Syst. Biol. 61(5):745–761, 2012

DOI:10.1093/sysbio/sys051
Advance Access publication on May 18, 2012

High-Throughput Identification of Informative Nuclear Loci for Shallow-Scale Phylogenetics and Phylogeography

ALAN R. LEMMON[1,*] AND EMILY MORIARTY LEMMON[2]

[1]*Department of Scientific Computing, Florida State University, 400 Dirac Science Library, Tallahassee, FL, 32306-4102, USA; and* [2]*Department of Biological Science, Florida State University, 319 Stadium Dr., P.O. Box 3064295, Tallahassee, FL, 32306-4295, USA;*
[*]*Correspondence to be sent to: Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, FL, 32306-4102, USA; E-mail: alemmon@fsu.edu.*

Received 17 November 2011; reviews returned 27 January 2012; accepted 15 May 2012
Associate Editor: Bryan Carstens

Abstract.—One of the major challenges for researchers studying phylogeography and shallow-scale phylogenetics is the identification of highly variable and informative nuclear loci for the question of interest. Previous approaches to locus identification have generally required extensive testing of anonymous nuclear loci developed from genomic libraries of the target taxon, testing of loci of unknown utility from other systems, or identification of loci from the nearest model organism with genomic resources. Here, we present a fast and economical approach to generating thousands of variable, single-copy nuclear loci for any system using next-generation sequencing. We performed Illumina paired-end sequencing of three reduced-representation libraries (RRLs) in chorus frogs (*Pseudacris*) to identify orthologous, single-copy loci across libraries and to estimate sequence divergence at multiple taxonomic levels. We also conducted PCR testing of these loci across the genus *Pseudacris* and outgroups to determine whether loci developed for phylogeography can be extended to deeper phylogenetic levels. Prior to sequencing, we conducted *in silico* digestion of the most closely related reference genome (*Xenopus tropicalis*) to generate expectations for the number of loci and degree of coverage for a particular experimental design. Using the RRL approach, we: (i) identified more than 100,000 single-copy nuclear loci, 6339 of which were obtained for divergent conspecifics and 904 of which were obtained for heterospecifics; (ii) estimated average nuclear sequence divergence at 0.1% between alleles within an individual, 1.1% between conspecific individuals that represent two different clades, and 1.8% between species; and (iii) determined from PCR testing that 53% of the loci successfully amplify within-species and also many amplify to the genus-level and deeper in the phylogeny (16%). Our study effectively identified nuclear loci present in the genome that have levels of sequence divergence on par with mitochondrial loci commonly used in phylogeography. Specifically, we estimated that ~7% of loci in the chorus frog genome are >3% divergent within species; this translates to a prediction of approximately 50,000 single-copy loci in the genome with >3% divergence. Moreover, successful amplification of many loci at deeper phylogenetic levels indicates that the RRL approach represents an efficient method for rapid identification of informative loci for both phylogenetics and phylogeography. We conclude by making recommendations for minimizing the cost and maximizing the efficiency of locus identification for future studies in this field. [Anonymous nuclear loci; Illumina HiSeq; next-generation sequencing; phylogenetics; phylogeography; *Pseudacris*; reduced-representation library.]

One of the major obstacles for phylogeography and shallow-scale phylogenetics is identifying appropriate loci that are informative for the question of interest (Hare 2001). Mitochondrial loci, for example, have been widely used due to the shorter coalescent times of haploid genomes compared with nuclear loci (Neigel and Avise 1986; Hudson 1991) and potentially higher levels of phylogenetic information. Furthermore, mitochondrial loci are recombination-free and have a smaller effective population size. Increasingly, researchers have incorporated nuclear loci into phylogenetic and phylogeographic studies to avoid some of the problems with relying on mitochondrial loci alone (e.g., selective sweeps, lack of clonality, nonneutrality, nonconstant mutation rates, and indirect selection caused by inherited symbionts: reviewed by Hurst and Jiggins 2005; Meiklejohn et al. 2007; Zink and Barrowclough 2008; Avise 2009; Brito and Edwards 2009; Galtier et al. 2009).

