Syst. Biol. 61(5):763–777, 2012

DOI:10.1093/sysbio/sys048
Advance Access publication on May 3, 2012

Deep Phylogeographic Structure and Environmental Differentiation in the Carnivorous Plant *Sarracenia alata*

AMANDA J. ZELLMER[1], MARGARET M. HANES[2], SARAH M. HIRD[1], AND BRYAN C. CARSTENS[1,*]

[1]*Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA; and* [2]*Department of Biology, Eastern Michigan University, Ypsilanti, MI 48197, USA;*
**Correspondence to be sent to: Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA. E-mail: bryan.c.carstens@gmail.com*

Received 9 September 2011; reviews returned 29 November 2011; accepted 25 April 2012
Associate Editor: Emily Lemmon

Abstract.—We collected ~29 kb of sequence data using Roche 454 pyrosequencing in order to estimate the timing and pattern of diversification in the carnivorous pitcher plant *Sarracenia alata*. Utilizing modified protocols for reduced representation library construction, we generated sequence data from 86 individuals across 10 populations from throughout the range of the species. We identified 76 high-quality and high-coverage loci (containing over 500 SNPs) using the bioinformatics pipeline PRGmatic. Results from a Bayesian clustering analysis indicate that populations are highly structured, and are similar in pattern to the topology of a population tree estimated using *BEAST. The pattern of diversification within *Sarracenia alata* implies that riverine barriers are the primary factor promoting population diversification, with divergence across the Mississippi River occurring more than 60,000 generations before present. Further, significant patterns of niche divergence and the identification of several outlier loci suggest that selection may contribute to population divergence. Our results demonstrate the feasibility of using next-generation sequencing to investigate intraspecific genetic variation in nonmodel species. [Carnivorous plants; local adaptation; next-generation sequencing; phylogeography; Roche 454; *Sarracenia alata*.]

The carnivorous pitcher plant, *Sarracenia alata*, has a disjunct distribution, with populations on either side of the Atchafalaya Basin/Mississippi River (Fig. 1). This wide biogeographic barrier (>100 km) has been implicated as a source for phylogeographic breaks in a variety of organisms (Soltis et al. 2006; Jackson and Austin 2010), and has inspired several previous investigations in *S. alata* (Sheridan 1991; Neyland 2008; Koopman and Carstens 2010). These studies did not identify substantial morphological (Sheridan 1991) or fixed genetic variation (Neyland 2008; Koopman and Carstens 2010) across the range of this pitcher plant species, presumably because of the paucity of genetic markers available at the time. However, with novel next-generation sequencing methods quickly becoming available for phylogeography, the question of diversification within *S. alata* is worth revisiting, primarily because much of what we know about the life history of *Sarracenia* suggests that these physically isolated populations should be genetically isolated as well.

Sarracenia alata are habitat specialists with a patchy distribution (Bayer et al. 1996), confined primarily to longleaf pine savannahs characterized by abundant sunlight and frequent fires. Seeds in the genus have limited dispersal and low rates of establishment (Ellison and Parker 2002). The seeds are small with no adornments for animal-mediated movement and do not float, suggesting that they are not transferred via floodwater (but see Neyland 2008). Furthermore, while little is known about pollination in this species, the distance separating the eastern and western ranges of *S. alata* is large enough that it is unlikely that animal-mediated long distance pollination is currently contributing to significant gene flow between these populations. Thus, the life history characteristics of *S. alata* strongly suggest that the eastern and western populations should be genetically isolated.

Human-induced habitat modification has led to dramatic reductions in longleaf pine savannah habitat over the last two centuries (Noss 1988). As a result, contemporary *S. alata* populations occupy isolated patches of habitat surrounded by land unsuitable for growth. Although previous results based on microsatellite data indicate that *S. alata* exhibits population genetic structure (Koopman and Carstens 2010), it is unclear if this structure is in response to recent habitat fragmentation, biogeographic barriers, or environmental variation between the disjunct portions of its range. Given the low levels of variation identified in commonly used markers such as ITS or chloroplast DNA (Neyland 2008; Koopman and Carstens 2010), we instead collected sequence data from anonymous portions of the nuclear genome. Loci were isolated via the creation of a reduced representation library (RRL), and sequenced using Roche 454 Titanium chemistry. RRLs allow for isolation of a specific, largely unbiased (e.g., including both coding and noncoding regions) subset of the genome (Altshuler et al. 2000; Whitelaw et al. 2003; Barbazuk et al. 2005; Williams et al. 2010). Using these data, we estimated the timing and pattern of divergence among populations of *S. alata* and explored the potential for adaptive differentiation using both genetic and environmental modeling approaches.

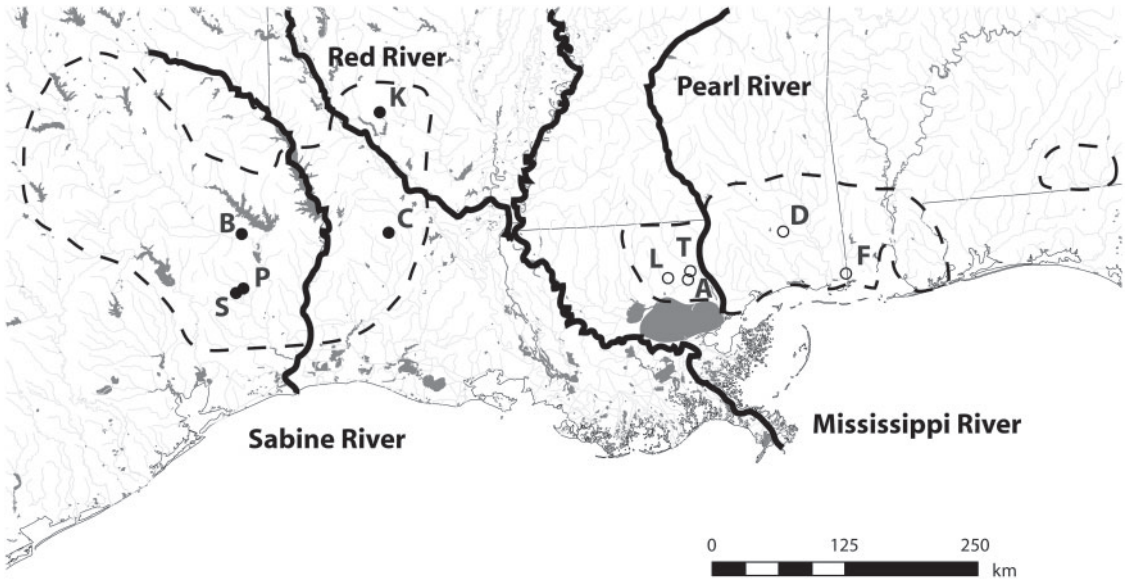

FIGURE 1. Distribution map of *Sarracenia alata* in the southern US. Dashed lines show the approximate range of the species. Western sampling localities from this study are marked with filled circles and include the following populations: Sundew (S), Pitcher Trail (P), Bouton Lake (B), Cooter's Bog (C), and Kisatchie (K). Eastern populations are marked with open circles and include the following populations: Abita Springs (A), Talisheek (T), Lake Ramsey (L), Franklin Creek (F), and DeSoto (D). Minor rivers (light grey) and major rivers and water bodies (dark gray and thick black lines) are shown.

METHODS

Field Collections and DNA Extraction

Tissue from 89 *S. alata* leaves was collected from 10 populations across the range of the species (Fig. 1; online Table S1, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6). An entire leaf was removed from the plant, all fluid from the pitcher was immediately drained, and the leaf was immediately placed in silica gel in the field. To avoid sequencing a large proportion of repetitive, plastid DNA, we removed plastid DNA by following a nuclear DNA extraction protocol, modified to allow for small-scale extractions from dried leaf material (Rabinowicz et al. 1999). Between 0.3 and 0.8 g of dried tissue was ground in liquid N_2 and homogenized with a handheld tissue homogenizer. Cetyl trimethylammonium bromide incubation was performed for 45 min on an orbital shaker. Isopropanol precipitation was conducted overnight. DNA was quantified on an ND 1000 Spectrophotometer (NanoDrop, Wilmington, DE, USA).

