eastern Value	Mean Western Value	p-value
Private_alleles	337	290.5714	0.8465
Perc_Polymorphic_Loci	0.11348	0.1201943	0.8824
Obs_Het	0.00042	0.00045714	0.7888
Exp_Het	0.00037	0.00037857	0.9463
Pi	0.0004	0.00041714	0.9036
Fis	-0.00001	-5.71E-05	0.7366

Fig. 1 RAxML tree of *Aconitum* using concatenated nrITS, nrETS, and 7 chloroplast markers with a model of GTR + Γ and 100 bootstrap replicates. Bootstrap support >75 is listed above the branches. The two subgenera are monophyletic, except for *A. moschatum*, which is typically placed in subg. *Lycoctonum*, although this is debated. The star indicates the divergence of *A. columbianum* with *A. delphiniifolium*.

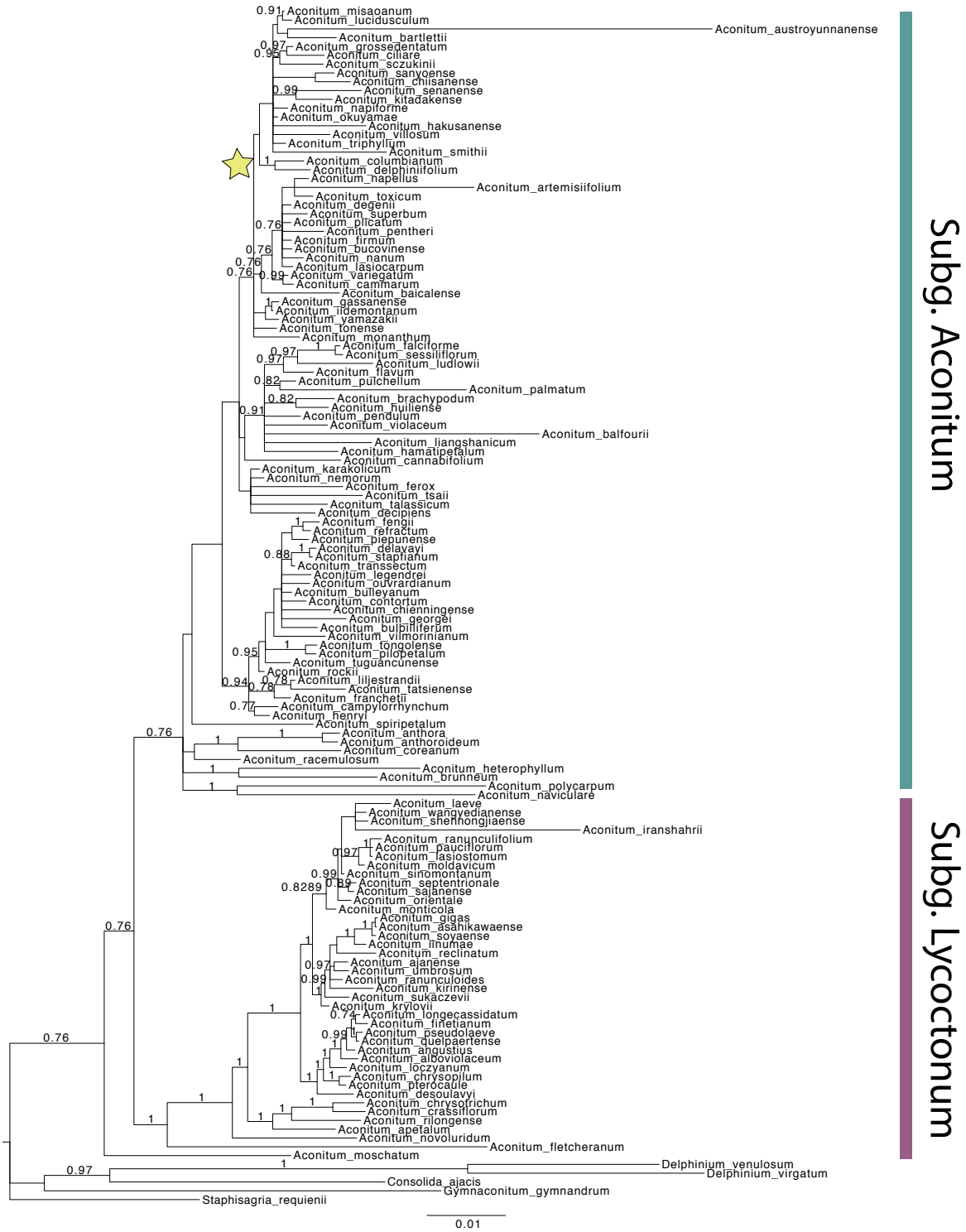


Fig. 2 MyBayes consensus tree of *Aconitum* using concatenated nrITS, nrETS, and 7 chloroplast markers and designating *Staphisagria requienii* as the outgroup. We ran a GTR model of substitution and ran100 bootstrap replicates. Posterior probability above 0.75 is listed above the branches. The two subgenera are monophyletic, except for *A. moschatum*, which is typically placed in subg. *Lycoctonum*, although this is debated. The star is to the left of the node where *A. columbianum* and *A. delphiniifolium* diverge.

