Resolving postglacial phylogeography using high-throughput sequencing

Kevin J. Emerson[1], Clayton R. Merz, Julian M. Catchen, Paul A. Hohenlohe, William A. Cresko, William E. Bradshaw, and Christina M. Holzapfel

Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, OR 97403-5289

Edited by David L. Denlinger, Ohio State University, Columbus, OH, and approved August 4, 2010 (received for review May 11, 2010)

The distinction between model and nonmodel organisms is becoming increasingly blurred. High-throughput, second-generation sequencing approaches are being applied to organisms based on their interesting ecological, physiological, developmental, or evolutionary properties and not on the depth of genetic information available for them. Here, we illustrate this point using a low-cost, efficient technique to determine the fine-scale phylogenetic relationships among recently diverged populations in a species. This application of restriction site-associated DNA tags (RAD tags) reveals previously unresolved genetic structure and direction of evolution in the pitcher plant mosquito, *Wyeomyia smithii*, from a southern Appalachian Mountain refugium following recession of the Laurentide Ice Sheet at 22,000–19,000 B.P. The RAD tag method can be used to identify detailed patterns of phylogeography in any organism regardless of existing genomic data, and, more broadly, to identify incipient speciation and genome-wide variation in natural populations in general.

genomics | restriction site-associated DNA tag | second-generation sequencing | *Wyeomyia smithii*

The increased accessibility of high-throughput, second-generation sequencing is closing the gap between what have been identified traditionally as model and nonmodel organisms for genetic studies (1, 2). This technology has led to the sequencing of large numbers of nearly complete transcriptomes (3), high-resolution genetic mapping (4), population genetic studies (5), and phylogeographic analysis of methicillin-resistant bacteria based on whole-genome sequences (6). For many species, however, such genomic resources are modest or not available, and fine-scale resolution of population dynamics or phylogeographies has been based on mitochondrial or chloroplast DNA, microsatellites, or a combination of these approaches (7). Here, we present a cost- and labor-effective method to resolve the postglacial phylogeography of a nonmodel organism, the North American pitcher plant mosquito, *Wyeomyia smithii* (Coq.). We use high-throughput sequencing of restriction-site–associated DNA tags (RAD tags) (4, 8) to identify 3,741 single nucleotide polymorphisms (SNPs) throughout *W. smithii*'s 836-Mb genome (9) that are fixed within and variable among populations. The phylogeography based on these SNPs indicates that following recession of the Laurentide Ice Sheet, refugial populations of *W. smithii* dispersed northward and then westward across North America. This level of postglacial phylogeographic resolution was not achieved with either allozymes (10) or mitochondrial *cytochrome oxidase subunit I* (*COI*) sequences. This RAD tag approach can easily be extended to other organisms, regardless of their available genomic resources.

Phylogeography, the study of the geographic distributions of genetic variation, originally was developed using single gene or tightly linked (mitochondrial) genetic markers (11). These approaches can be very costly, labor intensive, and, in cases of recent population differentiation such as postglacial range expansion, may fail to supply sufficient resolution to infer patterns of population relatedness with high degrees of certainty. More recent EST-based approaches and whole-genome phylogenomics are able to provide a greater number of markers for phylogenetic analysis (3, 12). However, these studies typically involve comparisons above the species level or, in the case of bacteria, whole-genome sequences (6). Hence, these approaches are not useful for determining the genetic similarity in closely related populations of nonmodel species or species for which whole-genome resequencing is not yet possible.

Phylogeography and phylogenetics have recently benefited from genome-wide SNP detection methods to elucidate patterns of variation in model taxa or their close relatives (7, 13). The limiting step in the applicability of multilocus datasets in nonmodel organisms has been generating the genetic markers to be used (14). Baird et al. (4) developed a second generation sequencing approach that allowed for the simultaneous discovery and typing of thousands of SNPs throughout the genome (5). This sequenced RAD tag technique is a general approach that does not require the prior development of any genomic resources for the study organism. Here, we show the power and efficiency of RAD tag genotyping to resolve differences among closely related populations in a nonmodel organism using the pitcher plant mosquito, *W. smithii*, the first animal to have shown an evolutionary (genetic) response to rapid climate change (15).

W. smithii is the single temperate species from a large neotropical genus (16, 17) and displays a geographical distribution that closely follows that of its host plant, *Sarracenia purpurea*, from the Gulf of Mexico north to Canada and from Labrador west to Alberta (10, 18). *W. smithii* separates into two geographic groups, a southern group at low elevations in the southeastern coastal plain from Mississippi to North Carolina and a northern group including populations in the southern Appalachian Mountains and populations extending from Maryland northward to Labrador and westward to northern Alberta. These two groups are fully interfertile and form a single species (18, 19) but can be distinguished by morphological, behavioral, physiological, and reproductive characters (18, 20, 21). This same suite of characters and allozymes (10) fails to discriminate relatedness of populations within the northern group and therefore leaves unresolved the patterns of postglacial range expansion following recession of the Laurentide Ice Sheet beginning 22,000–19,000 B.P (22). Here, we use the mitochondrial gene *COI* and two other *Wyeomyia* species as outgroups to root the *W. smithii* tree. We then use RAD tag technology to isolate SNPs that are fixed within populations and variable among populations (Fig. 1) to determine the phylogeographic history associated with the postglacial range expansion of *W. smithii*.