RRL Construction

A modified AFLP protocol (Vos et al. 1995) was performed in four steps following the protocol of Gompert et al. (2010; all enzymes from New England Biolabs unless otherwise noted): (i) digestion and ligation: 250 ng DNA was digested and adaptors were ligated to the digested DNA in an 11 μL reaction containing T4 Ligase buffer, NaCl, 5U EcoRI, 5U MseI, 10 μM EcoR1 adaptor, 10 μM MseI adaptor, 1 mg/mL BSA, and 0.5U T4 ligase for 2 h at 37°C (see Vos et al. 1995 for adaptor sequence). (ii) Preamplification: a 20 μL PCR was performed with 10 μL of a 10-fold dilution of ligation product, 2 μL of 25 mM $MgCl_2$, 2 μL of 5X Phusion buffer, 0.4 μL of 10 μM dNTPs, 0.06 μL of 100 μM each adaptor-specific primer (EcoR1: 5′ GACTGCGTACCAATTC; MseI: 5′ GATGAGTCCTGAGTAA), and 0.08 μL of 5 U/μL Phusion High Fidelity DNA Polymerase (Phusion products: Finnzymes Woburn, MA, USA). The conditions of the PCR program were: 98°C for 2 min, 20 cycles of: 98°C for 15 s, 56°C for 30 s, and 72°C for 30 s, with a final extension cycle at 72°C for 10 min. (iii) Size selection: all PCR product was separated by size on a 1.5% SeaKem gel (Lonza Rockland, ME, USA). A band was cut at 450–600 bp, purified using the Qiagen gel extraction kit (Valencia, CA, USA) and eluted with a final volume of 50 μL H_2O. (iv) Selective amplification: a selective amplification step was completed to add the 454 A and B priming sequences and the 10-bp MID barcodes (Roche) to each sample and to reduce the portion of the genome to be amplified and sequenced. Barcodes identify individuals and prevent the opportunity that substitution error would cause incorrect sample assignment (Hamady et al. 2008). All barcodes differed by at least two bases from any other barcode. We evaluated the effectiveness of one vs. two selective base pairs, where additional selective base pairs are expected to further reduce the portion of the genome to be sequenced and thus result in higher coverage per locus. The primers

were as follows—forward (A Fusion/MID-tag/EcoRI Sequence): 5′(CGTATCGCCTCCCTCGCGCCATCAG) (XXXXXXXXXX) (GACTGCGTACCAATTC)3′, where X refers to the MID-tag location; reverse (B Fusion/MseI Sequence/Selective Base(s)): 5′(CTATGCGCCTT GCCAGCCCGCTCAG) (CTATGCGCCTTGCCAGCCC GCTCAG)(C(G))3′. Test 454 sequencing runs were conducted by Engencore (Columbia, SC, USA). Final selective PCRs and 454 sequencing were performed and purified by gel as above by Research and Testing Laboratories (Lubbock, TX, USA).

Bioinformatic Processing

The sequences generated from 454 sequencing were quality controlled using the RDP Pipeline Initial Processor (Cole et al. 2007, 2009). We removed all sequence reads <100 bp, those with a quality score <20, and any containing an ambiguous base or lacking a forward primer, reverse primer, or individual tag. The remaining sequences were then sorted by MID barcodes and separated by individual. We then used the program PRGmatic (Hird et al. 2011) to construct a provisional reference genome (PRG) from our data. PRGmatic clusters all the reads within an individual then uses the high-coverage clusters as provisional alleles; it then reduces the provisional alleles into provisional loci by clustering at a lower percent identity to gather all alleles for a given locus together. These loci are then concatenated and used as a reference. The program can then quickly align all the original data to the reference and call diploid loci for all individuals that surpass coverage thresholds set by the user. PRGmatic aims to create a high confidence set of loci from species that lack a reference genome so that genotyping and/or SNP calling tools developed for application in model systems (where phase is known with high confidence in at least one individual) can be applied to nonmodel systems such as *S. alata*.

We were conservative throughout our bioinformatics processing in order to minimize errors in genotyping that might subsequently lead to errors in inference. For example, we required that an allele be sequenced at a minimum of 5X coverage before including it in the PRG so that we could be confident in the quality of the reference genome even given the relatively high error rates of 454 sequencing. This parameter has no effect on which loci were called in the individuals, but instead dictates the coverage required *within a single individual* for an allele to be used in our PRG. We chose 5X based on preliminary runs and the desire to minimize discarding good data while not including spurious or unsupported alleles. This represents a relatively conservative choice, since the same sequence needs to be almost identical in one individual at least five times. By using a high coverage threshold here, we can manipulate the coverage required to call the genotype in all individuals, since the PRG was constructed from a minimum of 5X coverage within a single individual. Note that any sequence reads included in the PRG were subject to SNP calling as described below.

We experimented with a variety of settings for clustering, alignment, and SNP calling. We explored four different clustering percentages (90%, 92%, 95%, and 98% similarity) to cluster reads into loci that will form the PRG. If the clustering level is set too high, the putative alleles at any given locus will be separated into different loci (oversplitting) and as a result few individuals will appear heterozygous. However, when reads are clustered at too low a value of percent similarity, separate loci will be falsely combined and a majority of individuals will appear heterozygous (overlumping). For each of the clustering levels, we employed a BLASTn search against all sequences in GenBank (accepting alignments with an E-value $< 1.0 \times 10^{-4}$). At a 90% clustering level, the resulting data aligned to the fewest loci in GenBank, whereas at a 98% clustering level, some loci that were called as independent by PRGmatic aligned to the same loci in GenBank. Results from the 92% and 95% levels were similar to one another, and we chose to analyze those sequences from the 95% clustering level because this level produced the fewest loci that had evidence of being oversplit or overlumped (based on BLASTn results and visual evaluation of alignments per the recommendation of Hird et al. 2011). Preliminary population structure analyses were run using each of the four clustering levels and the results were similar regardless of clustering level used. We also explored both liberal and conservative settings for the number of sequence reads required for calling SNPs within an individual (three versus. six sequences, respectively). Because the liberal settings often resulted in high levels of ambiguous bases, we used the conservative settings (i.e., a minimum of 6) for all analyses. Thus, every base in all of the genotypes produced by PRGmatic was represented by a minimum of six sequence reads in any particular individual.

After determining the individual genotypes for each locus, we performed additional quality control steps to remove paralogs and loci with low coverage. Due to the potential for paralogy, loci were not analyzed if they met one of the following criteria: (i) loci that had alleles of various lengths, which can be evidence of paralogy since this is indicative of mutations in one of the restriction cut sites; and (ii) loci that contained any nucleotide position with more than two nucleotide bases in an individual. Because the latter can occur from sequencing error as well, we set a very conservative threshold of 0.0005 sites with >2 bp per number of individuals sequenced per locus size. Only loci falling below this conservative threshold were retained for the analyses. Additionally, this quality control step provided an opportunity to address the issue of homopolymer errors. Although PRGmatic does not specifically isolate homopolymer errors, the sequences including homopolymer errors often resulted in individuals having >2 bp at a nucleotide site. As a result, many of the loci with homopolymer errors were discarded from the analyses during this step.

Together, these quality control steps resulted in a very conservative data set with only the highest quality loci, which we hereafter refer to as the *full data set*. We also calculated a set of basic summary statistics using the compute package in libsequence (Thornton 2003), including: Tajima's D, Fu and Li's D* and Fu and Li's F* (Fu and Li 1993). For all phylogenetic analyses (described below), we utilized a reduced data set including only loci that were found in all 10 sampled populations. Additionally, since the phylogenetic analyses assume no recombination within loci, we tested each locus for internal recombination sites using IMgc (Woerner et al. 2007). Loci that showed evidence of recombination were broken up into their two largest nonrecombining blocks.

Population Structure

Individual and population assignments were conducted using the program Structure (Pritchard et al. 2000). We explored an admixture model to determine the number of population clusters (k) from 1 through 11 using a burn-in of 2.0×10^6 and 2.0×10^6 replications. Analyses were repeated five times for each k value. The average and standard deviation (SD) of the likelihood of each model were used to calculate Δk (Evanno et al. 2005) using Structure Harvester (Earl and vonHoldt 2011), and the partitioning scheme with the highest Δk was selected as the model with the most support.

Outlier Loci

A Bayesian analysis of molecular variance (BAMOVA; Gompert and Buerkle 2011) was used to test for population subdivision (Φ_{ST}) and to identify outlier loci (i.e., loci that fall outside of the expectation for genome-level Φ_{ST}). The program Bamova (Gompert and Buerkle 2011) calculates both genome-level and locus-specific Φ_{ST} to determine the amount of genetic variation explained by population substructure taking into account the presence of loci potentially under selection and also to identify those loci that may be under selection. The Markov chain was sampled every 10 steps for 250,000 generations with the first 5000 samples discarded as burn-in. The random walk haplotype frequency vector was used, with additional parameters set to the default. The posterior distributions for the genome- and locus-level Φ_{ST} (as well as their associated 95% confidence intervals (CIs)) were summarized. Loci were considered to be outliers if their CIs were completely outside of the genome-level Φ_{ST} CIs. Loci with Φ_{ST} values below the genome-level CIs are thought to be under balancing or purifying selection, whereas loci above the genome-level CIs are either under positive selection within populations or divergent selection among populations (Gompert and Buerkle 2011).

Phylogeographic Analyses

We conducted two complementary sets of phylogeographic analyses to estimate the pattern and timing of diversification of *S. alata* populations. First, we generated a species tree estimate of the population phylogeny using *BEAST v 1.6.2 (Drummond and Rambaut 2007; Heled and Drummond 2010). Models of sequence evolution for each locus were chosen using DT-ModSel (Minin et al. 2003) and set to the closest match in *BEAST. We experimented with both strict and relaxed clock (exponential and lognormal) priors (Drummond et al. 2006) for each locus. Although both settings produced the same pattern of diversification, the final analyses were conducted with a relaxed clock to allow for heterogeneity in substitution rate across loci. *BEAST analyses were conducted using 5.0×10^8 steps in the Markov chain with the initial 5.0×10^6 steps discarded as burn-in. TreeAnnotator (Drummond and Rambaut 2007) in the BEAST package was used to visualize the maximum clade credibility phylogeny.

