INDEPENDENT REPLICATION OF PHYLOGEOGRAPHIES:

HOW REPEATABLE ARE THEY?

by

CLAYTON REED MERZ

A THESIS

Presented to the Department of Biology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Master of Science

September 2012

THESIS APPROVAL PAGE

Student: Clayton Reed Merz

Title: Independent Replication of Phylogeographies: How Repeatable Are They?

This thesis has been accepted and approved in partial fulfillment of the requirements for the Master of Science degree in the Department of Biology by:

William E. Bradshaw	Chairperson
Christina Holzapfel	Member
John Conery	Member

and

Kimberly Andrews Espy	Vice President for Research & Innovation/Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded September 2012

THESIS ABSTRACT

Clayton Reed Merz

Master of Science

Department of Biology

September 2012

Title: Independent Replication of Phylogeographies: How Repeatable Are They?

Herein we tested the repeatability of RAD-seq phylogeographic construction by creating a second, independent phylogeography of the pitcher-plant mosquito, *Wyeomyia smithii*. We sampled 25 populations drawn from different localities nearby previous collection sites and used these new data to construct a second, independent phylogeography to test the reproducibility of phylogenetic patterns. Our previous phylogeography was based on 3,741 phylogenetically informative markers from 21 populations and rooted with mitochondrial COI. The present phylogeography was based on 16,858 informative markers and rooted with RAD-seq. We found correspondence between clades at the extremes of *W. smithii*'s distribution; however, there were several discrepancies between the trees, including the refugium that gave rise to all post-glacial populations. We observed that combining all 46 populations resolved these discrepancies and, equally importantly, that extensive taxon sampling in areas of historical importance is more valuable than increasing the number of informative sites in establishing an accurate, robust phylogeography.

This thesis includes unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Clayton Reed Merz

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
New Mexico Institute of Mining and Technology, Socorro, NM

DEGREES AWARDED:

Master of Science, Biology, 2012, University of Oregon
Bachelor of Science, Biology, 2008, New Mexico Institute of Mining and Technology

AREAS OF SPECIAL INTEREST:

Insect Behavior and Evolution
Phylogenetics
Science Education

PROFESSIONAL EXPERIENCE:

Teaching Assistant, Department of Biology, University of Oregon, Eugene, 2008-2012

GRANTS, AWARDS, AND HONORS:

Graduate Teaching Fellowship, Biology, 2008-2012

PUBLICATIONS:

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16196-16200.

ACKNOWLEDGMENTS

I thank William Bradshaw and Christina Holzapfel for assisting in the creation of this project and for their assistance throughout. I thank Kevin Emerson for teaching me in the lab and in helping in innumerable ways. I thank Julian Catchen and Victor Hansen-Smith for their excellent collaboration on the digital side of this project. I thank John Conery, Joe Thornton and William Cresko for useful discussion. Funding was generously provided by NSF grants IOS-083998, IOS-1048276 and DEB-0917827 to WEB, NIH NRSA Fellowship 5F3GM095213 to JMC, NSF IGERT training grant DGE0504727 to VH-S, and NIH grant 1R24GM079486-01A1 to John Postlethwait. I am grateful to the Gros Morne National Park, The Nature Conservancy, the North Temperate Lakes LTER, Cheraw State Park, the Weymouth Woods-Sandhills Nature Preserve, the Tobyhanna State Park, the *Grand Bay* National Estuarine Research Reserve and Ada and George Simons for permission for and assistance in making collections on their lands. Both the Centers for Disease Control and the Department of Agriculture were most helpful in arranging import permits from Canada for *W. smithii*.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Each chapter of this work includes contributions from a number of collaborators. Julian Catchen performed the analysis and distillation of raw sequence files that transformed them into phylogenetically useful data. Kevin Emerson assisted with troubleshooting in the lab and during the analysis process. Victor Hansen-Smith provided Bayesian posterior probabilities on the combined tree and assisted in troubleshooting other analyses. William Bradshaw and Christina Holzapfel contributed to editing the manuscript and planning the experiments. I performed all the molecular techniques, analyzed phylogenetic data to yield all the trees, and did all the writing.

In recent years, a number of studies have appeared using next-generation sequencing (NGS) to construct phylogeographies and phylogenetic trees in novel ways (Lemmon and Lemmon 2012; McCormack et al. 2012ab; Rubin et al. 2012; Zellmer et al. 2012). Both of the most common forms of NGS (Illumina and 454 pyrosequencing) use large numbers of short reads to assemble contigs, from which SNPs can be identified (among other applications). NGS technologies have numerous benefits over older technologies, including high sequence-to-cost ratios, ease of incorporating numerous markers, and ease of implementation in non-model organisms. The utility of restriction-site associated DNA sequencing (RAD-seq) (Miller et al. 2007; Baird et al. 2008; Amores et al. 2011; Etter et al. 2011) in producing high-density, enriched, genome-wide markers for a variety of studies is now well-established in the literature (Davey et al. 2011, Cronn et al. 2012,

Rubin et al. 2012, McCormack et al. 2012a). Successful studies using RAD-seq are increasingly common in many types of non-model organisms, including the pitcher-plant mosquito (Emerson et al. 2010), the threespine stickleback (Hohenlohe et al. 2010) and the diamondback moth (Baxter et al. 2011).

The term "resampling" in phylogeography and phylogenetics almost exclusively refers to bootstrapping or other methods of subsampling a single data set. Herein, we present a different kind of resampling: a two-stage phylogeographic analysis using near-by but completely independent sets of populations. This approach allowed us to compare two separate analyses, which we used to test the reproducibility of results of closely related populations of the pitcher-plant mosquito, *Wyeomyia smithii*. In this study, we compared an initial tree based on 54bp RAD-seq reads and rooted with mitochondrial cytochrome oxidase subunit I (COI) (Emerson et al. 2010) with an independently created tree based on 80bp reads and rooted with RAD-seq. This two-tree approach allowed us to examine the robustness and reproducibility of the RAD-seq method, resolving a potential shortcoming in the use of RAD-seq for constructing phylogeographies (Twyford and Ennos 2012; McCormack et al. 2012a). As McCormack et al. (2012) note, the fact that RAD-seq samples are closely tied to specific restriction sites means that RAD-seq phylogeographies should be expected to be relatively reproducible.