Author contributions: K.J.E., W.E.B., and C.M.H. designed research; K.J.E. and C.R.M. performed research; J.M.C., P.A.H., and W.A.C. contributed new analytic tools; K.J.E., J.M.C., P.A.H., W.E.B., and C.M.H. analyzed data; and K.J.E., C.R.M., J.M.C., P.A.H., W.A.C., W.E.B., and C.M.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. HM136806–HM136827). Short reads can be found in Genbank's sequence read archive (accession nos. SRA012678 and SRP002409).

[1]To whom correspondence should be addressed. E-mail: kemerson@uoregon.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1006538107/-/DCSupplemental.

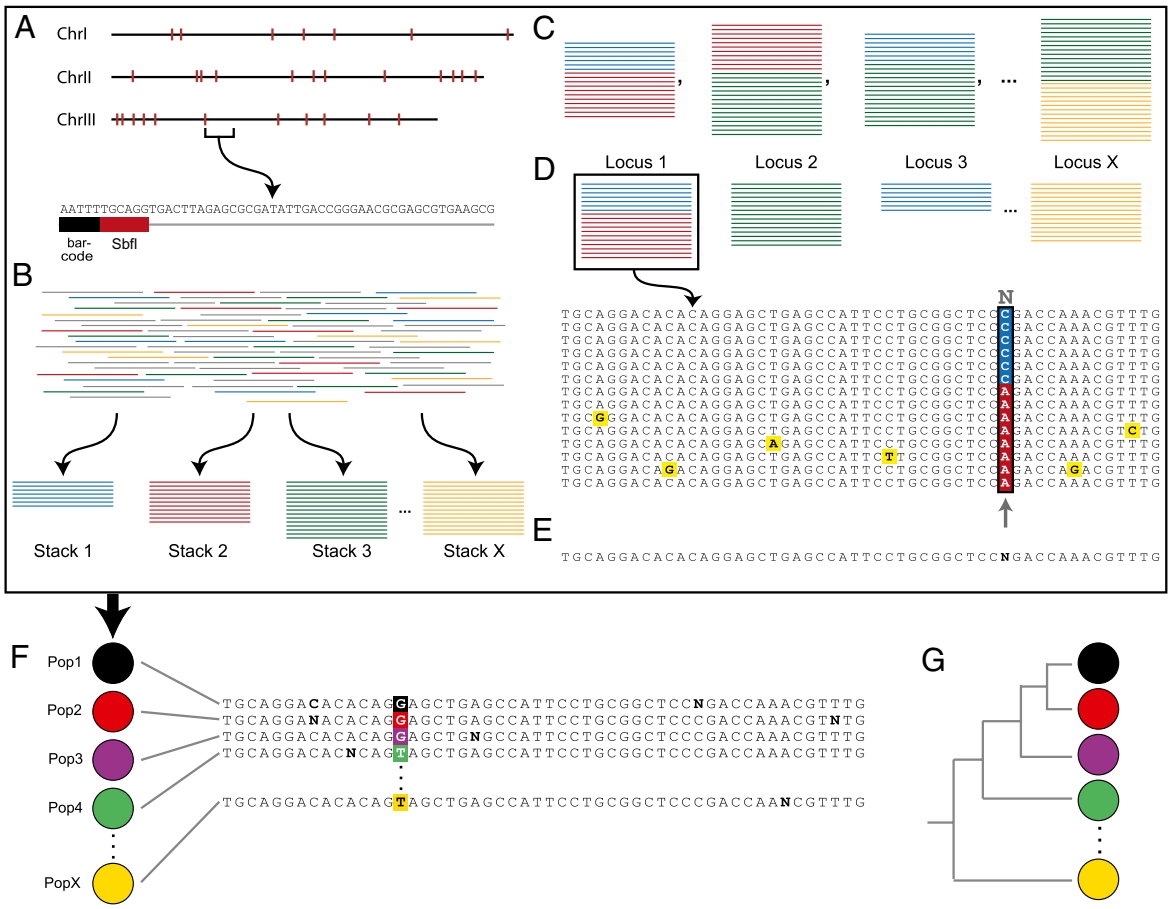

Fig. 1. *In silico* RAD tag genotyping within (*A*–*E*), and SNP discovery between (*F* and *G*), populations. (*A*) *W. smithii* has three nuclear chromosomes, each of which contains multiple SbfI cut sites (red marks). The genomic DNA is digested, barcoded with a population-specific sequence, and amplified, resulting in multiple sequence reads from each of the RAD tag sites in the genome. Each sequence consists of a population-specific 5-bp barcode (black), the enzyme-recognition sequence (red), and the downstream sequence. (*B*) The de novo RAD tag pipeline compares all the sequenced reads and builds stacks of exactly matching tags. (*C*) Pairwise comparisons are made between all stacks, i.e., blue vs. red, red vs. green, blue vs. green, and so on. (*D*) Loci were defined as a set of stacks such that for each stack, there is another stack in the locus that is at most one nucleotide divergent. Each locus is then examined one nucleotide position at a time. If the nucleotide at that position is at a significantly high frequency within the population, it is considered to be the consensus nucleotide; if not, it is replaced with an N, resulting in (*E*) the consensus sequence for that RAD tag site within the population. (*F*) This process is repeated for each of the populations. (*G*) The resulting RAD tag consensus sequences are then used for phylogenetic analysis.