Estimates of phylogeny under the species tree paradigm (Edwards 2009) can be biased by population dynamics such as gene flow (Eckert and Carstens 2008); consequently, we sought to compare the estimates of divergence generated by *BEAST with an estimate of population divergence that also incorporated gene flow. The program IMa2 (Hey 2010), which jointly estimates θ ($4N_e\mu$), migration, and population divergence, was utilized to estimate these parameters across the eastern and western populations. The run was conducted using a burn-in of 100,000 steps, prior values of $\theta = 30$, $M = 5$, $\tau = 10$, and a geometric heating scheme using 150 coupled Markov chains (as described in the user manual). Although the divergence estimates from *BEAST and IMa2 are not exactly equivalent, the comparison allows at the very least for a qualitative assessment of the degree to which the phylogenetic estimate of divergence may be misled by unaccounted-for gene flow. The comparison of divergence estimates across the Mississippi River using these approaches is further justified by the recent finding that population substructure does not strongly bias estimates of divergence time using IMa2 (Strasburg and Rieseberg 2011).

Environmental Niche Divergence

We developed ecological niche models for both eastern and western populations of *S. alata* using climate data for each of the 19 BIOCLIM variables downloaded from the WorldClim data set (Hijmans et al. 2004; Hijmans et al. 2005). MaxEnt v 3.3.3e (Phillips et al. 2006) was used to produce niche models for the eastern and western lineages using 24 known *S. alata* sites (Fig. 4): the 10 sampling locations plus 14 additional georeferenced *S. alata* sites either downloaded from the Global Biodiversity Information Facility Data Portal (www.gbif.org, last accessed May 25, 2012) or kindly provided by other researchers (Horner J., personal

communication). To prevent pseudo-replication, only sampling localities that were separated by at least 1 km were used in the analyses.

To test for divergence in the niches occupied on either side of the Mississippi River, we used a multivariate niche method (McCormack et al. 2010). This method utilizes a principal components analysis to describe environmental variation across multiple niche axes for each of the georeferenced *S. alata* localities and for randomly chosen background locations (including both suitable and nonsuitable habitat to illustrate the habitat available to either lineage). The average difference in the principal components axes are then calculated for the eastern and western populations and evaluated against 1000 jackknife comparisons based on the background data. If the average difference between the eastern and western populations for any axis is significantly greater than the average difference in the background points, then this suggests that niches in the eastern and western portions of *S. alata*'s range are significantly different. Similarly, niche conservatism is supported for any axis where eastern and western populations are significantly more similar to one another than the difference in the background points of each.

To conduct the analyses, climate data were extracted from the WorldClim data set for both the 24 actual *S. alata* localities as well as for 850 random background localities across the distribution of the species. For the background points, a minimum convex polygon was drawn between the most distal sampling localities for both the eastern and western *S. alata* lineages and 850 random background points were chosen within both of those polygons using Hawth's Tools (Beyer 2004) in ArcGIS V 9.2 (ESRI 2006). Because BIOCLIM variables can be highly correlated, we evaluated correlation among each of the variables and removed variables such that none of the remaining variables had a correlation higher than 0.9 or less than -0.9. Our resulting data set included 10 BIOCLIM variables describing temperature and precipitation variation across the range of *S. alata*. Statistical analyses were conducted in Stata v 11 (StataCorp 2009). Significance in niche differences was calculated using *t*-tests after evaluating for equality of variances and applying a Bonferroni correction ($P = 0.007$). All principal components axes that explained >1% of the data were retained.

Landscape Genetics Analyses

In order to determine whether population divergence should be attributed strictly to the distance separating populations or to barriers to dispersal, we compared genetic divergence among populations with both geographic distance (isolation by distance) and landscape-weighted geographic distance (isolation by resistance) among populations (where "landscape-weighted geographic distance" refers to a distance that accounts for the permeability of the landscape separating two populations). Genetic divergence was quantified using both F_{ST} (calculated in GENEPOP with the full data set; Raymond and Rousset 1995) and patristic distance (PD; calculated using the phylogeny estimate from *BEAST with the reduced data set).

Both geographic (GEO) and landscape-weighted resistance distance (RD) were calculated using Circuitscape v 3.5, which uses circuit theory to calculate the total resistance of the landscape separating pairs of populations (McRae 2006). To calculate resistance distance, GIS layers are used to create friction matrices describing the permeability of the landscape to dispersal, and each cell on the matrix is given a separate cost based on the permeability of the landscape features occupying that cell. The major landscape feature hypothesized to impact landscape permeability in this system is water. At a large scale, major rivers and water bodies (e.g., lakes, Mississippi River) are expected to impede gene flow and dispersal, whereas at a small scale, minor rivers and waterways may have little impact or may instead facilitate gene flow among populations (Fig. 1). In this way, we evaluated the effects of both major and minor rivers on genetic divergence among populations. Spatial landscape data were acquired from inland water shapefiles available through DIVA-GIS (www.diva-gis.org, last accessed May 25, 2012) and were converted into friction matrices using ArcGIS v 9.2 (ESRI 2006) by assigning a cost to the different landscape features. Because the choice of cost values can impact the correlation between genetic and landscape distance, we evaluated a range of costs (1–5000) for both the major and minor landscape features, starting with the lowest possible cost (1) and then up until increases in cost had little to no effect on the results (5000).

Landscape distances were considered to have a significant effect on genetic divergence if they were significantly correlated with genetic distance (Mantel test; Mantel 1967) and remained significant after controlling for geographic distance (partial Mantel test; Smouse et al. 1986). Both Mantel and partial Mantel tests were done in IBDWS (Jensen et al. 2005) with 30,000 randomizations. Sequential Bonferroni corrections were used to evaluate significance (Rice 1989). Partial Mantel tests were only performed if the initial Mantel test was significant or marginally nonsignificant, since partial Mantel tests on nonsignificant Mantel correlations are uninterpretable.

In addition, we evaluated whether environmental differences among the sampling localities have contributed to divergence of the sampled populations (isolation by environment) by assessing the correlation between genetic divergence and pairwise environmental differences between each of the 10 sampled populations. Environmental difference was calculated as the absolute pairwise difference among populations in principal components scores for each of the axes describing variation in climate data (see 'Environmental niche divergence' section above). Mantel tests were used to evaluate the correlation between genetic and environmental distances, and partial Mantel tests were used to evaluate whether this correlation remained

significant after controlling for geographic distance among populations. We used the results from the landscape genetics analyses to inform the isolation by environment analyses, using only the genetic and geographic distance measures with the highest R-value in the aforementioned landscape genetics tests for the Mantel and partial Mantel tests.

Results

Next-Generation Sequencing

Over all runs (2½ 454 plates; online Table S2, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6), we obtained 1,044,772 useable sequences (defined as high quality, tagged, and with both primer regions sequenced) across 86 individuals with an average length of 282.1 bp per sequence. At the 95% clustering level with 6X coverage for SNP calling, PRGmatic identified 1821 loci. Of these, 125 were aligned to existing GenBank sequences (E-value $< 1.0 \times 10^{-4}$) using BLAST (Altschul et al. 1990). After the quality control steps, removing loci with no variation (3 loci), with too few individuals (1576 loci), or with potential paralogs (166 loci), and removing individuals with low coverage (8 individuals), we ultimately analyzed 76 loci across 82 individuals. In total, these data represent ~29 kb of data, averaging 381 bp in length, and containing an average of 10 SNPs per locus (online Table S3, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6). Of the loci used in the 76 locus data set, 5 loci had significant BLASTn matches, and 100% of those were to plant sequences (online Table S4, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6). These 76 loci were utilized for population genetic analyses and for initial screening of the loci. In our final individual-by-locus matrix (86 individuals by 76 loci, online Table S5 and Fig. S1, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6), 1545 cells (23.6%) had coverages >6 sequences; 2472 cells (37.8%) had coverages greater than zero but <6 sequences (thus, were excluded from further analysis); 2519 cells (38.5%) had no sequencing reads recovered. Alignments of the 76 locus data set were deposited in GenBank under accession numbers JN665096–JN667881. Of those 76 loci, there were 10 loci that were sampled in all 10 sampling localities. These loci were all visually inspected. Three of the 10 loci showed significant evidence of containing recombination blocks using IMgc (Woerner et al. 2007), and were subsequently broken up into two loci each. A total of 13 loci were thus utilized for the phylogenetic analyses.