Beyond comparing the two data sets, we analyzed the combined set of all 46 populations as a whole, validating new conclusions differing from our earlier work on *W. smithii* phylogenetics (Emerson et al. 2010). Increasing the coverage of localities in the mid-Atlantic region (Maryland and New Jersey), where genetic relationships were not well defined, lead us to question our previous conclusions about the location of *W.*

smithii's glacial refugium and enabled us to confirm distinct northeastern and northwestern clades and their relative modes of post-glacial range expansion.

Wyeomyia smithii is especially well suited to this kind of phylogeographic analysis because of its unusually broad distribution, ranging from the Gulf of Mexico to northern Canada (30-54°N). Earlier studies from our lab involving allozymes or COI clearly established basal populations along the Gulf of Mexico and an ancient migration to the Carolina coastal plain and Piedmont, but left the relationships among more recently dispersed, post-glacial populations unresolved (Armbruster et al. 1998; Emerson et al. 2010). Given the position of the Laurentide ice sheet at the last glacial maximum ca. 22,000 - 19,000 y BP (Dyke et al. 2002; Colgan et al. 2003), all present-day populations north of ca. 41º N Latitude must have arisen within the last 19,000 y (Yokoyama et al. 2000). The first RAD-seq data set based on 3,741 phylogenticaly informative sites indicated that the glacial refugium of post-glacial populations resided in the southern Applachian Mountains, that post-glacial populations appeared to form two major clades diverging to the northeast and the northwest, and suggested that northwestern populations may have been founded along parallel longitudes due to anticyclonic winds along the retreating glacial front (Emerson et al. 2010). The replicated use of the RAD-seq approach in the second data set plus the combination of both replicates into a consensus phylogeography enabled us to validate or invalidate these conclusions.

CHAPTER II

METHODS

Populations

Sampled populations ranged from the Gulf Coast (30-31°N) to Newfoundland (50°N) and northwestward to Saskatchewan (54°N), from 10-1,000m elevation at 35°N in North Carolina, and from 10-595m elevation at 40-41°N in New Jersey and Pennsylvania. Throughout the text, populations are referred to by their state or province of origin, followed by an identifying number when more than one population was sampled in a given state or province. In each case, wild-caught individuals were used.

RAD library creation and sequencing

In order to test if similar clade structure is found in replicated phylogeographic datasets, we used two distinct datasets. The first represents 21 populations spanning much of the range of *W. smithii* (Emerson et al. 2010). The second, presented here, includes 25 distinct populations that represent populations geographically close to those in the previous analysis as well as populations expanding the sampled range of *W. smithii* to the western and eastern extremes of the Canadian range. The two datasets were first treated separately. When the two datasets were merged, all sequence data was truncated to 54 bp as *Stacks* requires sequences of the same length (Catchen et al. 2011).

For each of the 25 new populations sequenced for the present study, genomic DNA was extracted from pools of six adult *W. smithii* frozen at -80°C either dry or in alcohol, using a Qiagen DNeasy extraction column. DNA was digested with high-fidelity

SbfI (New England Biolabs). Illumina adapters, including a population-specific five-base barcode and partial *SbfI* sequence (Etter et al. 2011), were ligated to the cut ends. This DNA was sheared by sonication to reduce its size and a second primer-containing Y-shaped Illumina P2 adapter was ligated to the fragments. PCR was used to amplify the RAD libraries, which each included normalized amounts of RAD-tags from 9 or 10 populations. Three libraries were single-end sequenced (80bp) with three lanes of an Illumina GAIIx.

The *Stacks* pipeline (Catchen et al. 2011) was used to analyze the RAD data. The pipeline first applied stringent quality filters to the RAD sequences to remove potentially erroneous sequences. All exactly matching sequences were then grouped into stacks. Loci were then defined as sets of stacks such that for each stack in the locus there was another stack in the locus that is at most one nucleotide divergent. SNP detection was performed using a maximum likelihood framework (Catchen et al. 2011). All polymorphic stacks *within* populations were filtered out and for phylogeogenetic analyses, only SNPs that varied *between* at least two populations were included. This filtering of SNPs reduces the overall amount of data but increases the ability to identify clean phylogenetic signals. Once a SNP is identified, it is inserted into a data matrix that resembles an alignment file, but is produced from concatenated SNPs without surrounding sequences.

Phylogenetics

Each dataset was treated similarly for phylogenetic analysis, with each set being analyzed three ways – Maximum Parsimony (MP), Maximum Likelihood (ML), and

Bayesian Inference (BI) methods. Parsimony analysis used PAUP* (Swofford 2002) with 200 bootstrap replicates for node supports and a standard heuristic search. For BI and ML, our first analyses used jModelTest (Posada 2008); the new and combined data sets used PAUP*-based ModelTest (Posada and Crandall 1998). Both procedures selected TVM for the original and new data set, and TVM+Γ for the combined data set; in all three cases AIC chose the same model. For ML analysis, the model parameters found in ModelTest were input into PhyML v. 3.0 (Guindon and Gascuel 2003) to define a custom model to replicate TVM. The parameters were held constant during the analysis, without optimization. This ML topology and aLRT statistics (Anisimova and Gascuel 2006) are used in figures, because they provide a measure of node support of a more transparent meaning: the likelihood difference between the presented ML tree and the highest-likehood tree not containing that node. Bayesian inference was based on MrBayes v. 3.1.2 (Ronquist and Huelsenbeck 2003). As for ML, the ModelTest parameters were entered as priors and held constant for both the previous and current data sets. For the combined 46 populations, MrBayes was started from a random tree and run with eight chains for 5 million generations, sampling every 10 generations, to create 500,000 total samples. In order to avoid an "initial transient" that is unrepresentative of the true equilibrium distribution, an initial "burn-in" period of 59,000 samples was discarded. This burn-in period was determined by sliding a window, 1000 samples wide, across the function defined by "generations versus log-likelihood"; the end of the burn-in period was defined as the first point where the likelihood function within the sliding window did not significantly deviate from a linear regression with zero slope. Finally, from the remaining 441,000 samples, posterior probabilities of each clade on the ML tree

were calculated by counting the proportion of Bayesian samples containing that clade. The SumTrees program of the Dendropy package (Sukumaran and Holder 2010) was used to map the final posterior probability values onto the Newick-formatted tree.