Results

Rooting the Tree with *COI*. A 1,176-bp segment of *COI* from 20 populations of *W. smithii* and two outgroup species, *Wyeomyia mitchellii* and *Wyeomyia vanduzeei* (GenBank accession nos. HM136806–HM136827) was used to infer the phylogenetic relatedness of the populations using maximum parsimony, maximum likelihood, and topological empirical Bayes Markov chain Monte Carlo methods (23). The alignment included 167 variable sites. The three methods of phylogenetic inference agreed on the overall topology of the phylogenetic tree. There was overwhelming evidence that the root of *W. smithii* separated the southern from the northern groups (Fig. 2). The phylogenetic relationship among populations within both groups was not resolved by the *COI* sequences.

RAD Tag SNP-Based Tree. RAD tag libraries were created by individually barcoding and sequencing DNA from pools of six individuals from each of 21 different populations at SbfI cut sites throughout the genome. Two lanes of sequencing on an Illumina GAIIX resulted in a total of more than 27.5 million RAD tag sequences of which more than 14.9 million sequences passed several levels of quality filtering (*Methods*). The 21 populations of *W. smithii* were represented by an average of 711,702 ± 85,779 SE sequences (Fig. S1). Within each population, we identified an average of 20,868 ± 1,681 SE stacks (Fig. 1) spread across 13,627 ± 1,177 SE loci, resulting in an average of 1.53 stacks per RAD tag locus (Fig. S2). All raw sequence reads are available at the National Center for Biotechnology Information Short Read Archive (accession nos. SRA012678 and SRP002409). Throughout *W. smithii*'s genome, we identified 3,741 SNPs within the RAD tag sequences that were fixed within at least two populations and were variable among populations.

The RAD tag SNP dataset resolved four major clades in *W. smithii* (Fig. 3) within the two broad groups identified above. The southern group included two clades, one from along the Gulf Coast (black) and the other from the North Carolina coast (NC Coast, red). The northern group resolved into two major clades, a southern Appalachian Mountain clade (Appalachian, purple) and a northern clade (green and blue). Within the northern clade, there was consistent, clear resolution (node support ≥ 91) between the Maine and central Ontario populations, between the central Ontario and northern Wisconsin populations, and between the western Ontario and northern Manitoba populations. These results confirm the ancient divergence of the southern from the northern group and a recent, postglacial divergence among the two clades within the northern group. The RAD tag approach provided fine-scale resolution in the northern group and revealed sequential evolution along geographic gradients in postglacial populations.

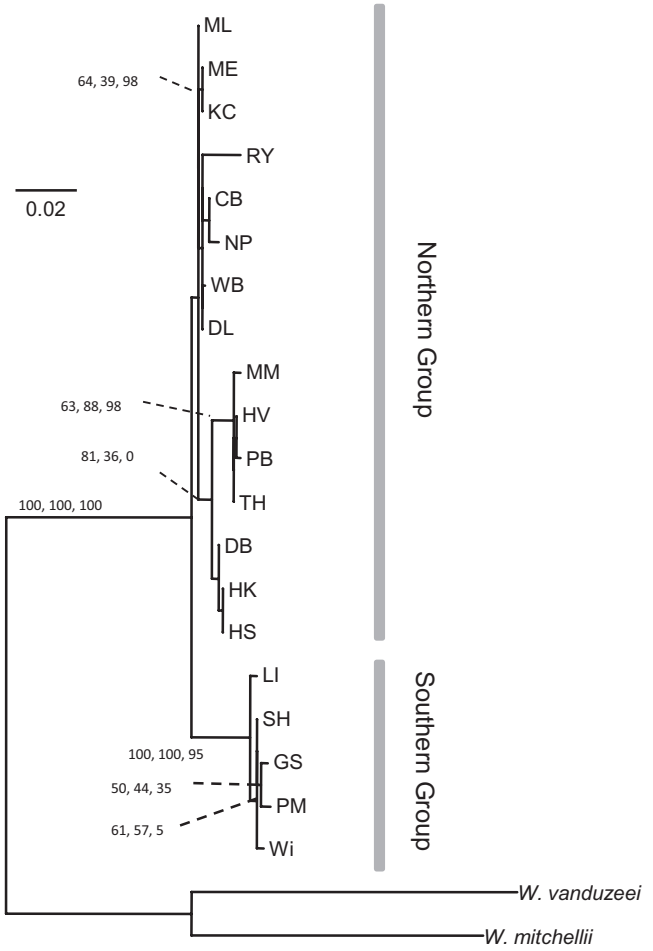

Fig. 2. Maximum likelihood phylogenetic tree of *W. smithii* populations based on *COI* sequences with *W. vanduzeei* and *W. mitchellii* as outgroups. Node support from the maximum parsimony bootstrap value (%), the approximate likelihood ratio test support, and empirical Bayes posterior probabilities (expressed as %), respectively, are shown for nodes that have at least one support value greater than 50. Bars delineate the southern and northern groups. Population names are those used in previous publications and are defined by their latitude, longitude, altitude, and state or province of origin as given for each locality in Table S1.

Discussion

Phylogeography of *W. smithii*. The RAD tag approach confirms the phylogeographic separation of *W. smithii* into southern and northern groups (Fig. 3). Importantly, this approach exposes a phylogeny that closely parallels the geographic distribution of *W. smithii* in North America. Within the southern group, genetic differences among populations of *W. smithii* determined using the RAD tag approach distinguished a Gulf Coast from a North Carolina coast clade. *W. smithii's* host plant, *S. purpurea*, is sparsely distributed in the coastal plain of Georgia and does not extend into the Florida Peninsula (24), retarding migration between the Gulf and North Carolina coasts. This distinction of Gulf Coast and North Carolina coast clades found in *W. smithii* is characteristic of numerous vertebrate and invertebrate species (25, 26).