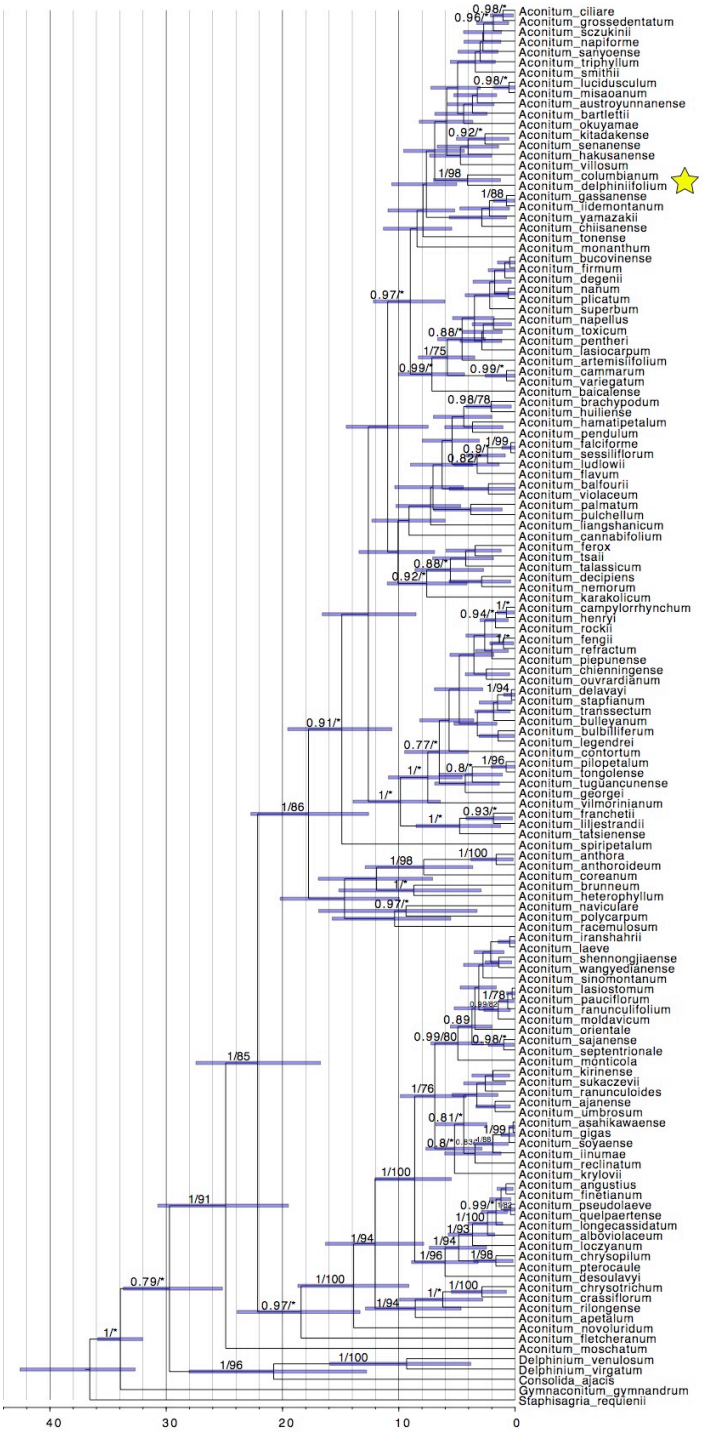


Fig. 3 BEAST chronogram of *Aconitum* from the nrITS + nrETS + 7 cpDNA partitioned data set. We used a secondary date calibration of 26.02 Ma at the stem of the MRCA of *Aconitum* + *Gymnaconitum* + *Delphinium* (*sensu* Jabbour and Renner 2012). Node bars indicate 95% HPD, time axis is in millions of years. Posterior probabilities and ML bootstrap support are shown for PP values >= 0.75 (PP/bootstrap). Asterisks indicate that the ML tree did not include this topology or that the ML bootstrap support was < 75. The star indicates the divergence of *A. columbianum* and *A. delphiniifolium*.

Fig 4. RAxML tree estimated with GTR + Γ using the INDIV data set of 1,882 concatenated SNPs and implementing the Lewis ascertainment bias correction. Bootstrap values above 70 are shown above the branches.