Population Genetic Structure

Likelihood values from the structure analyses increased as sample partitioning increased (Fig. 2a) and leveled off at approximately $k=7$, with the highest mean likelihood at $k=9$ and the highest likelihood for a single run at $k = 8$. The highest Δk value was found at $k=3$ (Fig. 2b), with the three identified clusters being: (i) all populations east of the Mississippi River; (ii) the Kasatchi sampling location (the most northern population in the west); and (iii) the remaining four western populations (online Fig. S2, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6). Additional Δk peaks were found at $k=7$ and $k=9$, suggesting there is hierarchical substructure (Fig. 2b).

Outlier Loci

The BAMOVA estimated a genome-level Φ_{ST} value of 0.63, suggesting that 63% of the genetic variation was partitioned among populations (95% CI: 0.59–0.67). Consistent with summary statistics calculated using compute, the BAMOVA suggested that some of the loci were under either positive or purifying selection (online Table S3, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6). However, only two loci were identified by at least one of the summary statistics and the BAMOVA, suggesting that many of the significant results from summary statistics may be false positives.

Phylogeographic Analyses

The *BEAST Bayesian phylogenetic analysis suggested that eastern and western populations are each monophyletic, with very high support (posterior probability = 1.0). Effective sample sizes (ESSs) were high (posterior ESS > 4217) suggesting that the Markov chains were convergent across parameter values. Population divergence mirrored the structure results in that the deepest nodes of the population phylogeny corresponded to the partitioning of samples at the lowest k values (online Fig. S2, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6). The diversification of *S. alata* populations also mirrored the physical division of the landscape by major rivers (Fig. 3). If one is willing to assume the genome-wide estimate of mutation rate from *Arabidopsis thaliana* of 7.0×10^{-9} substitutions per site per generation (Ossowski et al. 2010) for divergence dating in *S. alata*, divergence among the eastern and western lineages is estimated to be ~60,000 generations before present.

The two population (east and west) IMa2 analysis was conducted for ~1.1×10^6 generations following the burn-in. Effective samples sizes were moderate (Log[P] = 57, $\tau = 59$, remainder > 150), but plots of the sampled parameter values did not exhibit trends, suggesting that the Markov chain sampled from a stationary posterior distribution of parameter values. If mutation rates of *Arabidopsis thaliana* are assumed (Ossowski et al. 2010), population divergence is estimated to be ~279,567 (95% highest posterior density = 113,107–428,954) generations before present. Results also suggest

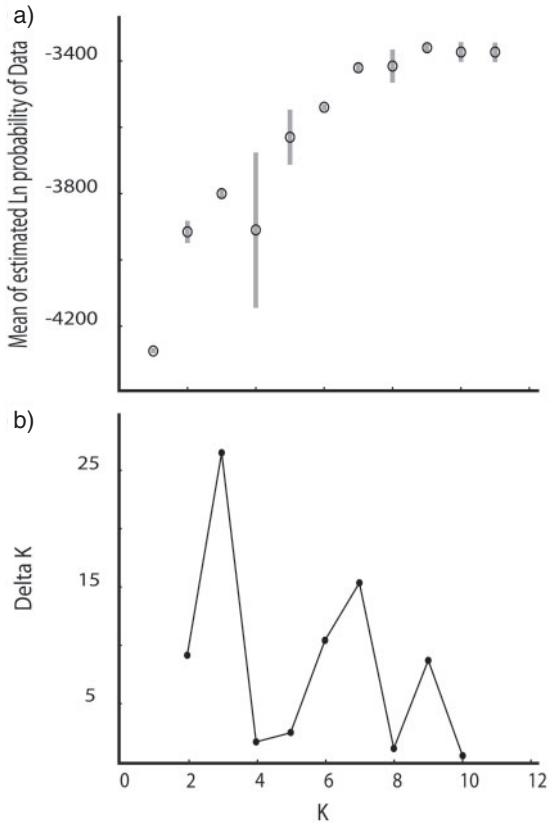

FIGURE 2. Population clustering analyses suggest significant population structure among *S. alata* populations. A) Likelihood scores for each value of k genetic clusters from STRUCTURE (Pritchard et al. 2000). B) Δk scores for each value of k genetic clusters following Evanno et al. (2005). Figures were generated using Structure Harvester (Earl et al. 2011).

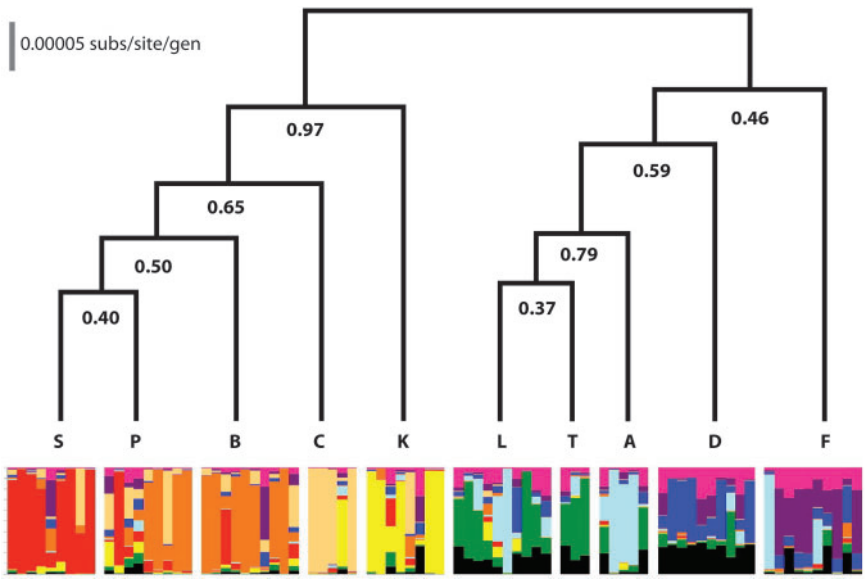

FIGURE 3. Maximum Clade Credibility tree for the 10 sampled populations generated using *BEAST (Drummond and Rambaut 2007; Heled and Drummond 2010). The population phylogeny is shown at top, with posterior probabilities of each node shown. Scale bar to left of phylogeny corresponds to 5.0×10^{-5} substitutions / site / generation. The bottom of the figure shows histograms generated by Structure (Pritchard et al. 2000) for the sampling localities (i.e., $k = 10$).

TABLE 1. Results from IMa2 analysis, showing estimates and 95% highest posterior density (HPD) values of divergence time ($\tau = t\mu$, where t is the number of generations of divergence) among eastern (E) and western (W) lineages, ancestral ($\theta_A = 4\mathrm{Ne}\mu$) and current ($\theta_E$ and θ_W) population sizes, as well as migration rates between populations

Parameter	95% HPD low	High point	95% HPD high
τ	0.15	0.66	1.47
θ_E	1.5	2.7	4.51
θ_W	0.18	0.42	1.62
θ_A	0	0.126	1.5
$m_{W>E}$	0	0.13	0.91
$m_{E>W}$	1.31	2.63	4.83

that effective population sizes in the eastern populations are larger than either the western or the ancestral populations, with low estimates of gene flow from west to east and slightly higher gene flow from east to west (Table 1).

Environmental Niche Divergence

Climate varied significantly between *S. alata* populations to the east and west of the Mississippi River. Niche models for the western populations predicted distribution in the west with small overlap into the eastern range, whereas the eastern distribution was predicted almost exclusively to the east of the Mississippi River (Fig. 4a). The principal components analysis also showed significant variation among the environmental conditions of the populations east and west of the Mississippi River. The first seven principal components accounted for ~99% of the variation in the 10 BIOCLIM variables (Table 2).

Using the multivariate niche method, we found evidence for both niche divergence and conservatism across the Mississippi River (Figs. 4b,c; Table 2). The differences among the eastern and western populations in both PC1 (mean divergence = 3.25) and PC6 (mean divergence = 0.70) were significant (*t*-tests with Bonferroni correction: $P < 0.007$) and significantly greater than expected based on the difference in the background data (background mean and 95% CIs: PC1 = 2.52 (2.39–2.63), PC6 = 0.09 (0.05–0.13); $P < 0.01$), suggesting divergence. In comparison, there was no significant difference between the eastern and western populations for either PC2 (0.631) or PC3 (0.338), and these differences were significantly less than expected based on the background data (PC2 = 2.13 (2.03–2.22), PC3 = 0.78 (0.71–0.85); $P < 0.01$), suggesting niche conservatism.

Landscape Genetics

There was a significant positive correlation between genetic divergence and geographic distance, regardless of whether F_{ST} or PD was used as the measure of genetic divergence (F_{ST}: $R = 0.36$, $P = 0.019$; PD: $R = 0.86$, $P < 0.0001$; online Table S6, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6), resulting in a significant pattern of isolation by distance. Although the trends were similar between F_{ST} and PD, PD showed a stronger correlation (higher R-value) than F_{ST} with each of the alternative models of geographic distance (online Table S6, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6). We, therefore, focus only on the results of the analyses using PD as the measure of genetic divergence.