CHAPTER III

RESULTS

Repeatability of the phylogeography

Figure 1 compares the initial phylogeography based on 54bp RAD-seq and rooted with COI (Emerson et al. 2010) with the independently-determined, replicate phylogeography based on and rooted with the 80 bp RAD-seq. Both trees were rooted using *Wyeomyia mitchelli* and *W. vanduzeei* as outgroups and indicate the Gulf Coast clade as basal to all northern populations. There is close correspondence between clades at the extremes of *W. smithii*'s distribution as represented by the Gulf Coast, North Carolina (NC) Lowland, and Northwest groupings. There are three major differences between the two trees. First, what appeared to be a basal clade in the NC Mountains in the first tree (NCmt1, 3-5) emerges within the Mid-Atlantic group in the second tree (NCmt2). Second, the first tree placed the Northeast populations (ME1-2) as basal to the Northwest clade while the second tree places a northeast clade (MA, NS1-2, NL1-2) within the Mid-Atlantic grouping. Third, a Pocono Mountain population (PA1) is basal to both the Northeast and Northwest clades in the first tree but the other Pocono Mountain population (PA2) is basal only to the northwest clade in the second tree.

Consensus (combined) phylogeography

The combined tree (Figure 2) rooted with RAD-seq and using *W. mitchelli* and *W. vanduzeei* as outgroups clearly separates into Gulf Coast, NC Lowland, North Carolina mountain, Northeast, and Northwest clades, with the Mid-Atlantic populations being

basal to the NC mountain, Northeast and Northwest clades. Within the Mid-Atlantic grouping, the Pocono Mountain populations (PA1-2) now are strongly supported as distinct from the remainder of the Mid-Atlantic populations and as basal to both the Northeast and Northwest clades (Figs.2- 3).

Within the NC mountain clade (Fig. 2), there is strong support for a distinction between the NC Mountain populations draining into the Savannah River Basin (NC mtn1-2) and those draining into the Tennessee River basin (NC mtn 3-5).

Within the Northwest clade (Figs. 2-3), the populations in southern and western Wisconsin as well as far-western Ontario (WI2-3, ON2) cluster more strongly with northern Manitoba populations (MB1-4) than with populations in northwestern Wisconsin, eastern Ontario, or western Québec (WI1, ON1, QC1-2).

CHAPTER IV

DISCUSSION

Repeatability of the phylogeography

The two phylogenetic trees (Fig. 1) are consistent at their extremes in both the northern (Northwest) and southern (Gulf Coast and NC Lowland) clades. Rooting with either the more conservative COI or with RAD-seq showed the Gulf Coast populations as being basal in the *W. smithii* lineage and more northern populations being progressively derived (Fig. 1). Furthermore, both trees are consistent in showing that the division between the Mid-Atlantic populations and the more southern coastal populations is more ancient than the division between the Mid-Atlantic and the post-glacial populations. The differences between the two trees involve associations in the mid-section of the trees. Most problematic is the lack of agreement in establishing the location of the glacial refugium that ultimately gave rise to the entire northern dispersal of *W. smithii* after recession of the Laurentide Ice Sheet, beginning some 20 Kya. Lesser inconsistencies notwithstanding, we have concluded that, even with a well-distributed sampling protocol including 20 to 25 discrete source populations, RAD-seq will not ensure correct phylogenetic inferences with complete fidelity. We believe, however, that these inconsistencies can be minimized as discussed below.

Consensus phylogeography of Wyeomyia smithii

After combining the two data sets, we constructed a consensus phylogeny for *W. smithii*. The consensus tree resolved all significant discordances that were observed in

comparisons of our two individual RAD-seq trees we used to test for repeatability (Fig. 2).

First, we resolved that the refugium of *W. smithii* during the last glaciation, from which the northern radiation of *W. smithii* occurred, lay near the glacial front, not in the southern Appalachians as we had earlier concluded. The consensus tree (Fig. 2) places the North Carolina mountain clade within, not basal to populations currently residing in Maryland and New Jersey. This conclusion is supported by levels of heterozygosity that remain high in the Gulf Coast, North Carolina coastal, and New Jersey populations, but decline northwards, indicating progressively more derived populations (Armbruster et al. 1998).

Second, present-day, more northern populations of *Sarracenia purpurea* are found in sphagnum peatlands associated with tamarack (*Larix larcina*) and black spruce (*Picea mariana*) (Johnson, 1985) and we have used the co-occurrence of the two tree species as good indicators of pitcher plants while searching for new northern populations of *W. smithii* over the last 40 years. During the last glacial maximum (ca 20-22 Kya), sphagnum-dominated peatlands east of the Appalachian Mountains ranged from northern North Carolina to southern Maryland, followed the glacial retreat northwards 26-18 Kya, westward south of the current Great Lakes 12-14 Kya, and then northwestwards 8-10 Kya (Halsey et al. 2000) approximating the draining of Lake Agassiz (Kleiven et al. 2008). The pattern of post-glacial colonization of tamarack and black spruce followed that of sphagnum peatlands (Halsey et al. 2000). The post-glacial spread of *Sphagnum* peatlands and their associated trees is reflected in the phylogeography of *W. smithii* (Fig. 3). Populations dispersed from Maryland and New Jersey, proceeding through the Pocono

Mountains of Pennsylvania northeastwards following the earlier glacial retreat and then westwards following the later recession of the Laurentide Ice Sheet.

Third, we had proposed that westward migration of *W. smithii* was abetted by the anticyclonic (westward) winds prevailing along the receding glacial front (Muhs and Bettis 2000; Bromwich et al. 2004). This proposition predicts that northern northwestern populations should be more closely related to eastern populations than to southern northwestern populations, i.e., *W. smithii* should exhibit parallel longitudinal zones of relatedness. This prediction is not borne out (Fig. 3). Northeastern and northwestern clades are clearly separate and divergent. Moreover, western Ontario and western Wisconsin populations share a more recent common ancestor with northern Manitoba populations than they do with the population in northeastern Wisconsin, eastern Ontario, or Québec (Figs. 2-3). This pattern is more consistent with the decline in midwestern peatlands ca 8-10 Kya following the draining of Lake Agassiz, their subsequent expansion into the Midwest 4-6 Kya (Halsey and Vitt 2000), and then independent colonization from the east and north.