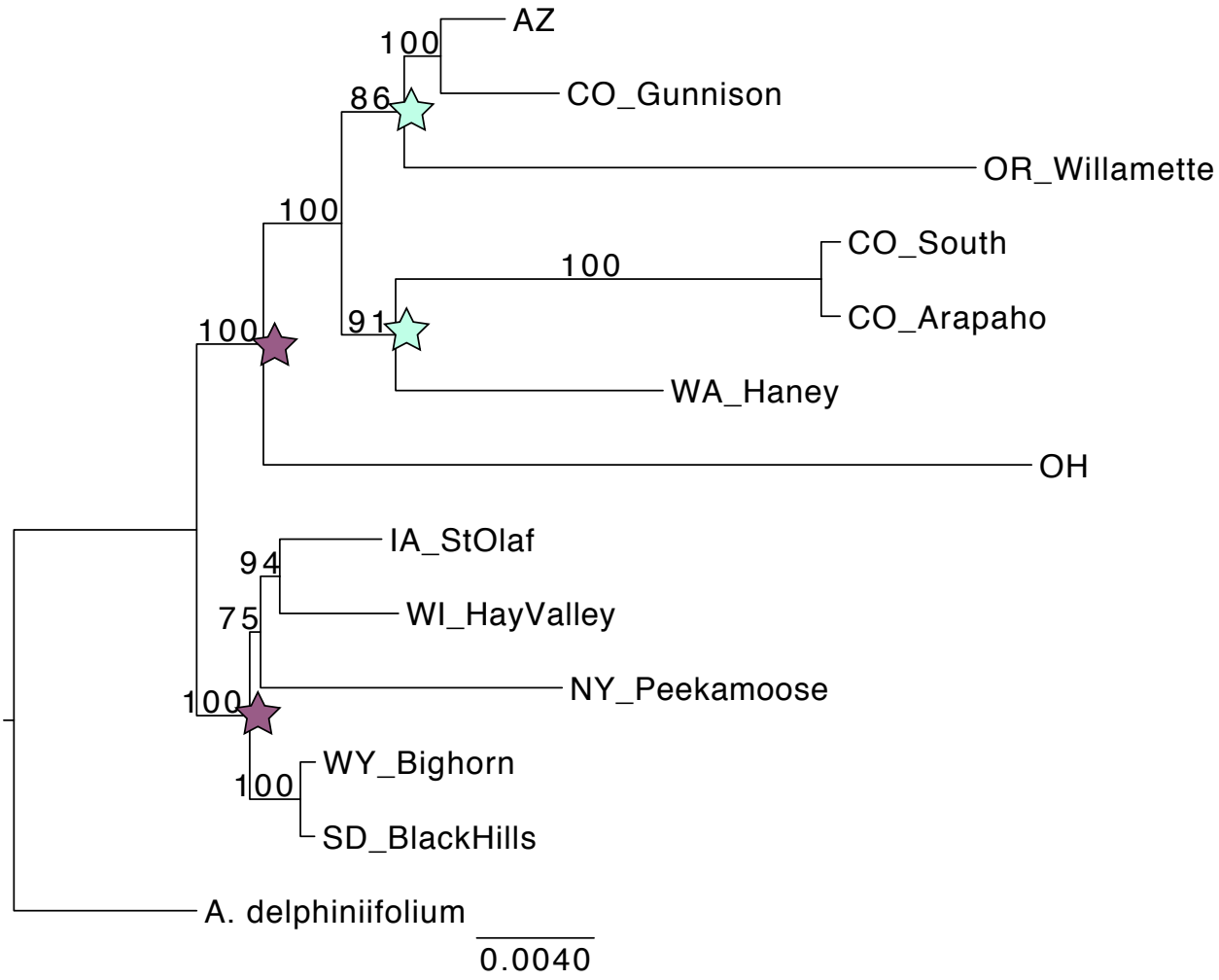


Fig. 5 RAxML tree from CONS data set of concatenated GBS SNPs using Lewis ascertainment bias correction. Bootstrap support values are written above the branches. Purple stars indicate two eastern North American disjunctions . Blue stars indicate two independent instances of dispersal or gene flow between a central Colorado range and western North America.

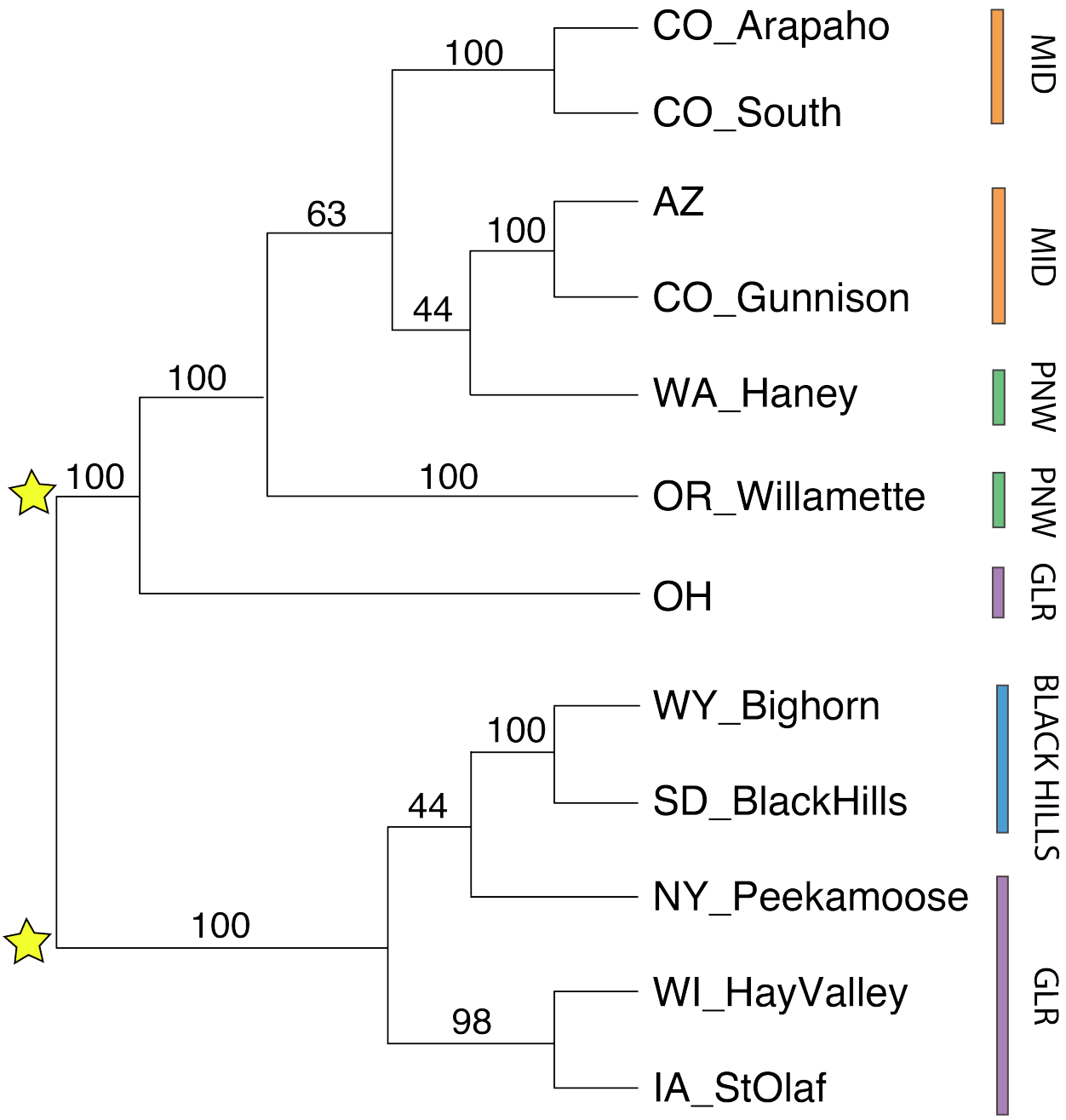


Fig. 6 SVDQuartets summary tree for all quartets across populations, without the outgroup. The tree was rooted with the RAxML topology. Individuals were partitioned into populations and ambiguity codes were distributed. 100 bootstrap replicates were conducted on subsets of 100,000 quartets and values are given above branches. Yellow stars indicate the clades containing two independent eastern dispersals.