Within the northern group, genetic differences among populations of *W. smithii* determined using the RAD tags distinguished an Appalachian and a northern clade (Fig. 3). The northern clade includes populations that have become established since recession of the Laurentide Ice Sheet over the last ~20,000 y. The northern clade shares a more recent common ancestor with the Appalachian clade than it does with either clade in the southern group. The distribution of *W. smithii*'s host plant, *S. purpurea*, and prevailing winds provide clues to the origin of populations within the northern group. First, natural populations of *S. purpurea* in unglaciated North America do not extend west of the Appalachian Mountains (24). Second, *W. smithii* are active as winged adults during the spring, summer, and fall when the prevailing winds in North America blow from southwest to northeast (27). The most likely source of the more northern populations is from more southern localities, i.e., the North Carolina coast or Appalachian clades (Fig. 3, red or purple nodes). The northern clade (blue node) shares a more recent common ancestor with the Appalachian clade (purple node) than it does with the North Carolina coast clade (red node). Finally, the climate in the southern Appalachians during the last glacial maximum was similar to that in present-day southern Manitoba (28), well within the current range of *W. smithii* and its host plant. We therefore conclude that the northern clade originated from glacial refugia in the southern Appalachian Mountains.

The Laurentide Ice Sheet retreated first along the eastern seaboard and thereafter northwestwards through the northern United States and Canada (29). At the same time, the retreating glacial anticyclone was generating easterly or northeasterly winds over the ice sheet and the area just south of it (30). Anticyclonic northeastern winds at the eastern glacial front initially would have inhibited northward migration of *W. smithii*. As the glaciers retreated northwestwards, however, the effect of the anticyclonic winds would have diminished along the eastern seaboard, permitting the southwest winds to prevail and allowing the northward expansion of *W. smithii*. At the same time, the northeast and eastern anticyclic winds at the retreating glacial front would have facilitated westward migration of the northern populations. Postglacial relationships within the northern clade of *W. smithii* support this wind-generated pattern because populations show increasing divergence toward Maine and then westward in the northern United States and Canada (Fig. 3).

RAD Tags for Phylogenetic Inference. Without any prior investment in genomic resources, high-throughput sequencing of RAD tags enabled us to resolve fine-scale genetic divergence among intraspecific populations that have been separated for less than 20,000 y. Had we invested more time and sequenced multiple individuals within as well as between populations, the complete *COI* gene might have provided a similar degree of resolution but then the tree would still have been based on variation in a single gene. RAD tag sequencing, even with conservative data filtration, produced 3,741 SNPs distributed among 13,627 loci in the nuclear genome, all within two lanes in a single run of an Illumina sequencer. This approach is generally applicable to traditional nonmodel organisms for elucidating population structure within a species. Here, we used a single restriction enzyme for identifying phylogenetic relationships among closely related populations within a single species. Even greater phylogenetic resolution could be achieved easily with this methodology by using additional restriction enzymes or enzymes requiring a shorter recognition sequence.

Phylogeographic analysis based on traditional genetic markers, such as microsatellites or mitochondrial sequences, assumes that these loci approximate neutral (i.e., genome-wide average) demographic and phylogeographic processes. However, the dozens to scores of markers typically assayed introduce sampling variation. A small number of outlier loci that are subject to selection or influenced by genetic hitchhiking can greatly influence the branch lengths and topology of inferred phylogenies (31) as well as estimates of population differentiation (14). One goal of a genomic approach such as RAD tag sequencing is to overwhelm sampling error and reduce the effect of outlier loci by providing a much denser genome-wide sample of genotype data, thus providing a more precise estimate of actual phylogeographic relationships. In the case of *W. smithii*, these phylogenetic relationships will allow us to determine the genetic patterns underlying the evolution of photoperiodic response relevant to rapid climate change as well as other