Informative sites vs. taxon sampling

Given the conclusion that our first RAD-based phylogeny was not as robust as we had assumed, the inevitable question arose: Could we have improved phylogenetic accuracy by increasing the number of phyogenetically informative sites or by increasing the number of populations (taxon sampling) we included in our study? This question is not peculiar to *W. smithii*, but has been the subject of considerable discussion (Havird and Miyamoto 2010; Townsend and Lopez-Giraldez 2010; Nabhan and Sarkar 2011;

Kawahara et al. 2011; Townsend et al. 2012b. In Figure 1, our first phylogenetic tree consists of 3,741 informative sites among 21 populations; our second tree consists of 16,858 informative sites among 25 populations. The former misplaced not only the position of the North Carolina mountain clade, but also nests the Pocono Mountain population (PA1) within and derived from the Maryland-New Jersey populations. The second phylogenetic tree misplaces the Massachusetts (MA) population as basal to the northeast clade and separately clusters the Saskatchewan populations (SK1, 2) with two of the Manitoba populations (MB3, 4). Hence, simply adding more than 13,000 informative sites modified ambiguities, but did not change the fact that we still had two phylogenies that were in substantial disagreement on a number of important points.

It was not until we combined all 46 populations that we arrived at a robust result that resolved discordances between the two individual trees (Figs. 1 vs. 2). The consensus phylogeny is comprised of twice the number of populations of either single tree. We saw above that simply adding more informative sites in approximately the same number of populations did little to improve our results. The important conclusion is that, with the advent of RAD-seq as a valuable tool for inferring phylogeographies, the number of populations sampled is going to be more important in creating a robust result than the addition of more informative sites.

Selection of populations

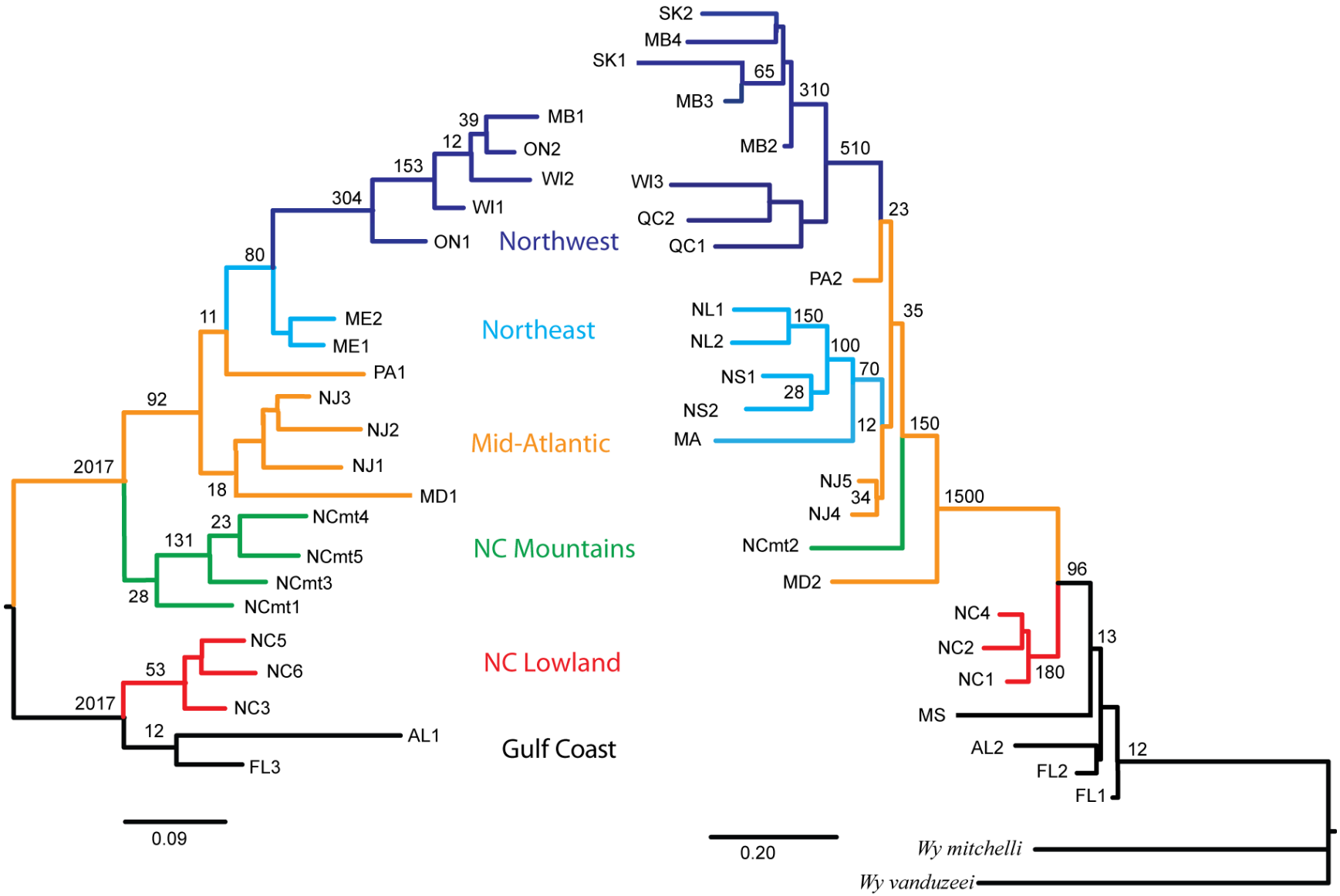
Since both of our independent phylogenies involved separate labor-intensive and expensive field collecting-trips, we have put considerable thought into how we could have stream-lined this process. The time and cost of collecting independent populations

far exceeded the time and cost of RAD-sequencing and the subsequent analyses. Our work with *W. smithii* tells us that we would have achieved a robust, credible phylogeography more efficiently by collecting populations over the species' range while focusing most intensively on the populations likely to have encountered historical or geographic barriers to gene flow. In our case, these barriers were likely the southern boundary of the Laurentide Ice Sheet, Lake Agassiz, and the Great Lakes. The goal would have been to collect from as many populations as possible, even if seemingly excessive at the outset. Since samples can be stored at -80°C indefinitely, a single collecting trip would have provided a library of populations for both current and future projects. We would have made an initial phylogeography based on our knowledge of the organism and its likely geographic history to answer our initial questions. We would then have drawn from our frozen library additional populations from regions of historical complexity, from regions of phylogenetic uncertainty, or from regions appropriate for answering new questions that may have arisen in the interim.