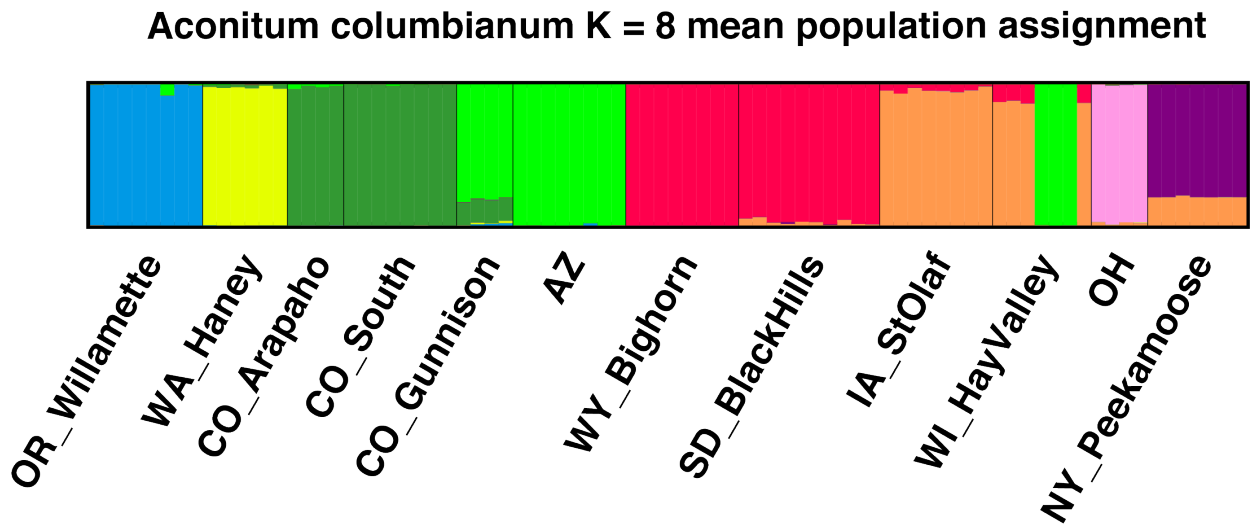


Fig. 7 Plot of Structure population assignments using best K = 8. The red admixture in the Iowa and Wisconsin populations indicates one instance of eastern gene flow. The bright green found in Arizona + Colorado as well as Wisconsin indicates a second instance of eastern gene flow. The minor bright green admixture in the Oregon population indicates one instance of western gene flow. The minor dark green admixture in the Washington population indicates a second instance of western gene flow.

REFERENCES

Anderson, B.M. et al., 2017. Genotyping-by-Sequencing in a Species Complex of Australian Hummock Grasses (Triodia): Methodological Insights and Phylogenetic Resolution G. Sun, ed. *PLoS ONE*, 12(1), pp.e0171053–34.

Anderson, C.L., 2005. Dating phylogenetically basal eudicots using rbcL sequences and multiple fossil reference points. 92(10), pp.1737–1748.

Association, D.S.S. et al., *PAUP* Phylogeny Analysis Using Parsimony (* and other methods), version 40b10*.

Bell, C.D., Soltis, D.E. & Soltis, P.S., 2010. The age and diversification of the angiosperms re-revisited. *American Journal of Botany*, 97(8), pp.1296–1303.

Blischak, P.D. et al., 2018. HyDe: a Python Package for Genome-Scale Hybridization Detection. D. Posada, ed. *Systematic Biology*, 66, p.620.

Iltis, H.H. 1965. The genus Gentianopsis (Gentianaceae): transfers and phytogeographic comments. *Sida, Contributions to Botany* 2, pp.129-154.

Bouckaert, R. et al., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. A. Prlic, ed. *PLoS computational biology*, 10(4), p.e1003537.

Brigham-Grette, J. & Gualtieri, L.M., 2003. Chlorine-36 and 14C chronology support a limited last glacial maximum across central Chukotka, northeastern Siberia, and no Beringian ice sheet. *Quaternary …*, 59, pp.386–398.

Brink, D.E., 1980. Reproduction and Variation in Aconitum columbianum (Ranunculaceae), with Emphasis on California Populations. *American Journal of Botany*, 67(3), pp.263–273.

Brink, D., 1982. Tuberous Aconitum (Ranunculaceae) of the Continental United States: Morphological Variation, Taxonomy and Disjunction. *Bulletin of the Torrey Botanical Club*, 109(1), p.13.

Brink, D. & deWet, J.M.J., 1980. Interpopulation variation in nectar production in Aconitum columbianum (Ranunculaceae). *Oecologia*, 47(2), pp.160–163.

Brink, D.E. & Woods, J.A., 1997. Aconitum. *Flora of North America*, 3, pp.191-195.

Britton, T. et al., 2007. Estimating divergence times in large phylogenetic trees. F. Anderson, ed. *Systematic Biology*, 56(5), pp.741–752.