The approach taken by many workers to identify nuclear loci usually involves: (i) borrowing loci from another system in which they have been found to amplify successfully (e.g., Spinks et al. 2010; Fijarczyk et al. 2011), (ii) testing new loci generated from the nearest genome or set of genomic resources (e.g., Townsend et al. 2008; Thompson et al. 2008), or (iii) identifying anonymous nuclear loci by sequencing clones from a genomic library (e.g., Jennings and Edwards 2005). Typically PCR followed by Sanger sequencing is used to obtain sequence data, and frequently target loci are selected based on how well they amplify rather than according to how informative they will be for the research question. With the above approaches, loci that are chosen must at least have conserved priming regions across species to allow amplification, before levels of sequencing variation can be assessed. This requirement, in fact, may bias the set of loci used toward those with lower levels of variation and thus affect our perception of the utility of nuclear loci for resolving shallow divergences (Hare 2001; Zhang and Hewitt 2003; Avise 2009; Brito and Edwards 2009).

Here, we use a reduced-representation library (RRL) approach (reviewed by Altshuler et al. 2000; Barbazuk et al. 2005; Van Tassell et al. 2008; Davey et al. 2011)

combined with next-generation sequencing (NGS) of multiple individuals to discover orthologous, single-copy loci that have high levels of sequence divergence within a species and between closely related species. RRLs are created by digesting genomic DNA with one or more restriction enzymes, running the digested sample on a gel, and then selecting the same fragment size range across multiple individuals. RRL sequencing and related methods (e.g., restriction-site associated DNA (RAD) sequencing, Baird et al. 2008) have been used, for example, to successfully identify single nucleotide polymorphisms for various genetic mapping and population genetic applications (e.g., Wiedmann et al. 2008; Kerstens et al. 2009; Emerson et al. 2010). The RRL approach allows the researcher to select a subset of the genome that, in principle, should contain largely the same genomic regions across different individuals; this subset includes both coding and noncoding regions (Altshuler et al. 2000). This approach partially alleviates the constraint of conserved priming regions as short restriction sites (e.g., 6 bp) rather than longer priming sites (e.g., >18 bp) must be conserved to generate data for preliminary estimates of sequence divergence. We utilize paired-end sequencing on the Illumina HiSeq 2000 platform to identify appropriate paired priming regions for subsequent PCR amplicon sequencing. This approach circumvents a substantial problem in phylogenetics and phylogeography by providing a fast method for development of potentially informative loci from the nuclear genome. Utilization of high-throughput paired-end sequencing helps to avoid the problem of allele phasing (Avise 2009, see Discussion section). In addition to developing this new pipeline, we address two questions in this paper: (i) How variable at different taxonomic scales are the loci that can be obtained? (ii) How useful is this approach for deeper phylogenetic scales, such as across species and across genera?

Ideally, we would like to obtain estimates of rates of evolution that could be used to improve locus selection. Estimating rates of evolution is very difficult on a shallow phylogenetic scale, however, because the quantity we can directly measure (sequence divergence) is a function of both the rate of evolution of the locus and the time to coalescence (Hudson 1991). Here, we profile newly identified loci for levels of sequence divergence, while fully acknowledging that sequence divergence may not be a perfect predictor of rates of evolution on shallow time scales. We contend that although the average rate of evolution of the nuclear genome is indeed slower than that of the mitochondrial genome, there is, in fact, a broad distribution of rates within the much larger nuclear genome, suggesting that it may indeed contain a very large number of useful loci for phylogeography. One of the purposes here is to estimate the distribution of sequence variation across single-copy nuclear loci at several time scales. We reserve the dissection of sequence variation of these loci into coalescent stochasticity and evolutionary rate for future studies.