APPENDIX

FIGURES

Figure 1 (next page). Replication of phylogenetic relationships using RAD-seq. **Left**, phylogenetic tree from Emerson et al. (2010) using 21 populations and 54 bp reads, generating 3,741 informative sites. The tree was rooted with mitochondrial COI. **Right**, phylogenetic tree from this study using 25 populations, 80 bp reads, generating 16,858 informative sites. The tree was rooted with RAD-seq with the long terminal branches leading to *W. mitchelli* and *W. vanduzeii* abbreviated to clarify presentation. Color code indicates region of geographic origin. Topologies and branch lengths are based on maximum likelihood (PhyML). Node supports are given as rounded aLRT scores for nodes with aLRT score of at least 10. Bayesian and maximum parsimony support for the left tree are provided in Emerson et al. (2010) and for the right tree are provided in Figure S1. Two-letter abbreviations identify each state or province. Where two or more sites come from the same state or province, they are identified by number.


SK2
MB4
SK1
65
MB3
310
MB2
510
39
MB1
12
ON2
153
WI2
304
WI1
ON1
WI3
QC2
QC1
23
80
PA2
Northwest
11
ME2
ME1
Northeast
NL1
150
NL2
100
70
NS1
28
NS2
12
MA
35
PA1
92
NJ3
NJ2
Mid-Atlantic
2017
NJ1
18
MD1
NJ5
34
NJ4
150
1500
NCmt4
23
131
NCmt5
NCmt3
NC Mountains
NCmt2
28
NCmt1
MD2
96
NC4
NC5
NC2
13
53
NC6
NC1
180
NC Lowland
2017
NC3
MS
AL1
AL2
12
12
Gulf Coast
FL2
FL3
FL1
0.09
0.20
Wy mitchelli
Wy vanduzeei

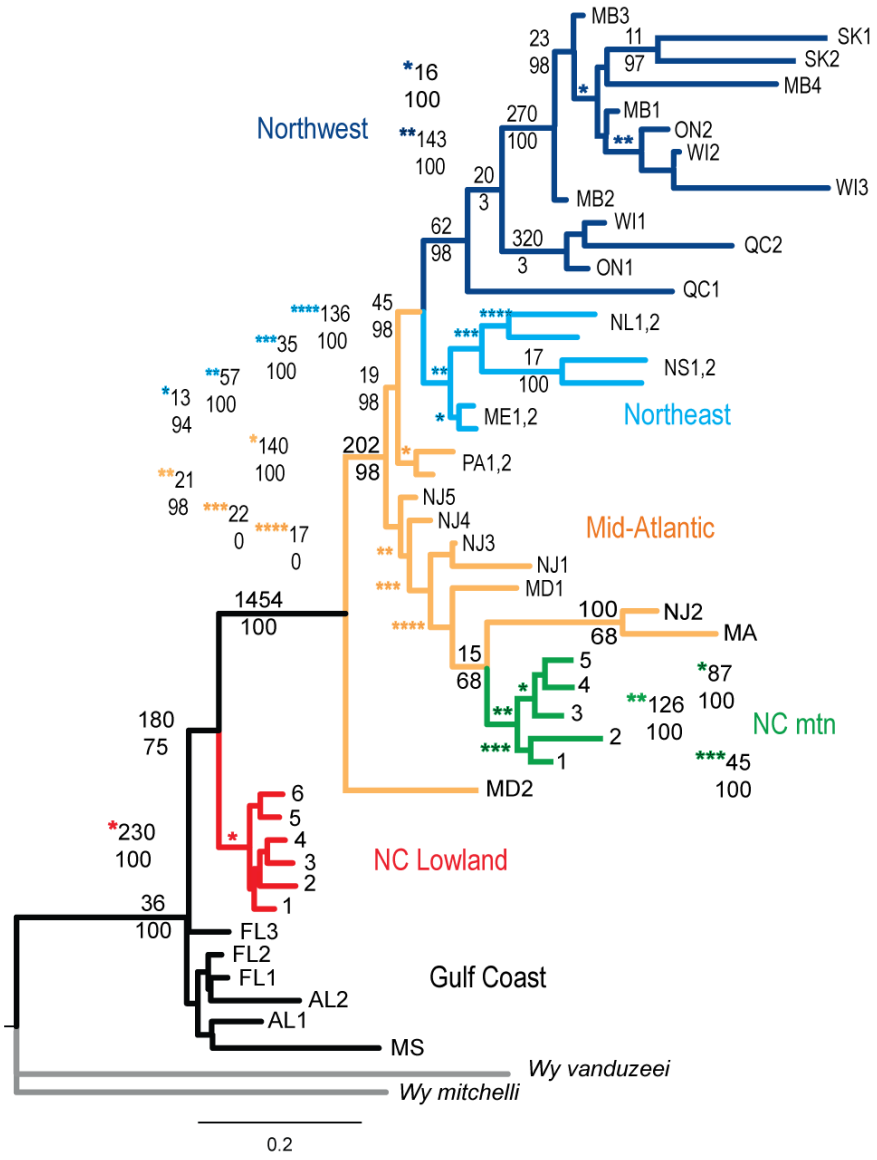


Figure 2. Maximum-likelihood phylogenetic tree for all 46 populations. The tree is based on 54 bp reads, generating 18,680 phylogenetically informative sites. Node support is shown for aLRT values ≥10 (upper value) with their corresponding Bayesian support (lower value). Note that the asterisks are used to connect aLRT and Bayesian support with specific nodes. The corresponding maximum parsimony tree is provided in Figure S2. Color codes are the same as in Figure 1.