Catchen, J. et al., 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22(11), pp.3124–3140.

Chifman, J. & Kubatko, L., 2014. Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, 30(23), pp.3317–3324.

Cole, C.T. & Kuchenreuther, M.A., 2001. Molecular markers reveal little genetic differentiation among Aconitum noveboracense and A. columbianum (Ranunculaceae) populations. *American Journal of Botany*, 88(2), pp.337–347.

DeChaine, E.G. et al., 2013. On the evolutionary and biogeographic history of Saxifraga sect. Trachyphyllum (Gaud.) Koch (Saxifragaceae Juss.). S. Lavergne, ed. *PLoS ONE*, 8(7), p.e69814.

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

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

Eaton, D.A.R. et al., 2016. Misconceptions on Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants. *Systematic Biology*, 56, p.syw092.

Emadzade, K. et al., 2011. The biogeographical history of the cosmopolitan genus Ranunculus L. (Ranunculaceae) in the temperate to meridional zones. *Molecular Phylogenetics and Evolution*, 58(1), pp.4–21.

Federal review, 2008, citation: 73 FR 21643 21645.

Fior, S. et al., 2013. Spatiotemporal reconstruction of the Aquilegia rapid radiation through next-generation sequencing of rapidly evolving cpDNA regions. 198(2), pp.579–592.

Guest, H.J., 2009. Dissertation: Systematic and phylogeographic implications of molecular variation in the western North American roseroom, Rhodiola integrifolia (Crassulaceae). University of Victoria.

Guest, H.J. & Allen, G.A., 2014. Geographical origins of North American Rhodiola(Crassulaceae) and phylogeography of the western roseroot, Rhodiola integrifolia M. Carine, ed. *Journal of Biogeography*, 41(6), pp.1070–1080.

Hardin, J.W., 1964. Variation in Aconitum of Eastern United States. *Brittonia*, 16(1), pp.80–94.

Harvey, M.G. et al., 2016. Sequence Capture versus Restriction Site Associated DNA Sequencing for Shallow Systematics. *Systematic Biology*, 65(5), pp.910–924.

Hong, Y. et al., 2017. Phylogeny and reclassification of Aconitum subgenus Lycoctonum (Ranunculaceae). Z.-J. Liu, ed. *PLoS ONE*, 12(1), p.e0171038.

Hultén, E., 1941-1950. Flora of Alaska and Yukon. Vol. 4

Jabbour, F. & Renner, S.S., 2012. A phylogeny of Delphinieae (Ranunculaceae) shows that Aconitum is nested within Delphinium and that Late Miocene transitions to long life cycles in the Himalayas and Southwest China coincide with bursts in diversification. *Molecular Phylogenetics and Evolution*, 62(3), pp.928–942.

Jaramillo-Correa, J.P. et al., 2009. Inferring the past from the present phylogeographic structure of North American forest trees: seeing the forest for the genes. *Canadian Journal of Forest Research*, 39(2), pp.286–307.

Kadota, Y., 2001. Systematic studies of Asian Aconitum (Ranunculaceae): 7. A new species and a new form of subgenus Lycoctonum from Hokkaido, Japan. *J. Jap. Bot*, 76(1),pp.20-27.

Kearse, M. et al., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), pp.1647–1649.

Kuchenreuther, M. 1991. Dissertation: Life history, demography and genetics of Aconitum noveboracense: implications for preservation and management of a threatened species. University of Wisconsin-Madison.

Kuchenreuther, M. 1996. The natural history of Aconitum noveboracense Gray (northern monkshood), a federally threatened species. *Journal of the Iowa Academy of Science: JIAX*, 103(3-4)pp.57-62.

Larget, B.R. et al., 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, 26(22), pp.2910–2911.

Leaché, A.D. et al., 2015. Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. *Systematic Biology*, 64(6), pp.1032–1047.

Luo, Y., Zhang, F.-M. & Yang, Q.-E., 2005. Phylogeny of Aconitum subgenus Aconitum (Ranunculaceae) inferred from ITS sequences. *Plant Systematics and Evolution*, 252(1-2), pp.11–25.

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

Merkel, D., 2014. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014(239), p2.

Miller, M.A., Pfeiffer, W. & Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In 2010 Gateway Computing Environments Workshop (GCE). IEEE, pp. 1–8.

Mitton, J.B., Kreiser, B.R. & Latta, R.G., 2000. Glacial refugia of limber pine (Pinus flexilis James) inferred from the population structure of mitochondrial DNA. *Molecular Ecology*, 9(1), pp.91–97.

Papadopoulou, A. & Knowles, L.L., 2016. Toward a paradigm shift in comparative phylogeography driven by trait-based hypotheses. *Proceedings of the National Academy of Sciences*, 113(29), pp.8018–8024.

Paris, J.R., Stevens, J.R. & Catchen, J.M., 2017. Lost in parameter space: a road map for stacks S. Johnston, ed. *Methods in Ecology and Evolution*, 8(10), pp.1360–1373.

Pierce, K.L., 2003. Pleistocene glaciations of the Rocky Mountains. In *The Quaternary Period in the United States*. Developments in Quaternary Sciences. Elsevier, pp. 63–76.

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

Rambaut, A. et al., 2018. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. E. Susko, ed. *Systematic Biology*, 22, p.341.

Reaz, R., Bayzid, M.S. & Rahman, M.S., 2014. Accurate Phylogenetic Tree Reconstruction from Quartets: A Heuristic Approach R. Wu, ed. *PLoS ONE*, 9(8), pp.e104008–13.

Rochette, N.C. & Catchen, J.M., 2017. Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Publishing Group*, 12(12), pp.2640–2659.