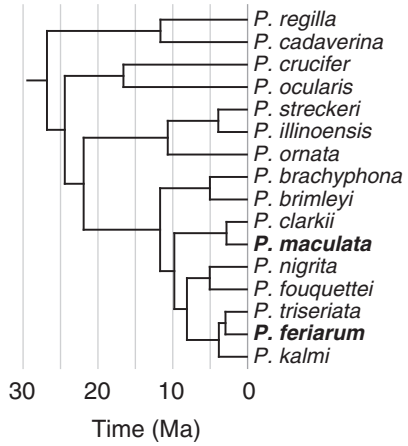

FIGURE 1. Chronogram of the target taxonomic group in this study (treefrog genus *Pseudacris*). One goal was to generate highly polymorphic loci for studying the phylogeography of one species, *P. feriarum*. A second goal was to identify loci to be used across the 17 species in the genus for phylogenetic studies. Species in bold were utilized in the locus identification portion of this study, whereas all species were used in the PCR screen portion of the study. Phylogeny and divergence times are from Lemmon et al. (2007).

METHODS

Overview

The primary motivation for this study was to develop informative nuclear loci for studying the phylogeographic history of the North American chorus frog species *Pseudacris feriarum* (Fig. 1). The two major clades within this species are estimated to share a common ancestor at approximately 2.6 million years ago (Ma; Lemmon et al. 2007; Lemmon and Lemmon 2008). We also desired informative loci that could be used to resolve the phylogeny of the containing genus *Pseudacris*, the basal node of which is estimated at 27 Ma (Fig. 1; Lemmon et al. 2007). Using one individual from each *P. feriarum* clade and one individual from a second species, *P. maculata*, we aimed to obtain a large set of nuclear loci that are homologous across the three individuals, assess levels of sequence variation in the identified loci, and test primers designed from these loci. The general strategy for locus discovery was to sequence the ends of thousands of loci approximately 600 bp in length (this length was chosen based on perceived limitations with respect to the length of fragments that could be sequenced on the Illumina platform). Since we desired also to obtain homologous loci across the three individuals, we prepared RRLs by digesting genomic DNA using restriction enzymes, sorting the fragments by size on a gel, and selecting the same size range for the three individuals. Paired-end sequencing was performed to obtain sequence data for 100 bp of each sequenced fragment. After assembling reads into single-copy loci for each individual, sequences for each locus were aligned across individuals and primers were designed. Primer testing utilized a larger panel containing individuals from all species of *Pseudacris* in addition to several outgroups.

Three features of the experimental design used in this study mitigated the difficulties of developing a new molecular technique in a nonmodel organism: (i) several different restriction enzymes were used during library preparation, (ii) multiple individuals with different degrees of relatedness were sequenced, and (iii) a large amount of sequence data was collected. One of the objectives of this study was to identify which of these elements are necessary so that we can provide recommendations to researchers wishing to use RRLs identify nuclear loci on nonmodel organisms. Because we desired to identify loci in a genus (*Pseudacris*) for which the nearest available genome was quite distant (divergence time with *Xenopus tropicalis* is ~200 Ma; Pyron 2011), and the utility of a restriction enzyme is likely to vary with genomic content, we used eight different restriction enzymes predicted to produce different numbers of loci spanning four orders of magnitude. Since we wished to determine the distribution of sequence variation across nuclear loci at different time scales, we applied the RRL approach to two individuals representing different clades of one species and one individual from another species, separated by ~9.6 million years (Lemmon et al. 2007). Lastly, we collected a large amount of sequence data (three Illumina HiSeq 2000 lanes) to ensure that enough coverage was obtained to yield a large number of loci for which heterozygous genotypes could be accurately identified. In the discussion, we use our results to suggest the most cost-effective techniques for identifying numerous informative loci in other nonmodel systems. A simplified workflow of the method is given in Figure 2.