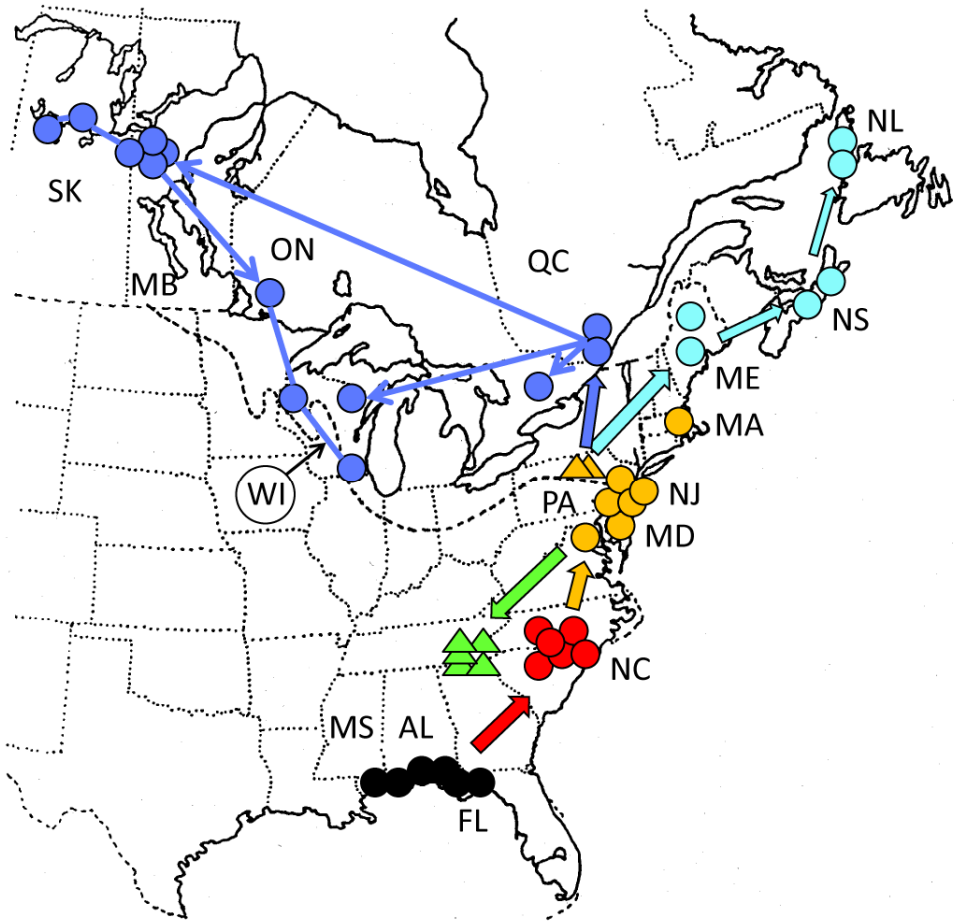


Figure 3. Phylogeography of *Wyeomyia smithii* based on the combined 46-population tree. Arrows indicate likely direction of expansion based on phylogeny in Figure 2. Maximum extent of the Laurentide Ice Sheet at the last glacial maximum is plotted as a dotted line (Colgan 2003). Two-letter abbreviations identify each state or province. Color codes are the same as in Figures 1-2.

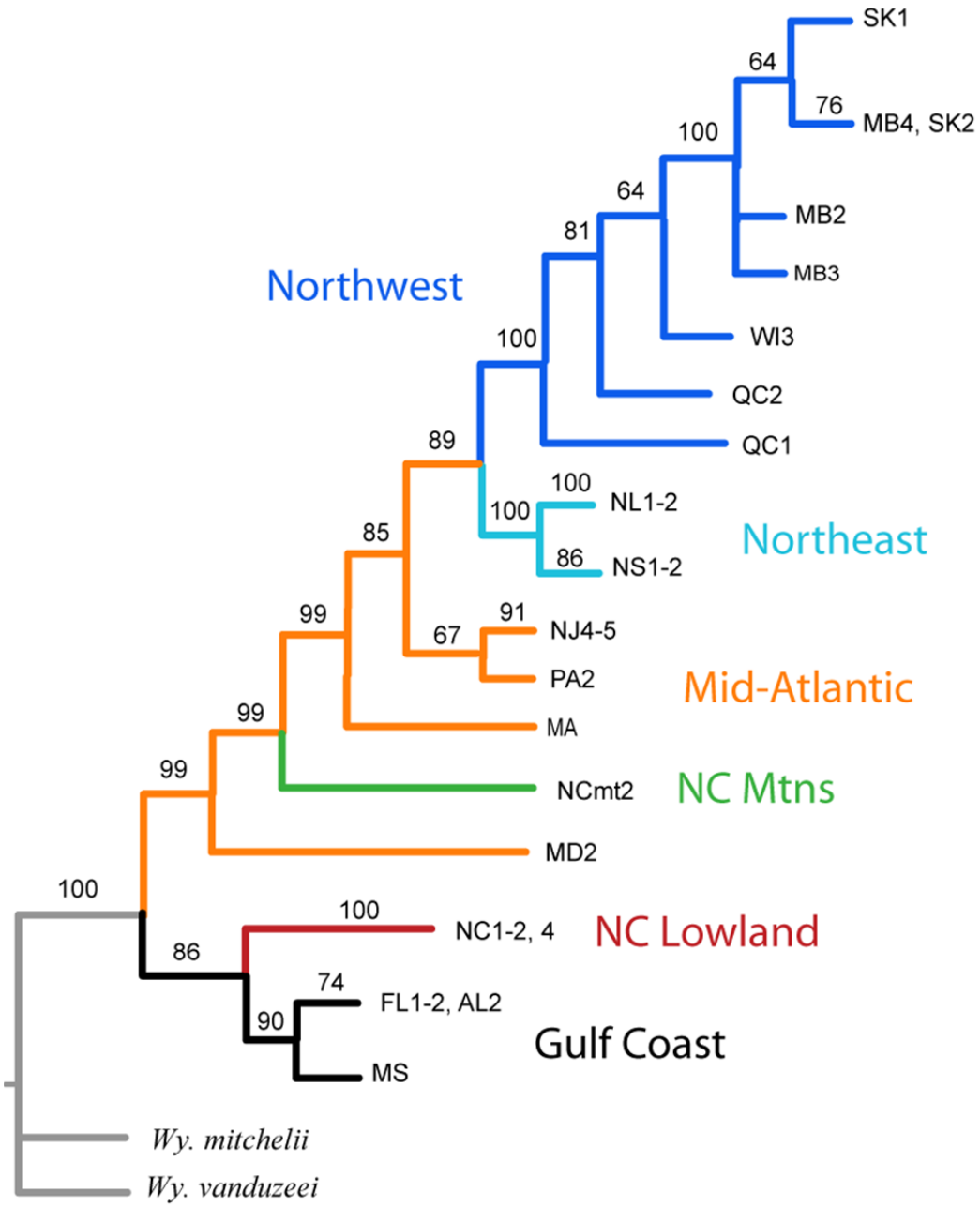


Figure S1a. Maximum parsimony bootstrap values for all resolved nodes > 50 in the 80bp tree