Ronquist, F. et al., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3), pp.539–542.

Rosenberg, N.A. 2004. *Distruct*: a program for the graphical display of population structure. *Molecular Ecology Notes*, 4,pp.137-138.

Satler, J.D. & Carstens, B.C., 2016. Phylogeographic concordance factors quantify phylogeographic congruence among co-distributed species in the Sarracenia alatapitcher plant system. *Evolution*, 70(5), pp.1105–1119.

Shafer, A.B.A. et al., 2010. Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Molecular Ecology*, 19(21), pp.4589–4621.

Smith, S.A. & Brown, J.W., 2018. Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany*, 105(3), pp.302–314.

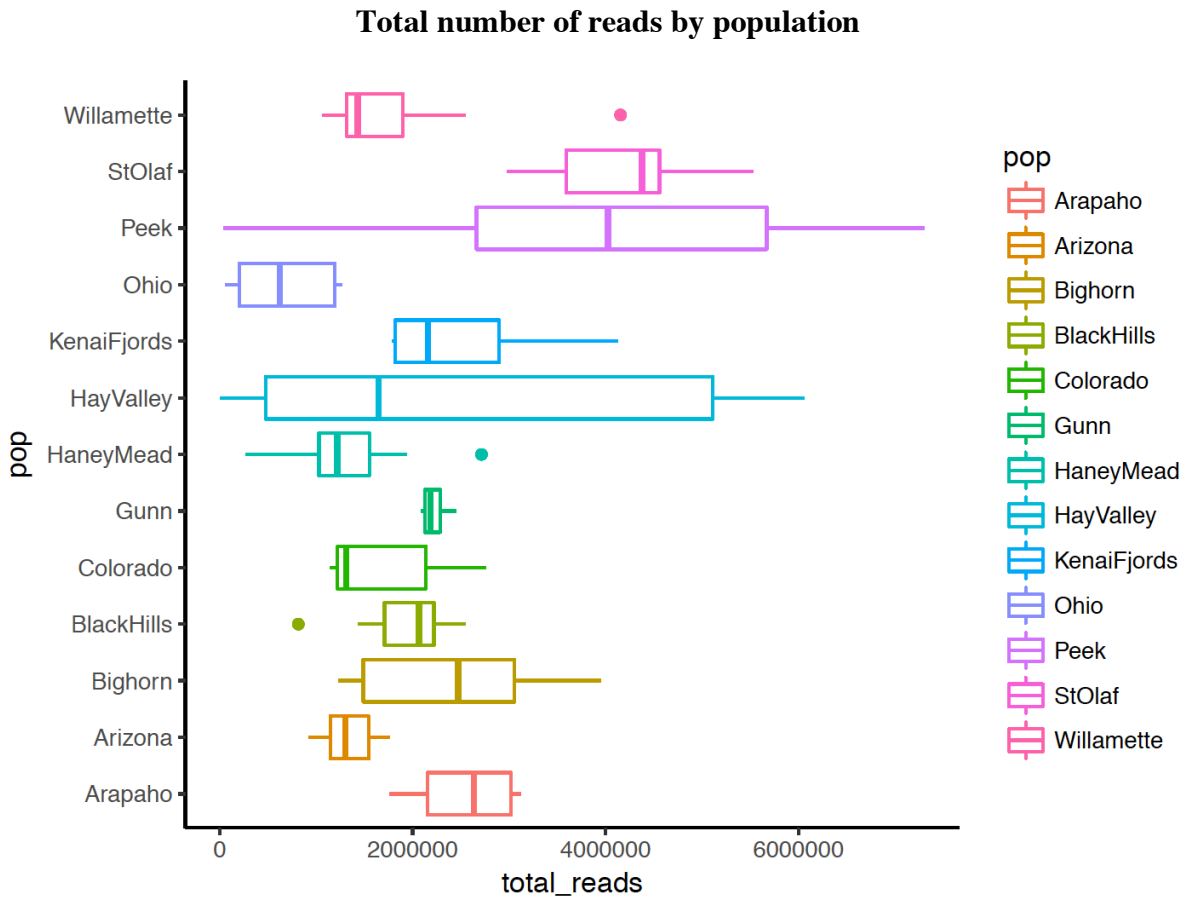
Smith, S.A., Brown, J.W. & Walker, J.F., 2018. So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. H. Escriva, ed. *PLoS ONE*, 13(5), p.e0197433.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), pp.1312–1313.