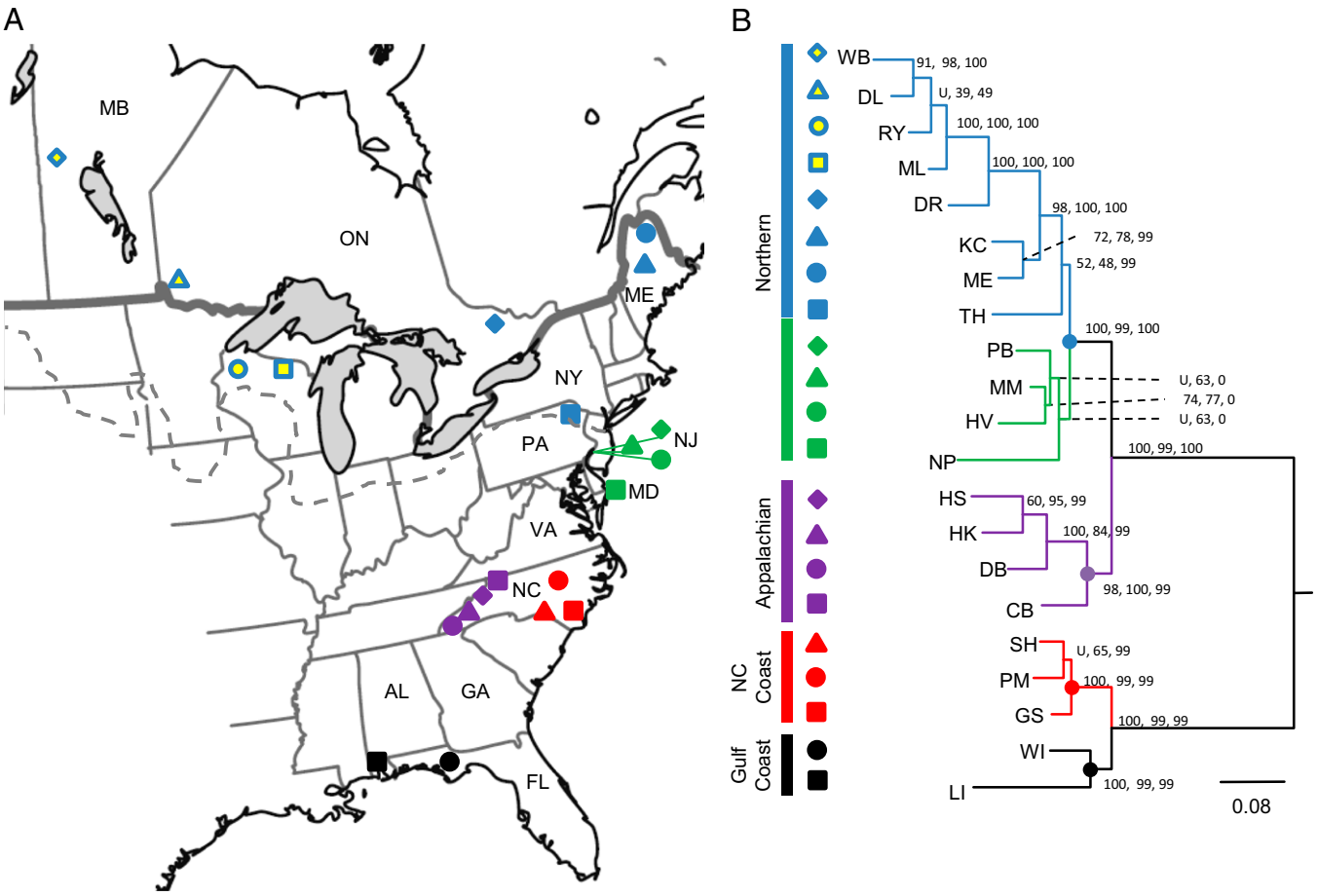

Fig. 3. (*A*) Map of eastern North America showing the geographic range of *W. smithii* used in this study. Dashed line shows the maximum extent of the Laurentide Ice Sheet at ~20,000 B.P. State and province abbreviations used: MB, Manitoba; ON, Ontario; ME, Maine; NY, New York; PA, Pennsylvania; NJ, New Jersey; MD, Maryland; VA, Virginia; NC, North Carolina; GA, Georgia; AL, Alabama; FL, Florida. (*B*) Maximum likelihood phylogenetic tree of *W. smithii* using the RAD tag SNPs. Node support is given as the maximum parsimony bootstrap value (U, unresolved node), the maximum likelihood approximate likelihood ratio test support, and the empirical Bayes posterior probabilities. The tree was rooted with the *COI* data presented in Fig. 2. Colored bars identify the four major clades within *W. smithii*. Population names are those used in previous publications and are defined by their latitude, longitude, altitude, and state or province of origin as given for each locality in Table S1.

physiological, morphological, developmental, and reproductive processes that have taken place over millennial time scales.

Methods

Amplification and Sequencing of *COI*. A 1,237-bp segment of *COI* was amplified from a single individual from each of 20 populations of *W. smithii* (Table S1) and one individual each from *W. mitchellii* and *W. vanduzeei* (collected from Collier County, FL) using the forward primer GGAGCACCTGATATAGCTTTCCC and the reverse primer TTCATTGCACTAATCTGCCATATTA. Alignment and end-trimming resulted in a 1,176-bp alignment, which corresponds to bases 1,790–3,004 of the *Drosophila melanogaster* mtDNA sequence (GenBank acc no.: U37541).

Creation and Sequencing of the RAD Tag Library. The genomic RAD tag libraries were created as in ref. 5. Genomic DNA was collected from a pool of six individuals from each of 21 separate populations of *W. smithii* (Table S1). Each population sample was digested with high-fidelity SbfI (New England Biolabs) and barcoded with a unique 5-bp sequence (4). All barcodes differed by at least two nucleotides to aid in the quality of *in silico* filtering. Adaptors designed for sequencing were ligated to the samples as in ref. 5. Two libraries were prepared with approximately equal amounts of DNA from 10 and 11 populations (Table S1). Each library was sequenced in one lane of an Illumina GAIIX sequencer.

Inferring Phylogenetically Informative SNPs from Sequenced RAD Tags. Sequence reads were assigned to each population based on a population-specific barcode and three additional quality checks were conducted. First, the beginning of each read was checked for the end of an intact cut site (TGCAGG). RAD tags containing a single sequencing error in the cut site were corrected; those with more errors in the cut site were removed from the dataset. Second, any read containing an uncalled nucleotide base was discarded. Third, the quality scores of each read were examined using a sliding window analysis, keeping reads with isolated, low-quality nucleotides, whereas discarding sequencing reads that show a sustained decrease in quality with sequence length (Fig. S1).