There was also significant evidence of isolation by resistance. Although there was a significant correlation between PD and resistance distance (RD) for each of the alternative geographic distance models (online Table S6, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6), the results suggest that only major rivers and water bodies have a significant impact on genetic divergence. First, the R-values for all RD models including major rivers were greater than the R-value for geographic distance (GEO), whereas the R-values for RD models including minor rivers were all less than the R-value for geographic distance. Second, increasing cost values resulted in increasing R-values for RD models with major rivers, whereas for RD models including minor rivers increasing cost values led to decreasing R-values (online Table S6, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6). Lastly, partial Mantel tests confirm that RDs based on major rivers are a better predictor of genetic divergence. For RDs based on major rivers (except in the case of the lowest cost model), the correlation between genetic distance and RD remained significant after controlling for GEO, whereas the correlation between genetic distance and GEO was not significant after controlling for RD (online Table S6, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6). The opposite pattern was observed for partial Mantel tests of RDs based on minor rivers: the correlation between genetic distance and RD was not significant after controlling for GEO, whereas the correlation between genetic distance and GEO was significant after controlling for RD (online Table S6, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6). Overall, the model with the best support was the RD model including major rivers at a cost of 5000 ($R = 0.89$, $P < 0.0001$). This RD model remained significantly correlated with genetic divergence after controlling for geographic distance ($R = 0.51$, $P = 0.008$), whereas geographic distance was not significantly correlated with genetic divergence after controlling for this RD model (online Table S6, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6). Consequently, this RD model was used for subsequent analyses.

The results for the test of isolation by environment were inconclusive. Genetic divergence (PD) and environmental distance (pairwise absolute difference in principal components scores) among populations were significantly positively correlated for one principal component axis, PC1 (Mantel test: $R = 0.32$, $P < 0.002$), marginally nonsignificant for two axes, PC5

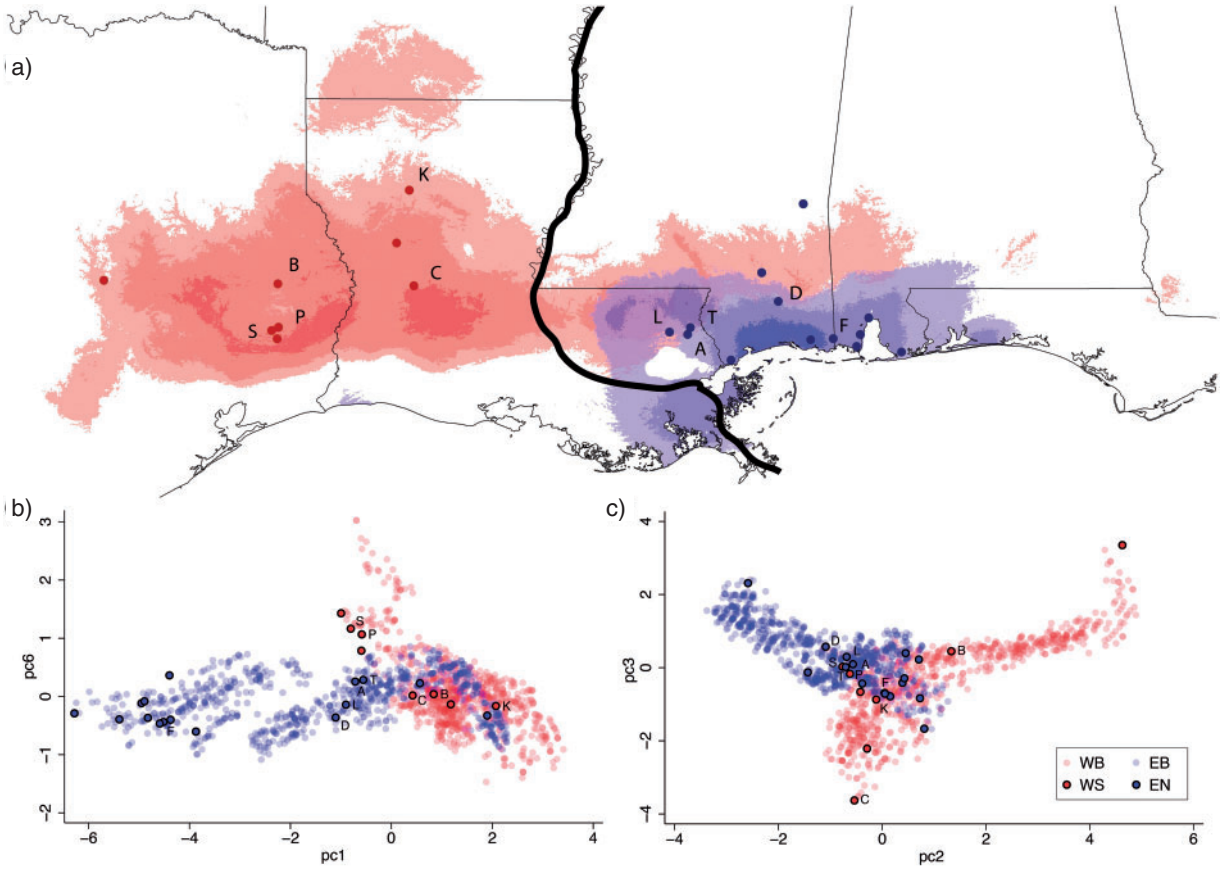

FIGURE 4. Environmental niche models and environmental variation for eastern (blue) and western (red) populations of *S. alata*. a) Divergence in niches across the Mississippi River. Predictions were calculated using MAXENT v 3.3.3e (Phillips et al. 2006), with darker colors showing greater prediction scores. The thick black line denotes the Mississippi River. b) Principal components axes 1 and 6 show significant niche divergence, c) whereas axes 2 and 3 show significant niche conservatism. WS: western sampled points; WB: western background points; ES: eastern sampled points; EB: eastern background points.

($R = 0.21, P < 0.053$) and PC7 ($R = 0.26, P < 0.055$; Table 3), and nonsignificant for the other four PC axes. However, partialMantel correlations show that these correlations are mostly due to the distance separating populations rather than the environmental differences among populations. The correlation between genetic distance and landscape distance (RD) remained significant after controlling for environmental differences (e.g., PC1, PC5, and PC7), whereas the correlation between genetic distance and environmental distance was either nonsignificant (e.g., PC1) or was marginally nonsignificant (e.g., PC5 and PC7; Table 3) after controlling for RD.

DISCUSSION

Similar to other Gulf Coast organisms (e.g., Soltis et al. 2006), the range of *Sarracenia alata* is divided by the Atchafalaya Swamp and Mississippi River. In contrast to previous investigations, which did not identify diagnosable differences of either morphological characters (Sheridan 1991) or ITS sequences (Neyland 2008), here we identified substantial divergence (at least 60,000 generations before present; Fig. 3) between eastern and western populations of *S. alata* and determined that these populations have been isolated by major rivers. Our results were consistent with previous work that identified population genetic structure using microsatellite markers (Koopman and Carstens 2010); however, it is still prudent to explore our approach in some detail given that the application of next-generation sequencing to phylogeographic investigations is not as well understood as genotyping microsatellites or sequencing genes such as ITS or chloroplast DNA.

Phylogeography of S. alata

Early phylogeographic investigations applied 'tree thinking' to population level variation (e.g., Avise 2000) by estimating genealogies from organellar genes and assuming that the pattern of coalescence reflected the pattern of population diversification (e.g., Avise et al. 1987). Over time, phylogeographers became convinced that single locus data were unlikely to adequately track the history of population divergence (Edwards and Beerli 2000; Hudson and Coyne 2002; Knowles 2004). It is

TABLE 2. Environmental variation among eastern and western populations

Variable	Description	PCI (0.45)	PC2 (0.30)	PC3 (0.11)	PC4 (0.05)	PC5 (0.03)	PC6 (0.03)	PC7 (0.02)
Biol	Annual mean temperature	−0.4039	0.2211	−0.0385	0.2644	−0.1943	0.2812	0.4057
Bio2	Mean diurnal range (mean(period max–min))	0.4457	−0.0805	0.1888	0.1151	0.1636	0.2938	0.0691
Bio3	Isothermality (Bio2/Bio7)	0.2278	−0.3956	0.3685	0.423	−0.0454	0.2702	0.3911
Bio7	Temperature annual range (Bio5-Bio6)	0.441	0.1591	0.0259	−0.1092	0.2125	0.2435	−0.1247
Bio8	Mean temperature of wettest quarter	−0.3267	0.2352	0.2772	0.4779	0.6587	−0.2156	−0.0815
Bio9	Mean temperature of driest quarter	0.3406	0.3048	−0.0327	−0.0926	0.0205	−0.5795	0.6659
Bio10	Mean temperature of wannest quarter	−0.1131	0.5266	−0.113	−0.0756	−0.1081	0.4965	0.212
Biol2	Annual precipitation	−0.2641	−0.4553	−0.0313	0.0631	−0.263	−0.1599	0.1714
Biol4	Precipitation of driest period	−0.0302	−0.3214	−0.72	−0.0211	0.5228	0.1703	0.2606
Biol5	Precipitation seasonality (coefficient of variation)	−0.2927	−0.1634	0.4651	−0.692	0.3159	0.1377	0.2644
	Mean niche difference	3.25[a]*D*	0.63*C*	0.34*C*	0.44	0.59	0.70[a]*D*	0.05
	Mean background difference	2.52	2.13	0.78	0.03	0.30	0.09	0.05
	95% CI background difference	(2.39 − 2.63)	(2.03 − 2.22)	(0.71 − 0.85)	(0.00 − 0.08)	(0.27 − 0.34)	(0.05 − 0.13)	(0.01 − 0.08)

Notes: Loadings for each variable on the seven retained principal components axes for the multivariate niche assessment. Niche differences among the eastern and western lineages as well as among the background environments are listed at the bottom of the table. The proportion of variance explained by each axis is listed in parentheses. The test for significance in niche differences between the eastern and western populations based on locality data alone is indicated by an asterisk. The test for niche divergence versus niche conservatism based on the background environmental data is indicated by either a D (axes with significant divergence relative to the background) or a C (axes with significant niche conservatism relative to the background).
[a]Niches differ significantly (*t*-test, Sequential Bonferroni adjustment (Rice et al. 1989)).