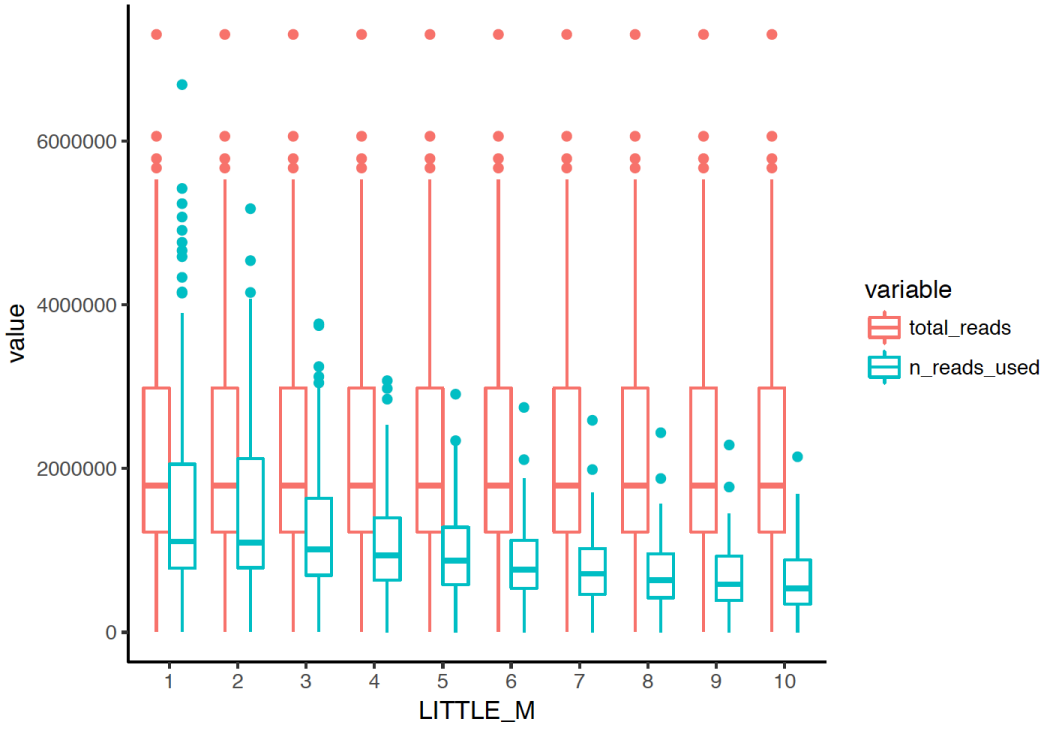
Wang, L. et al., 2009. History and evolution of alpine plants endemic to the Qinghai-Tibetan Plateau: Aconitum gymnandrum (Ranunculaceae). *Molecular Ecology*, 18(4), pp.709–721.

Wang, W. et al., 2016. The rise of angiosperm-dominated herbaceous floras: Insights from Ranunculaceae. *Nature Publishing Group*, 6(1), pp.1–8.

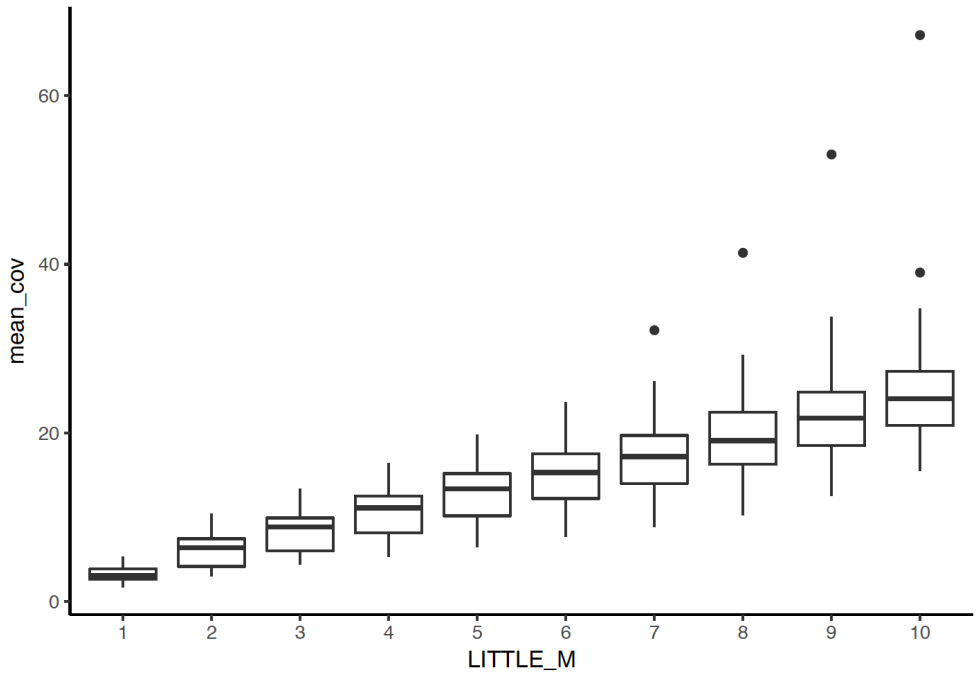
Appendix 7: Results from Stacks assembly optimization.



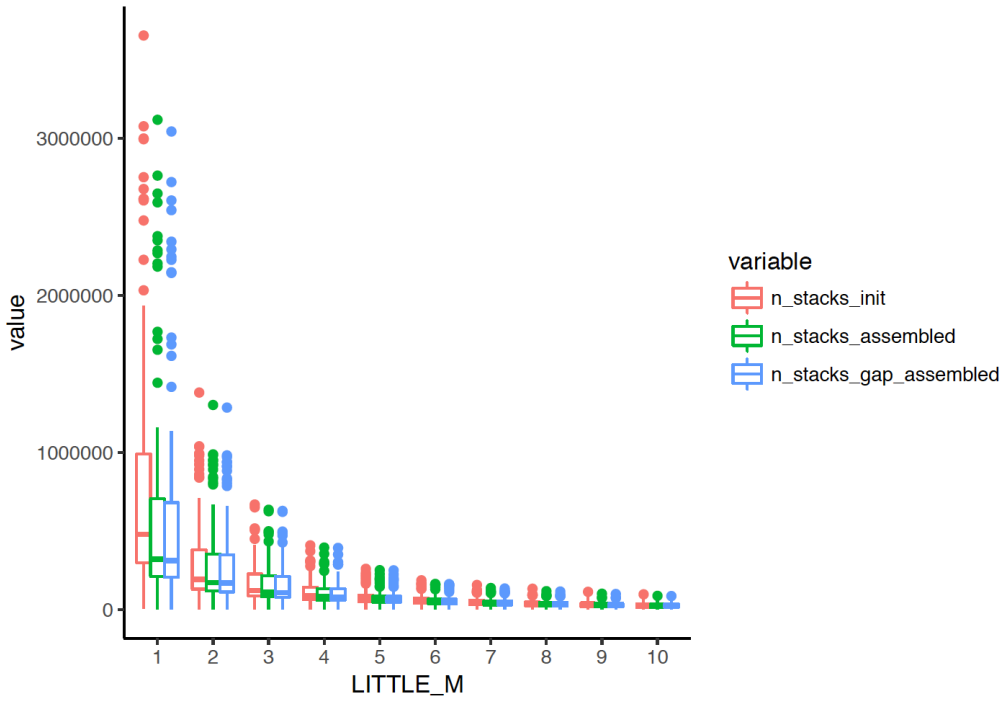
Number of reads used based on 'm' assembly parameter