W. smithii does not have a reference genome sequence against which to align sequence reads. Instead, we used a multistep process to identify RAD tag loci within populations, assign a consensus sequence to each population at each RAD tag locus, and align consensus sequences across populations (Fig. 1). The goal was to identify nucleotide positions that are fixed, or nearly fixed, within populations and variable among populations to use as informative SNPs in subsequent phylogenetic analysis.

Within each population, identical reads were aligned together into stacks (Fig. 1*B*). The pairwise sequence divergence among stacks was used to group them into putative loci (Fig. 1*C*). Loci were defined as a set of stacks such that for each stack, there is another stack in the locus that is at most one nucleotide divergent. Lumberjack stacks, stacks containing excessive numbers of sequence reads, can occur when multiple, repetitive sites in the genome are all within a single nucleotide of one another. For this analysis, all stacks with a depth of coverage greater than two SDs above the mean stack depth were removed and the remaining stacks were merged into a locus.

For each nucleotide site in a locus, a likelihood ratio test of the read counts of alternative nucleotides was used to test whether the allele frequency of

the most-observed nucleotide was significantly larger than a threshold $\widehat{p}$. The likelihood of the observed read counts is:

$$L(p) = \frac{n!}{n_1!n_2!n_3!n_4!}\left(p(1-\varepsilon)+\frac{\varepsilon}{4}\right)^{n_1}\left((1-p)(1-\varepsilon)+\frac{\varepsilon}{4}\right)^{n_2}\left(\frac{\varepsilon}{4}\right)^{n_3+n_4} \quad [1]$$

where p is the nucleotide frequency, n_1 is the read count of the most-observed nucleotide, n_2 is the count of the second-most-observed nucleotide, and so on; n is the total read count, and ε is the sequencing error rate. To calculate a global estimate of ε, we used the observed error in the known barcode sequences. Assuming a Poisson error process in the 5-bp barcode sequence, the error rate estimate is $\widehat{\varepsilon} = -(1\mathrm{n}(1\text{-}x)/5)$, where x is the proportion of barcodes with at least one error. This procedure resulted in an estimate of 0.0372 for this dataset, although the estimate of $\widehat{\varepsilon}$ actually has a small effect on the likelihood ratio test below.

We tested whether the likelihood of the observed frequency $P = n_1/(n_1 + n_2)$ was significantly higher than that of a threshold frequency $\widehat{p}$ using the likelihood ratio test statistic:

$$LR = n_1 1\mathrm{n}\left[\frac{4n_1(1-\varepsilon)+(n_1+n_2)\varepsilon}{4\widehat{p}\ (n_1+n_2)(1-\varepsilon)+(n_1+n_2)\varepsilon}\right] + n_2 1\mathrm{n}\left[\frac{4n_2(1-\varepsilon)+(n_1+n_2)\varepsilon}{4(1-\widehat{p}\ )(n_1+n_2)(1-\varepsilon)+(n_1+n_2)\varepsilon}\right] \quad [2]$$

The population consensus sequence was assigned the most-observed nucleotide if both $p > \widehat{p}$ and $2*LR > 3.84$ (significance level $\alpha = 0.05$). If one or both of these conditions was not met, the consensus sequence was given N at this position (Fig. 1 *D* and *E*).

We tested each nucleotide position against a threshold allele frequency of $\widehat{p} = 0.5$. Note that this test does not mean we are considering any nucleotide nearly fixed if its observed frequency is simply greater than this value. Rather, our method combines information on observed frequency and depth of coverage into a single test. For example, at the observed mean, coverage depth of 29 sequence reads per locus, and the error rate above, a nucleotide site with read counts of (n_1,n_2,n_3,n_4) = (20,8,1,0) would be assigned allele 1, whereas (19,9,1,0) would be assigned N. At lower coverage, (6,1,0,0) would be assigned allele 1 and (5,1,0,0) would be assigned N.

Homologous RAD tag loci consensus sequences were then aligned among populations (Fig. 1*F*). Any locus that was present in at least two populations was retained and used in subsequent phylogenetic analysis (Fig. 1*G*).

Phylogenetic Analysis. The resulting two datasets (*COI* and SNP) were analyzed using maximum parsimony, maximum likelihood, and Bayesian methods. Maximum parsimony analysis was conducted using phylogenetic analysis using parsimony (PAUP) (32) with 100 replicates of tree bisection and reconnection branch swapping. Node support was estimated with 200 bootstrap replicates. Maximum likelihood and topological empirical Bayes analyses were conducted using PhyML (33, 34) with Bayesian implementation (23, 35) using approximate likelihood ratio tests scaled with a Shimodaira-Hasegawa–like conversion (36) and posterior probabilities as estimates of node support, respectively. Several criteria for choosing an appropriate model of nucleotide evolution (including Akaike's information criterion and Bayesian information criterion) agreed that a generalized time reversible (GTR) model was appropriate for the SNP dataset and a GTR+I+Γ model was appropriate for the *COI* dataset; all maximum likelihood and Bayesian implementation analyses used these models (37).

ACKNOWLEDGMENTS. We thank Victor Hanson-Smith and Geeta Eick for discussions and assistance with the phylogenetic analysis, John Postlethwait for logistical support, Joe Thornton and two anonymous reviewers for insightful comments, and Prithiviraj Fernando and Lucien Jacky for assistance in generating the *COI* sequences. This research was supported by generous funding through National Science Foundation Grants DEB-0413573, IOB-0445710, DEB-0917827, and IOB-0839998 (to W.E.B.); DEB-0919090 (to W.A.C.); and IOS-0843392 (to P.A.H. and Steven Arnold); and National Institutes of Health Grants 1R24GM079486-01A1 (to W.A.C.) and 3R01RR020833-04S1 (to John Postlethwait).

1. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* 24:133–141.
2. Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Mol Ecol Resour* 8:3–17.
3. Hittinger CT, Johnston M, Tossberg JT, Rokas A (2010) Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. *Proc Natl Acad Sci USA* 107:1476–1481.
4. Baird NA, et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3:e3376.
5. Hohenlohe PA, et al. (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* 6:e1000862.
6. Harris SR, et al. (2010) Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327:469–474.
7. Brito PH, Edwards SV (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica* 135:439–455.
8. Miller MR, et al. (2007) RAD marker microarrays enable rapid mapping of zebrafish mutations. *Genome Biol* 8:R105.
9. Rao PN, Rai KS (1990) Genome evolution in the mosquitoes and other closely related members of superfamily Culicoidea. *Hereditas* 113:139–144.
10. Armbruster PA, Bradshaw WE, Holzapfel CM (1998) Effects of postglacial range expansion on allozyme and quantitative genetic variation in the pitcher-plant mosquito, *Wyeomyia smithii*. *Evolution* 52:1697–1704.
11. Avise JC, et al. (1987) Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annu Rev Ecol Syst* 18:489–522.
12. de la Torre-Bárcena JE, et al. (2009) The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data. *PLoS ONE* 4:e5764.
13. Whitfield CW, et al. (2006) Thrice out of Africa: Ancient and recent expansions of the honey bee, *Apis mellifera*. *Science* 314:642–645.
14. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: From genotyping to genome typing. *Nat Rev Genet* 4:981–994.
15. Bradshaw WE, Holzapfel CM (2008) Genetic response to rapid climate change: It's seasonal timing that matters. *Mol Ecol* 17:157–166.
16. Lane J (1953) *Neotropical Culicidae* (Univ of São Paulo, São Paulo, Brazil).
17. Stone A, Knight KL, Starke H (1959) *A Synoptic Catalog of the Mosquitoes of the World (Diptera: Culicidae)* (Entomological Society of America, Washington, DC).
18. Bradshaw WE, Lounibos LP (1977) Evolution of dormancy and its photoperiodic control in pitcher-plant mosquitoes. *Evolution* 31:546–567.
19. Armbruster P, Bradshaw WE, Holzapfel CM (1997) Evolution of the genetic architecture underlying fitness in the pitcher-plant mosquito, *Wyeomyia smithii*. *Evolution* 51:451–458.
20. Bradshaw WE (1986) Variable iteroparity as a life history tactic in the pitcher plant mosquito *Wyeomyia smithii*. *Evolution* 40:471–478.
21. Lounibos LP, Vandover C, O'Meara GF (1982) Fecundity, autogeny, and the larval environment of the pitcher-plant mosquito, *Wyeomyia smithii*. *Oecologia* 55:160–164.
22. Yokoyama Y, Lambeck K, De Deckker P, Johnston P, Fifield LK (2000) Timing of the Last Glacial Maximum from observed sea-level minima. *Nature* 406:713–716.
23. Kolaczkowski B, Thornton JW (2008) A mixed branch length model of heterotachy improves phylogenetic accuracy. *Mol Biol Evol* 25:1054–1066.
24. McDaniel S (1971) *The Genus Sarracenia (Sarraceniaceae)*. *Bulletin of the Tall Timbers Research Station No. 9*. (Tall Timbers Research Station, Tallahassee, FL).
25. Avise JC (2000) *Phylogeography: The History and Formation of Species* (Harvard Univ Press, Cambridge, MA).
26. Soltis DE, Morris AB, McLachlan JS, Manos PS, Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Mol Ecol* 15:4261–4293.
27. Baldwin JL (1968) *Climatic Atlas of the United States* (Department of Commerce, Environmental Science Services Administration, Environmental Data Service, Washington, D.C.).
28. Jackson ST, et al. (2000) Vegetation and environment in Eastern North America during the Last Glacial Maximum. *Quat Sci Rev* 19:489–508.
29. Šibrava V, Bowen DQ, Richmond GM, eds (1986) *Quaternary Glaciations in the Northern Hemisphere* (Pergamon Press, Oxford).
30. Muhs DR, Bettis EA (2000) Geochemical variations in Peoria Loess of western Iowa indicate paleowinds of midcontinental North America during last glaciation. *Quat Res* 53:49–61.
31. Landry PA, Koskinen MT, Primmer CR (2002) Deriving evolutionary relationships among populations using microsatellites and (deltamu)(2): All loci are equal, but some are more equal than others. *Genetics* 161:1339–1347.
32. Swofford D (2003) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)* (Sinauer Associates, Sunderland, MA).
33. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
34. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* 59:307–321.
35. Kolaczkowski B, Thornton JW (2009) Long-branch attraction bias and inconsistency in Bayesian phylogenetics. *PLoS ONE* 4:e7891.
36. Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol* 55:539–552.
37. Posada D (2008) jModelTest: Phylogenetic model averaging. *Mol Biol Evol* 25:1253–1256.