TABLE 3. Correlation between environmental distance (pairwise absolute difference in principal components scores, PC1-PC7) and genetic distance (Patristic Distance)

Envt variable	Mantel test		Partial Mantel tests			
	EnvtDis vs. Genetics		RD vs. Genetics (controlling for EnvtDist)		EnvtDist vs. Genetics (controlling for RD)	
	R	*P*	R	*P*	R	*P*
PC1	0.315	*0.002**	0.879	*0.000**	0.097	0.211
PC2	–0.018	0.487	—	—	—	—
PC3	0.053	0.316	—	—	—	—
PC4	0.125	0.163	—	—	—	—
PC5	0.207	0.053	0.893	*0.000**	0.241	*0.028*
PC6	0.090	0.248	—	—	—	—
PC7	0.263	0.055	0.894	*0.000**	0.310	*0.041*

Partial Mantel results show the correlation between landscape distance (RD) and genetic distance controlling for environmental distance (EnvtDist) as well as for EnvtDist and genetic distance controlling for RD. The RD model used included major rivers with a cost of 5000, which was the best-fit model from the landscape genetics analyses. Significance at $P < 0.05$ is indicated by italics. Significance after sequential Bonferroni adjustment is indicated with an asterisk. Partial Mantel tests were only performed on axes that had significant or marginally nonsignificant Mantel correlations.

time to revisit this question. Methodological advances in species tree phylogeny estimation (e.g., Edwards 2009), particularly when coupled with the larger data sets made possible by new sequencing technologies, allow us to directly estimate population divergence to a degree previously impossible. Although the general utility of tree thinking (e.g., Smouse 1998) for any given system should still be explored, it seems particularly applicable to our focal taxa. In *S. alata*, rivers appear to divide the species into several isolated populations, as evidenced by the structure analyses and the correlation between genetic isolation and resistance distance, and it seems

reasonable to represent the divergence among these populations as a phylogenetic tree. Further, our results are also consistent with aspects of *S. alata*'s life history (e.g., low rates of seed dispersal and establishment) as well as previous investigations using microsatellites (Koopman and Carstens 2010).

The deep divergence among populations using *BEAST implies that a phylogenetic model is appropriate for these data. Divergence between eastern and western populations is estimated to be at least 60,000 generations before present (Fig. 3) on the basis of the phylogeny estimate from *BEAST. In addition, analysis using a coalescent model suggests that population divergence across the Mississippi River is substantial (Table 1). It is worth emphasizing that for our system these models are each inadequate in their own manner. *BEAST does not parameterize gene flow, and while estimates of this parameter from IMa2 are low, it is possible that allele sharing across lineages due to gene flow could lead to estimation error in species tree methods. Although simulations indicate that topology can be accurately estimated when gene flow occurs among sister lineages (Eckert and Carstens 2008), increased polymorphism that results from unaccounted for gene flow may lead to errors in estimation of branch lengths (and thus the estimates of divergence). IMa2, on the other hand, estimates divergence and gene flow simultaneously and can accommodate more than two populations; however, we were unable to achieve good results when attempting to estimate parameters using the phylogeny of the 10 populations investigated here. Given these shortcomings, we are not satisfied with absolute estimates of divergence generated by either method, but still hope to make some broad inferences regarding diversification within *S. alata*.

Given that population divergence between the eastern and western *S. alata* populations is estimated to be at least 60,000 generations, what can we infer about the absolute timing of diversification in this group? Although there are no direct estimates of generation length in *Sarracenia*, the plants are long-lived (Brewer 2001), so even a conservative estimate of generation length (e.g., 2 years) would indicate that eastern and western *S. alata* began to diverge well into the Pleistocene. Although the deepest subdivision within *S. alata* does not predate the formation of the Mississippi River, which is thought to have originated well before the Pleistocene (Mann and Thomas 1968; Cox and Van Arsdale 1997), it may correspond to dramatic shifts in the river's course that occurred during the Pleistocene. Finally, our finding that the Mississippi River represents a major phylogeographic break in *S. alata* is consistent with findings in co-distributed species (Soltis et al. 2006; Jackson and Austin 2010).

At a smaller scale, sampled populations also show significant divergence that likewise corresponds to major rivers (online Table S6, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6), particularly for the more basal splits, and to habitat boundaries, for the shallower divergences (online Fig. S2, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6). Additionally, the BAMOVA estimates indicate that over 63% of the genetic variation can be attributed to among-population variation for the 10 sampled populations. This amount of genetic variation is very high compared with previous studies using BAMOVA (e.g., Gompert et al. 2010), further suggesting that *S. alata* populations are highly structured. The general congruence among these results supports the previously proposed hypothesis (Koopman and Carstens 2010) that riverine barriers are the major factor that promotes population divergence in *S. alata*.

These results also have implications for evolution throughout the genus *Sarracenia*. Historical biogeographic hypotheses for this genus suggest potential expansion along the east coast of the United States following the retreat of the Pleistocene glaciers (Oard 1997). Many *Sarracenia* species are distributed along the eastern Gulf and Atlantic coasts of the United States; in this region, there are many broad rivers, and each of these rivers has the potential to bisect a *Sarracenia* species. Species such as *S. psittacina*, *S. oreophila*, *S. flava* and *S. minor* are all bisected by major rivers; it is therefore reasonable to hypothesize that each may contain a large amount of cryptic genetic subdivision.

Environmental Differentiation within S. alata

Although genetic structure among *S. alata* populations appears to be largely due to neutral divergence across major rivers and with increasing geographic distance, there are significant environmental differences among populations, which may contribute to population divergence. At a broad scale, we found significant niche differences among the eastern and western lineages for two axes (PC1, PC6) of environmental variation (Fig. 4b; Table 2), and these differences are significantly greater than differences in randomly sampled background localities (Table 2). Despite the fact that this species usually inhabits longleaf pine savannahs, the climatic conditions of the habitats on either side of the Mississippi River are significantly different. Surprisingly, however, the strongest axis of environmental difference does not appear to be along a longitudinal axis across the Mississippi River. The divergence across PC1 (tradeoff between temperature annual/diurnal range and annual mean temperature), which accounts for most of the variation among populations, is primarily due to differences between the coastal populations and the interior populations. This is further reflected in the niche models, where the western niche model encompasses the western populations plus all interior eastern populations, whereas the eastern niche model is centered on the coastal populations, with the interior eastern populations showing low prediction scores (Fig. 4). Thus, it appears that if selection is occurring in this system due to abiotic factors, it is likely due to differences between coastal and inland climatic conditions as opposed to arbitrarily bisecting into east

and west habitats. Future studies focusing on the inland versus coastal populations will be necessary to further investigate the effects of these environmental differences on genetic divergence.

When considering genetic variation among the 10 sampled populations, rather than the more broadly defined eastern and western habitats, there is limited evidence for the presence of local adaptation. Populations with greater environmental differences (absolute difference in PC5 and PC7) show only a marginally nonsignificant trend toward increased genetic divergence (PD) after controlling for landscape distance (Table 3). At this smaller scale, PC5 and PC7 (differences in the mean temperature of the wettest quarter and mean temperature of the driest quarter, respectively; Table 2), rather than PC1 or PC6, may be more important as drivers of diversification among local populations. The lack of strong evidence for a correlation between local environmental and genetic divergence after controlling for landscape distance may indicate that there either has not been sufficient time for accumulation of genome-wide population subdivision due to environmental differences or else that 10 populations may not be sufficient to evaluate correlations in local environmental variation and genetic divergence.

Lastly, the presence of significant values for the various molecular tests of selection (D, D*, and F* and Φ_{ST} outliers; online Table S3, available from Dryad data repository; doi:10.5061/dryad.hk25q4d6) also indicates that natural selection may be influencing the pattern of genetic variation across these populations. Although the results of these analyses are subject to type I errors and can be confounded by demographic forces such as population size change (e.g., Hammer et al. 2003) or population substructure (Excoffier et al. 2009), the consistency among the various tests of selection suggest that our results are robust. Future research will focus on screening these loci in order to identify the nature of selection operating within *S. alata*, with a particular focus on responses to environmental differences between the coastal and inland populations.