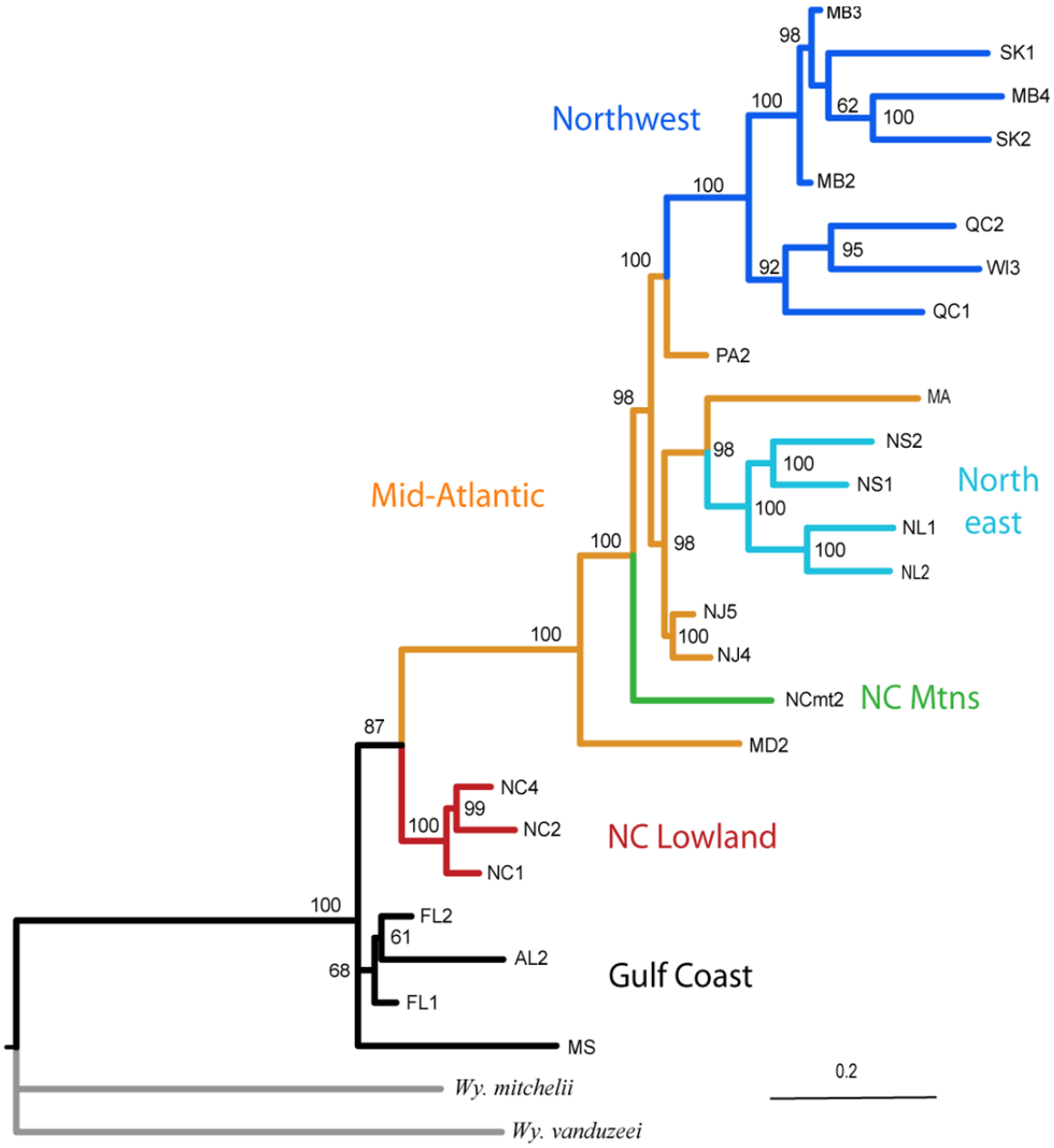


Figure S1b. Empirical Bayesian posterior probabilities for the 80 bp tree

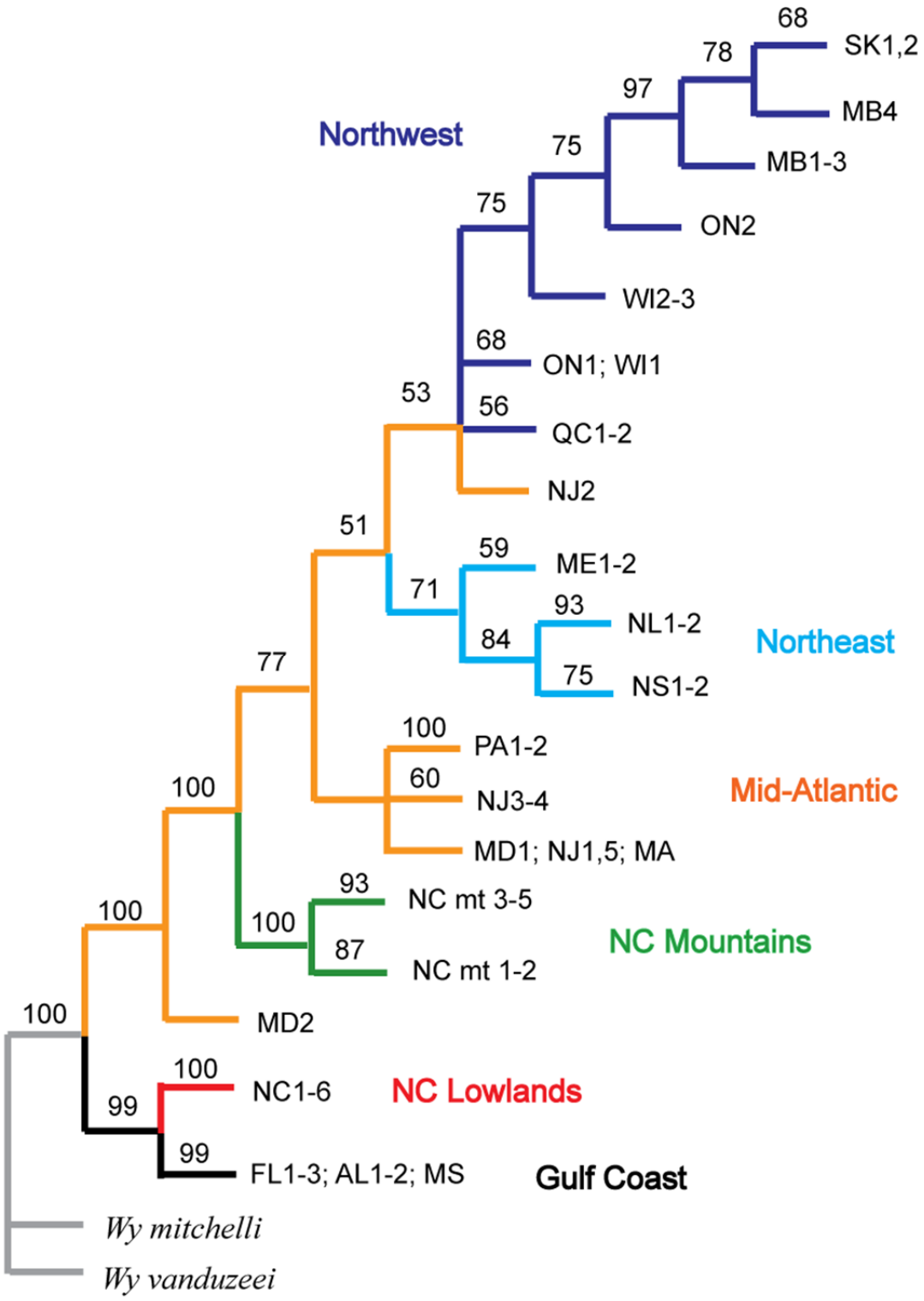


Figure S2. Maximum parsimony bootstrap values for all resolved nodes > 50 in the combined tree.

REFERENCES CITED

Amores A, Catchen J, Ferrera A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics*, **188**, 799-808.

Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Systematic Biology,* **55**, 539-552.

Armbruster, P, Bradshaw WE, Holzapfel CM. 1998. Effects of postglacial range expansion on allozyme and quantitative genetic variation of the pitcher-plant mosquito, *Wyeomyia smithii*. *Evolution*, **52**, 1697-1704.

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

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, Jiggins CD, Blaxter ML (2011) Linkage mapping and comparative genomics using next-generation RAD Sequencing of a Non-Model Organism. *PLoS ONE*, **6**, e19315.

Bromwich DH, Toracinta ER, Wei H, Oglesby RJ, Fastook JL, Hughes TJ (2004) Polar MM5 Simulations of the winter climate of the Laurentide ice sheet at the LGM. *Journal of Climate*, **17**, 3415–3433.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) *Stacks*: building and genotyping loci *de novo* from short-read sequences. *G3* **1**, 71-182.

Colgan PM, Mickelson DM, Cutler PM (2003) Ice-marginal terrestrial landsystems: southern Laurentide Ice Sheet. In: *Glacial Landsystems* (eds Evans DA, Rea BR), pp. 111-142. Edwin Arnold, London.

Cronn R, B. J. Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J (2012) Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany*, **99**, 291-311.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Baxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499-510.

Dyke AS, Andrews JT, Clark PU, England JH, Miller GH, Shaw J, Veillette JJ (2002) The Laurentide and Innuitian ice sheets during the last glacial maximum. *Quaternary Science Reviews*, **21**, 9-31.

Emerson KJ, C.R. Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16196-16200.

Etter PD, Bassham S, Hohenlohe PA, Johnson A, Cresko WA (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. *Methods in Molecular Biology,* **772**, 157-178.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**, 696-704.

Halsey LA, Vitt DH, Gignac L (2000) *Sphagnum*-dominated peatlands in North America since the last glacial maximum: their occurrence and extent. *The Bryologist*, **103**, 334-352.

Havird JC, Miyamoto MM (2010) The importance of taxon sampling in genomic studies: An example from the cyclooxygenases of teleost fishes. *Molecular Phylogenetics and Evolution*, **56**, 451-455.

Hohenlohe PA, Bassham S, Etter PD, Stiffler S, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Hovenkamp P (2006) Can taxon-sampling effects be minimized by using branch supports? *Cladistics*, **22**, 264-275.