Supporting Information

Emerson et al. 10.1073/pnas.1006538107

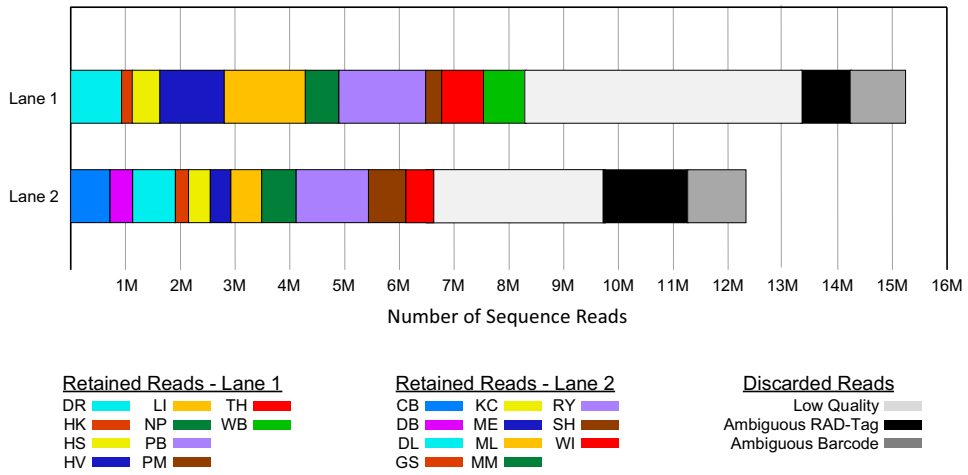

Fig. S1. Distribution of sequence reads among populations of *W. smithii*. Two lanes of an Illumina GAIIX were used to generate the sequence data, each containing a RAD tag library from 10 to 11 populations, identifiable by a 5-bp barcode (*Methods*). Reads were discarded before assembly steps for three reasons: (*i*) low quality—low overall quality scores for a large portion of the sequence, (*ii*) ambiguous RAD tag—at least two low-quality base pairs in the Sbf1 cut site, or (*iii*) ambiguous barcode—at least two low-quality base pairs in the population-specific barcode. Barcodes were separated by at least two mismatches, making the probability of incorrect population assignment very low. Population names are those used in previous publications and are defined by their latitude, longitude, altitude, and state or province of origin as given for each population in Table S1.

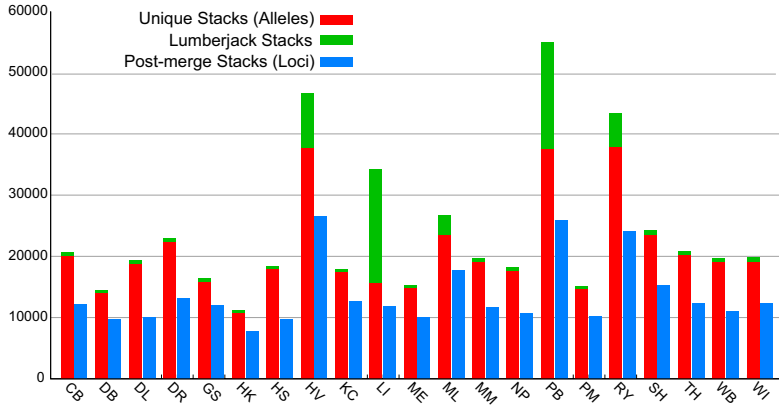

Fig. S2. Frequency distribution of alleles and loci for each population in the *W. smithii* RAD tag dataset. The first step of the RAD tag analysis was to identify all stacks of exactly matching sequences (Fig. 1*B*); these stacks are then merged into loci—sets of stacks such that for each stack, there is another member of the set that differs by at most one nucleotide. Lumberjack stacks can occur when duplicate regions in the genome are within a single nucleotide of one another, resulting in spuriously large sequencing depth. In this analysis, all stacks with a depth of coverage greater than two SDs above the mean stack depth were removed and the remaining stacks were merged into a locus. Population names are those used in previous publications and are defined by their latitude, longitude, altitude, and state or province of origin as given for each population in Table S1.

Table S1. Origin of *W. smithii* populations

Population*	State/province	North latitude, °	West longitude, °	Elevation, m
Northern clade				
WB	MB	54	101	305
DL	ON	50	94	406
RY	WI	46	92	295
ML	WI	46	90	500
DR	ON	46	78	154
KC	ME	46	68	365
ME	ME	46	69	60
TH	PA	41	75	596
Mid-Atlantic clade				
PB	NJ	40	74	10
MM	NJ	40	75	10
HV	NJ	40	75	10
NP	MD	38	75	18
Appalachian clade				
HS†	NC	35	83	1190
HK	NC	35	83	900
DB	NC	35	83	900
CB	NC	35	83	840
NC Coast clade				
SH	NC	35	80	107
PM	NC	35	80	107
GS	NC	34	78	20
Gulf Coast Clade				
WI	FL	30	85	10
LI	AL	30	87	15

*Each two-letter acronym corresponds to a specific population collected within <1 km over the last 30+ y and has been used consistently to denote that population in all previous publications from this laboratory.

†*S. purpurea* at Highlands Biological Station (HS) was transplanted from Horse Cove (HK), ca. 2 km southeast and 270 m lower in elevation around 1900. We included the HS *W. smithii* because the HK locality has been drained and neither pitcher plants nor *W. smithii* were found at HK in 2004. Both HS and HK samples were collected in 1996 when the plants still existed at HK. To our knowledge, HS represents the only transplanted pitcher plants among all localities used in this study.