Roche 454 Sequencing of an RRL

Next-generation sequencing has been quickly transforming many areas of evolutionary biology, yet these new technologies have been slower to come to the field of phylogeography. This delay may be due to difficulties bringing these novel sequencing methods to nonmodel organisms, obtaining orthologous loci among individuals, affordably sequencing multiple individuals, and the need for long sequence reads for gene-tree analyses (McCormack et al. 2011). Next-generation sequencing of RRLs (e.g., Gompert et al. 2010) using individual barcodes added to sequences via PCR (Williams et al. 2010), in conjunction with bioinformatics tools that reduce the need for reference genomes (e.g., Hird et al. 2011), provides an efficient and inexpensive method to generate a large amount of sequence data for nonmodel organisms. However, several challenges must be met before this approach can be widely utilized in phylogeographic analyses, including both preventing and identifying paralogous loci, distinguishing sequencing error from rare variants, and reducing the presence of and understanding the consequences of missing data in these large data sets.

Because the restriction sites used to reduce the representation of the genome are inherited from common ancestors, the preparation of RRLs using restriction enzymes is susceptible to the misidentification of loci that are actually paralogous as single copy. This misidentification would have substantial effects on various analyses, as levels of heterozygosity would be inflated, which in turn may inflate estimates of important parameters such as divergence time. The use of a double-digest method, as done here, aids in the avoidance of paralogous loci. Regardless, we were extremely conservative when we examined the initial set of loci and removed a large number of loci from the data set, including all loci that showed any evidence of being paralogous (see 'Methods' section). Although we reduced the size of our data set and ultimately analyzed data derived from fewer than 18% of the sequence reads, we are also confident that those loci that were analyzed represent high-quality data that are not likely to mislead our inferences regarding the phylogeography of *S. alata*. For study systems with lower levels of repetitive DNA, this method has much potential for identifying a large number of loci across multiple individuals. However, for systems where high levels of paralogy are expected, other methods of library construction may be preferable, such as PCR amplicon sequencing (e.g., Binladen et al. 2007; Meyer et al. 2008; Tewhey et al. 2009) or target enrichment (e.g., Albert et al. 2007; Okou et al. 2007; Gnirke et al. 2009).

Another major issue with anonymous loci generated via next-generation sequencing methods is that it is difficult to identify high-quality, high-coverage errors as errors or low-quality, low-coverage polymorphism as real variants. Currently, both probability and quality scores are used to construct a working data set from next-generation sequencing data, in an attempt to minimize both the number of errors retained and good reads discarded. Unfortunately, setting thresholds (e.g., for coverage and quality scores) may bias analyses and inferences, since using values too low will include errors, but setting cutoffs too high will exclude real variation. Thresholds are admittedly imperfect, but the genome is heterogeneous and different values apply to different loci; it is currently impossible to determine individual parameters for each locus. Here, we set several different thresholds and analyzed the results to determine which set of parameters led to the most conservative and robust data set (see 'Methods' section). For an enhanced picture of the genome and increased confidence in the data, SNPs could be validated through resequencing or alternate restriction enzymes could be used to evaluate a second set of loci from

some individuals. However, these methods are costly, and we therefore encourage researchers to carefully consider the distribution of their resources with these issues in mind.

Although the RRL method utilized here allowed us to recover sequences from 76 high confidence loci in some individuals, including 13 loci from all sampling localities, there was incomplete coverage across individuals for the much larger set of ~1800 loci identified by PRGmatic. The low coverage is an issue for two reasons: because of the possible impact on the genetic analyses and because it represents an inefficient use of the sequence reads. The presence of missing data within a data matrix can have significant consequences on the results obtained from some analyses, such as estimates of summary statistics (e.g., π or Tajima's D) as well as estimates of both branch lengths and tree topology (Lemmon et al. 2009). However, since coalescent theory assumes that alleles are sampled at random from a population, estimates generated using coalescent models (e.g., *BEAST, IMa2) should not be affected by missing data *per se*, although the quality of such estimates may be correlated with the number of sampled alleles. Similarly, because BAMOVA accounts for stochastic sampling in the likelihood function, estimates of Φ_{ST} should not be biased by missing data, but may result in lower confidence (although simulations verifying this have, to our knowledge, not been published). A bigger concern for researchers with limited budgets is the inefficiency that results from these missing data. Regardless, our results were largely consistent with both biological predictions for this organism as well as previous genetic work with complete data sets (e.g., Koopman and Carstens 2010), suggesting that incomplete data sets due to variation in next-generation sequencing efforts among individuals and loci may not prohibit the recovery of phylogeographic patterns in other studies using similar methods.

With roughly 1.0×10^6 sequence reads resulting from our sequencing effort divided among 80 individuals, we would expect $\sim 1.25 \times 10^4$ reads per individual and ~160X coverage over 76 loci, assuming equal molar concentrations of input samples and sequencing of only these loci. Although we do not yet know the genome size of *S. alata*, the closely related diploid *S. flava* has been reported to have an unexpectedly large genome (Hanson et al. 2005). If *S. alata* has a similarly large genome, it could explain our difficulty in obtaining a more complete data set for all loci. For species with smaller genome sizes, this method may provide much more complete data sets. In addition, we note there are some aspects of library construction that could be improved to this end. One possibility is to select a smaller base pair size range to excise from the agarose gel in order to increase the coverage across loci. However, a smaller size range requires high precision in order to insure overlap in the fragments excised for each sample. Such precision can be difficult to achieve. Second, methods that reduce the number of PCR steps needed prior to next-generation sequencing (e.g., Kozarewa et al. 2009) will have fewer issues with PCR bias and as a result should provide much more complete data sets. Finally, for samples that require amplification via PCR, emulsion methods (e.g., Williams et al. 2006) produce libraries that are far less biased and therefore improve representation across loci.

Conclusions

As in many organisms that occupy the Gulf Coast of North America, the distribution of *S. alata* is divided by the Mississippi River and Atchafalaya swamp. Analysis of sequence data collected using next-generation sequencing suggests that populations in the eastern and western portions of the range have been isolated for tens of thousands of generations. Furthermore, it appears that there is deep divergence within the eastern and western populations and that this divergence is associated with major rivers. Although recent human-induced habitat fragmentation has likely reduced the census size of *S. alata* populations and contributed to the formation of population genetic structure at a local scale, landscape processes that predate human settlement of the region appear to be the dominant factor responsible for structuring genetic diversity in this species.

Supplementary Material

Data and other supplementary material have been deposited in the Dryad data repository under doi:10.5061/dryad.hk25q4d6 (www.datadryad.org).

Funding

This work was supported by grants from the Louisiana Board of Regents and from the National Science Foundation [DEB - 0956069 to B.C.C].

Acknowledgments

We thank V. Smith and M. Olinde for assistance with collection permits, J. Horner for providing spatial data describing *S. alata* populations, S. Dowd and J. McCormack for assistance with sequencing, and D. Fuselier for assistance with field collection of *S. alata*. We thank the Nature Conservancy for access to sampling localities, and members of the Carstens lab for their assistance and discussions regarding various aspects of this investigation. We thank E. Moriarty Lemmon and two anonymous reviewers for helpful comments on the manuscript.

References

Albert T.J., Molla M.N., Muzny D.M., Nazareth L., Wheeler D., Song X., Richmond T.A., Middle C.M., Rodesch M.J., Packard C.J. 2007.