Kawahara AY, Oshima I, Kawakita A, Regier JC, Mitter C, Cummings MP, Davis DR, Wagner DL, De Prins J, Lopez-Vaamonde C (2011) Increased gene sampling strengthens support for higher –level groups within laf-mining moths and relatives (Lepidoptera: Gracillariidae). *BMC Evolutionary Biology*, **11**, 182.

Kleiven HF, Kissel C, Laj C, Ninnemann US, Richeter TO, Cortijo E (2008) Reduced North Atlantic deep water coeval with the glacial Lake Agassiz freshwater outburst. *Science*, **319**, 60-64.

Johnson CW (1985) *Bogs of the Northeast*. University Press of New England, Hanover and London.

Lemmon AR, Lemmon EM (2012) High-throughput identification of informative loci for shallow-scale phylogenies and phylogeography. *Systematic Biology* (in press: doi 10 1093/sysbio/sys051).

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2012a) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* (In press. doi:10.1016/j.ympev.2011.12.007).

McCormack JE, Maley JM, Hird SM, Derryberry EP, Graves GR, Brumfield RT. (2012b) Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. *Molecular Phylogenetics and Evolution,* **62**, 397-406.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Muhs DR, Bettis EA (2000) Geochemical variations in Peoria Loess of western Iowa indicate paleowinds of midcontinental North America during last glaciations. *Quaternary Research*, **53**, 49-61.

Nabham, AR, Sarkar IN (2011) The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in Bioinformatics,* **13**, 122-134.

Posada D (2008) jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253-1256.

Posada D, Crandall, KA (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics,* **14**, 817-818.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics,* **19**, 1572-1574.

Rubin B E R, Ree RH, Moreeau CS (2012) Inferring phylogenies from RAD sequence data. *PLoS ONE* **7**, e33394.

Sukumaran J, Holder MT (2010) DendroPy: A Python library for phylogenetic computing. *Bioinformatics,* **26**, 1569-1571.

Swofford DL (2002) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.* Sinauer Associates, Sunderland, Massachusetts.

Townsend JP, Lopez-Giraldez F (2010) Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. *Systematic Biology*, **59**, 446-457.

Townsend JP, Su Z, Tekle YI (2012) Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. *Systematic Biology* (in press, doi: 10.1093/sysbio/sys036).

Tyford AD, Ennos RA (2012) Next-generation hybridization and introgression. *Heredity* **108**, 179-189.

Yokoyama Y, Lambeck K, De Deckker P, Johnston P, Fifield LK (2000) Timing of Last Glacial Maximum from observed sea-level minima. *Nature*, **406**,713-716.

Zellmer AJ, Hanes MM, Hird SM, Carstens BC. 2012. Deep phylogenetic structure and environmental differentiation in the carnivorous plant *Sarracenia allata*. *Systematic Biology* (in press: doi 10 1093/sysbio/sys048)