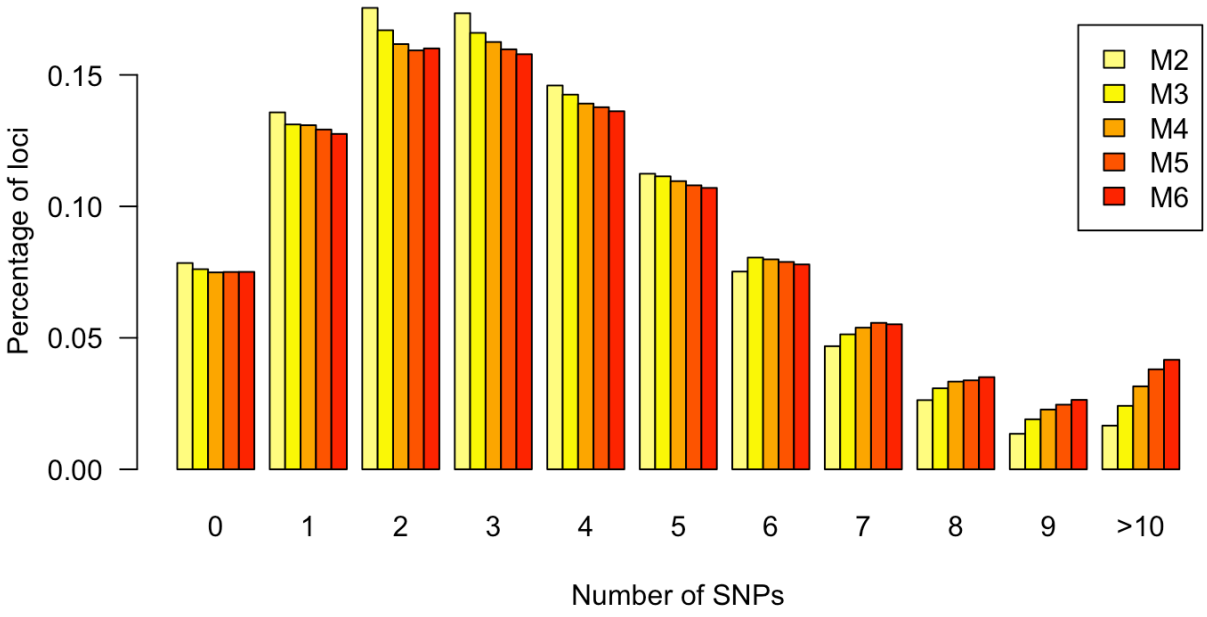
Mean coverage per locus based on 'm' assembly parameter



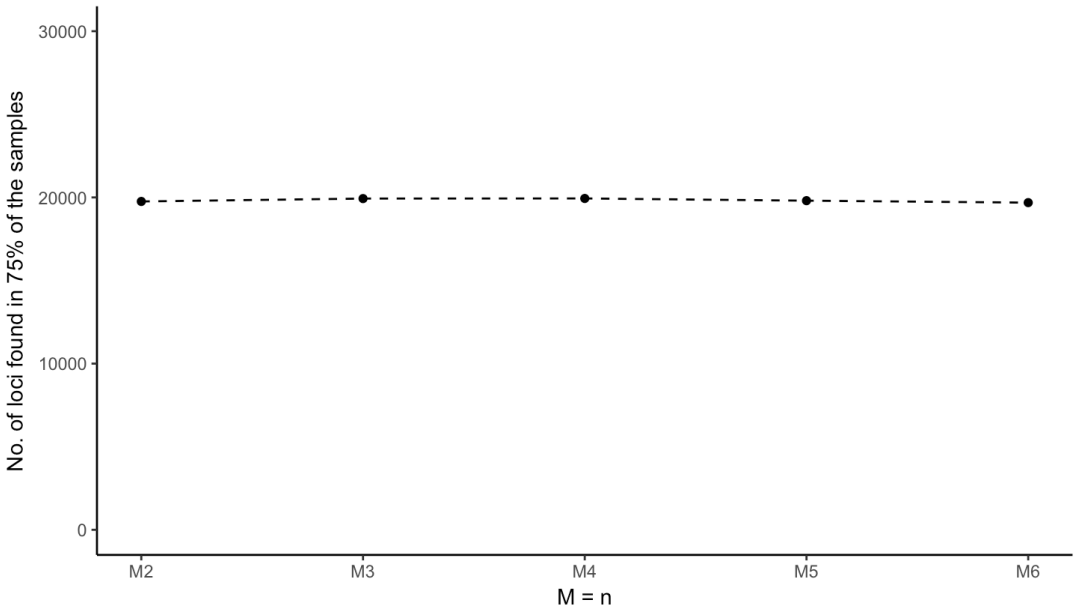
Number of loci (stacks) formed based on 'm' assembly parameter



Proportion of loci with n-SNPs based on "M" assembly parameter



Number of loci in 75% of samples for each value of M==n



Proportion of heterozygotes at each locus (SNP) for M = n = 4