Direct selection of human genomic loci by microarray hybridization. Nat. Methods, 4:903–905.
Altschul S., Gish W., Miller W., Myers E., Lipman D. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.
Altshuler D., Pollara V.J., Cowles C.R., Van Etten W.J., Baldwin J., Linton L., Lander E.S. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407:513–516.
Avise J.C. 2000. Phylogeography: the history and formation of species. Cambridge (MA): Harvard University Press.
Avise J.C., Arnold J., Ball R.M., Bermingham E., Lamb T., Neigel J., Reeb C.A., Saunders N.C. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu. Rev. Ecol. Evol. Syst. 18:489–522.
Barbazuk W.B., Bedell J., Rabinowicz P.D. 2005. Reduced representation sequencing: a success in maize and a promise for other plant genomes. Bioessays 27:839–848.
Bayer R.J., Hufford L., Soltis D.E. 1996. Phylogenetic relationships in Sarraceniaceae based on rbcL and ITS sequences. Syst. Bot. 21: 121–134.
Beyer H.L. 2004. Hawth's Analysis Tools for ArcGIS. Available from: URL http://www.spatialecology.com/htools.
Binladen J., Gilbert M.T.P, Bollback J.P., Panitz F., Bendixen C., Nielsen R., Willerslev E. 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS One 2:e197.
Brewer J.S. 2001. A demographic analysis of fire-stimulated seedling establishment of *Sarracenia alata* (Sarraceniaceae). Am. J. Bot. 88:1250–1257.
Cole J.R., Chai B., Farris R.J., Wang Q., Kulam-Syed-Mohideen A.S., McGarrell D.M., Bandela A.M., Cardenas E., Garrity G.M., Tiedje J.M. 2007. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res. 35:D169–D172.
Cole J.R., Wang Q., Cardenas E., Fish J., Chai B., Farris R.J., Kulam-Syed-Mohideen A.S., McGarrell D.M., Marsh T., Garrity G.M., Tiedje J.M. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37:D141–D145.
Cox R.T., Van Arsdale R.B. 1997. Hotspot origin of the Mississippi embayment and its possible impact on contemporary seismicity. Eng. Geol. 46:201–216.
Drummond A., Ho S., Phillips M., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.
Drummond A., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214–221.
Earl D.A., vonHoldt B.M. 2011. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. 4:359–361.
Eckert A.J., Carstens B.C. 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Mol. Phylogen. Evol. 49:832–842.
Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.
Edwards S.V., Beerli P. 2000. Perspective: Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54:1839–1854.
Ellison A.M., Parker J.N. 2002. Seed dispersal and seedling establishment of *Sarracenia purpurea* (Sarraceniaceae). Am. J. Bot. 89:1024–1026.
Evanno G., Regnaut S., Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14:2611–2620.
Fu Y.X., Li W.H. 1993. Statistical tests of neutrality of mutations. Genetics 133:693–709.
Gnirke A., Melnikov A., Maguire J., Rogov P., LeProust E.M., Brockman W., Fennell T., Giannoukos G., Fisher S., Russ C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27:182–189.
Gompert Z., Buerkle C.A. 2011. A hierarchical Bayesian model for next-generation population genomics. Genetics 187:903–917.
Gompert Z., Forister M.L., Fordyce J.A., Nice C.C., Williamson R.J., Buerkle C.A. 2010. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. Mol. Ecol. 19:2455–2473.
Hamady M., Walker J., Harris J., Gold N., Knight R. 2008. Error-correcting barcoded primers allow hundreds of samples to be pyrosequenced in multiplex. Nat. Methods 5:235–237.
Hanson L., Boyd A., Johnson M.A.T., Bennett M.D. 2005. First nuclear DNA C-values for 18 eudicot families. Ann. Bot. 96:1315–1320.
Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.
Hey J. 2010. Isolation with Migration Models for More Than Two Populations. Mol. Biol. Evol., 27:905–920.
Hijmans R.J., Cameron S., Parra J. 2004. WORLDCLIM 1.2 [Online]. Museum of Vertebrate Zoology University of California Berkeley. Available from: URL http://biogeo.berkeley.edu/worldclim/methods.htm.
Hijmans R., Cameron S., Parra J., Jones P., Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25:1965–1978.
Hird S.M., Brumfield R.T., Carstens B.C. 2011. PRGmatic: an efficient pipeline for collating genome-enriched second-generation sequencing data using a 'provisional-reference genome'. Mol. Ecol. Resour. 11:743–748.
Hudson R.R., Coyne J.A. 2002. Mathematical consequences of the genealogical species concept. Evolution 56:1557–1565.
Jackson N.J., Austin C.C. 2010. The combined effects of rivers and refugia generate extreme cryptic fragmentation within the common ground skink (*Scincella lateralis*). Evolution 64:409–428.
Jensen J.L., Bohanak A.J., Kelley S.T. 2005. Isolation by distance, web service. BMC Genet. 6:13–18.
Knowles L.L. 2004. The burgeoning field of statistical phylogeography. J. Evol. Biol. 17:1–10.
Koopman M., Carstens B. 2010. Conservation genetic inferences in the carnivorous pitcher plant *Sarracenia alata* (Sarraceniaceae). Conserv. Genet. 11:2027–2038.
Kozarewa I., Ning Z., Quail M.A., Sanders M.J., Berriman M., Turner D.J. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat. Methods 6:291–295.
Lemmon A.R., Brown J.M., Stanger-Hall K., Lemmon E.M. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58:130–145.
Mann C.J., Thomas W.A. 1968. The ancient Mississippi River. Gulf Coast Association. Geol. Soc. T. 18.
Mantel N. 1967. Detection of disease clustering and a generalized regression approach. Cancer Res. 27:209–220.
McCormack J.E., Hird S.M., Zellmer A.J., Carstens B.C., Brumfield R.T. 2011. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogen. Evol. http://dx.doi.org/10.1016/j.ympev.2011.12.007
McCormack J.E., Zellmer A.J., Knowles L.L. 2010. Does niche divergence accompany allopatric divergence in aphelocoma jays as predicted under ecological speciation?: insights from tests with niche models. Evolution 64:1231–1244.
McRae B.H. 2006. Isolation by resistance. Evolution 60:1551–1561.
Meyer M., Stenzel U., Hofreiter M. 2008. Parallel tagged sequencing on the 454 platform. Nat. Protoc. 3:267–278.
Minin V., Abdo Z., Joyce P., Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:674.
Neyland R. 2008. Inraspecific systematic relationships of *Sarracenia alata* Wood. (Sarraceniaceae) inferred from nuclear ribosomal DNA sequences. J. Miss. Acad. Sci. 53:238–245.
Noss R.F. 1988. The longleaf pine landscape of the Southeast: almost gone and almost forgotten. Endangered Spec. Update 5:1–8.
Oard M.E. 1997. The evolution of landscapes and lineages in pitcher plants and their moths. Entomology. Baton Rouge (LA): Louisiana State University and A&M. [dissertation].
Okou D.T., Steinberg K.M., Middle C., Cutler D.J., Albert T.J., Zwick M.E. 2007. Microarray-based genomic selection for high-throughput resequencing. Nat. Methods 4:907–909.

Ossowski S., Schneeberger K., Lucas-Lled J.I., Warthman U.N., Clark R.M., Shaw R.G., Weigel D., Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science 327:92–94.

Phillips S.J., Anderson R.P., Schapire R.E. 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190: 231–259.

Pritchard J.K., Stephens M., Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Rabinowicz P.D., Schultz J., Dedhia N., Yordan C., Parnell L.D., Stein L., McCombie W.R., Martienssen R.A. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat. Genet. 23:305–308.

Raymond M., Rousset F. 1995. Genepop (Version-1.2): Population genetics softwarer for exact tests and ecumenicism. J. Hered. 86:248–249.

Ribbands C.R. 1951. The flight range of the honey-bee. J. Anim. Ecol. 20:220–226.

Rice W.R. 1989. Analyzing tables of statistical tests. Evolution 43: 223–225.

Sheridan P.M. 1991. What is the identity of the West Gulf Coastal pitcher plant, *Sarracenia alata*? Carnivorous Plant Newsletter. 20:102–110.

Smouse P. 1998. To tree or not to tree. Mol. Ecol. 7:399–412.

Smouse P., Long J.C., Sokal R.R. 1986. Multiple-regression and correlation extensions of the Mantel test of matrix correspondence. Syst. Zool. 35:627–632.

Soltis D.E., Morris A.B., McLachlan J.S., Manos P.S., Soltis P.S. 2006.
Comparative phylogeography of unglaciated eastern North America. Mol. Ecol. 15:4261–4293.

Strasburg J.L., Rieseberg L.H. 2011. Interpreting the estimated timing of migration events between hybridizing species. Mol. Ecol. 20: 2353–2366.

Tewhey R., Warner J.B., Nakano M., Libby B., Medkova M., David P.H., Kotsopoulos S.K., Samuels M.L., Hutchison J.B., Larson J.W. 2009. Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat. Biotechnol. 27:1025–1031.

Thornton K. 2003. libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics 19:2325–2327

Vos P., Hogers R., Bleeker M., Reijans M., Lee T., Hornes M., Friters A., Pot J., Paleman J., Kuiper M. 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 23:4407–4414.

Whitelaw C.A., Barbazuk W.B., Pertea G., Chan A.P., Cheung F., Lee Y., Zheng L., van Heeringen S., Karamycheva S., Bennetzen J.L., SanMiguel P., Lakey N., Bedell J., Yuan Y., Budiman M.A., Resnick A., Van Aken S., Utterback T., Riedmuller S., Williams M., Feldblyum T., Schubert K., Beachy R., Fraser C.M., Quackenbush J. 2003. Enrichment of Gene-coding sequences in maize by genome filtration. Science 302:2118–2120.

Williams L.M., Ma X., Boyko A.R., Bustamante C.D., Oleksiak M.F. 2010. SNP identification, verification, and utility for population genetics in a non-model genus. BMC Genet. 11:32–45.

Williams R., Peisajovich S.G., Miller O.J., Magdassi S., Tawfik D.S., Griffiths A.D. 2006. Amplification of complex gene libraries by emulsion PCR. Nat. Methods 3:545–550.

Woerner A.E., Cox M.P., Hammer M.F. 2007. Recombination-filtered genomic datasets by information maximization. Bioinformatics 23:1851–1853.