In Silico Predictions of Restriction Enzyme Utility

Restriction enzymes were chosen based on the results of *in silico* predictions using the *Xenopus tropicalis* genome (Xentr4, v.4.1, August 2005, Joint Genome Institute, http://genome.jgi-psf.org/Xentr4/Xentr4.info.html). After downloading the genome sequence, we estimated the distribution of fragment lengths that would be obtained by separately applying each of 25 different Type II, palindromic restriction enzymes with corresponding restriction sites greater than 5 bp in length (Supplementary Table 1). For each of the restriction enzymes, we predicted the fragment length distribution that would result from 100% digestion of the *Xenopus* genome (Supplementary Table 1). In short, the script performing the *in silico* digest: (i) loads the genomic sequence data, (ii) splits the sequences into smaller sequences at each restriction site, and (iii) computes the lengths of the resulting subsequences. We chose eight restriction enzymes (SrfI, PmeI, SnaBI, NruI, BstZ17I, StuI, EcoRV, and PsiI) that represented the range of distributions observed (Fig. 3) but that also would produce sufficient variation in base composition at each site when pooled. The latter requirement was necessary since the Illumina HiSeq sequencing software requires variation in each of the first few bases for efficient cluster identification. Using these predicted fragment length distributions, we identified a size range, 514–634 bp that would produce a large number of total loci with adequate coverage when the resulting library was sequenced in a single Illumina HiSeq 2000 lane (Supplementary Tables 2–4). Note that the eight enzymes are expected to produce a different number of loci under the chosen size range since the fragment size distribution produced by each enzyme is different. The script used to perform the *in silico* digest experiment is available through Dryad (http://datadryad.org, doi:10.5061/dryad.vh151q1c).

DNA Extraction, Library Preparation, and Sequencing

Three chorus frogs (diploid) were field collected under ACUC protocol #0905, and DNA was extracted from liver tissue using the E.Z.N.A.® Tissue DNA Kit (Omega Bio-Tek, Norcross, GA, USA). The samples included *Pseudacris maculata* (ECM7278; Boone Co., MO, USA), a *P. feriarum* from the Inland Clade (ECM7210; Lafayette Co., MS, USA), and a *P. feriarum* from the Coastal Clade (ECM7144; Prince Edward Co., VA, USA). All three samples were collected from regions of allopatry relative to other closely related species to decrease the likelihood of sequencing heterospecific alleles as a consequence of hybridization.

Each DNA sample was divided into eight aliquots and digested separately with eight different restriction enzymes (EcoRV, NruI, PsiI, StuI, BstZ17I, PmeI, SnaBI, and Srf1). *In silico* predictions (see above) were verified by running digested DNA on an Experion automated electrophoresis system (Bio-Rad). *In silico* predictions were considered verified if the corresponding empirical distribution had a similar shape. Each DNA digest of the three chorus frog samples was performed twice per enzyme. The protocol recommended by the enzyme supplier (New England Biolabs) was followed: 2 μL restriction enzyme, 2 μg DNA, 10 μL 10× New England BioLabs (NEB) buffer, 1.0 μL Bovine Serum Albumin (BSA) (included if recommended for particular enzyme), and enough H_20 to bring total reaction volume to 100 μL. Samples were incubated at 37°C for 2 hours, then held at 4 °C. The two digestions per enzyme were then pooled by enzyme within an individual, purified with the MinElute PCR purification kit (QIAGEN), eluted into 20 μL elution buffer, and stored at 4 °C.

Library preparation and sequencing were performed at HudsonAlpha Institute for Biotechnology in Huntsville, AL, USA. The 24 samples (8 enzymes for 3 individuals) were quantified using both a broad range and high sensitivity assay on a Qubit™ fluorometer (Invitrogen, Carlsbad, CA, USA) and combined in equimolar ratios across enzymes by individual. Paired-end Illumina library preparation was performed on each of the three samples (individuals) separately. The three libraries were then individually size-selected by running samples for 105 min on a 2% 1 × Tris-Acetate-EDTA (TAE) agarose gel, excising a

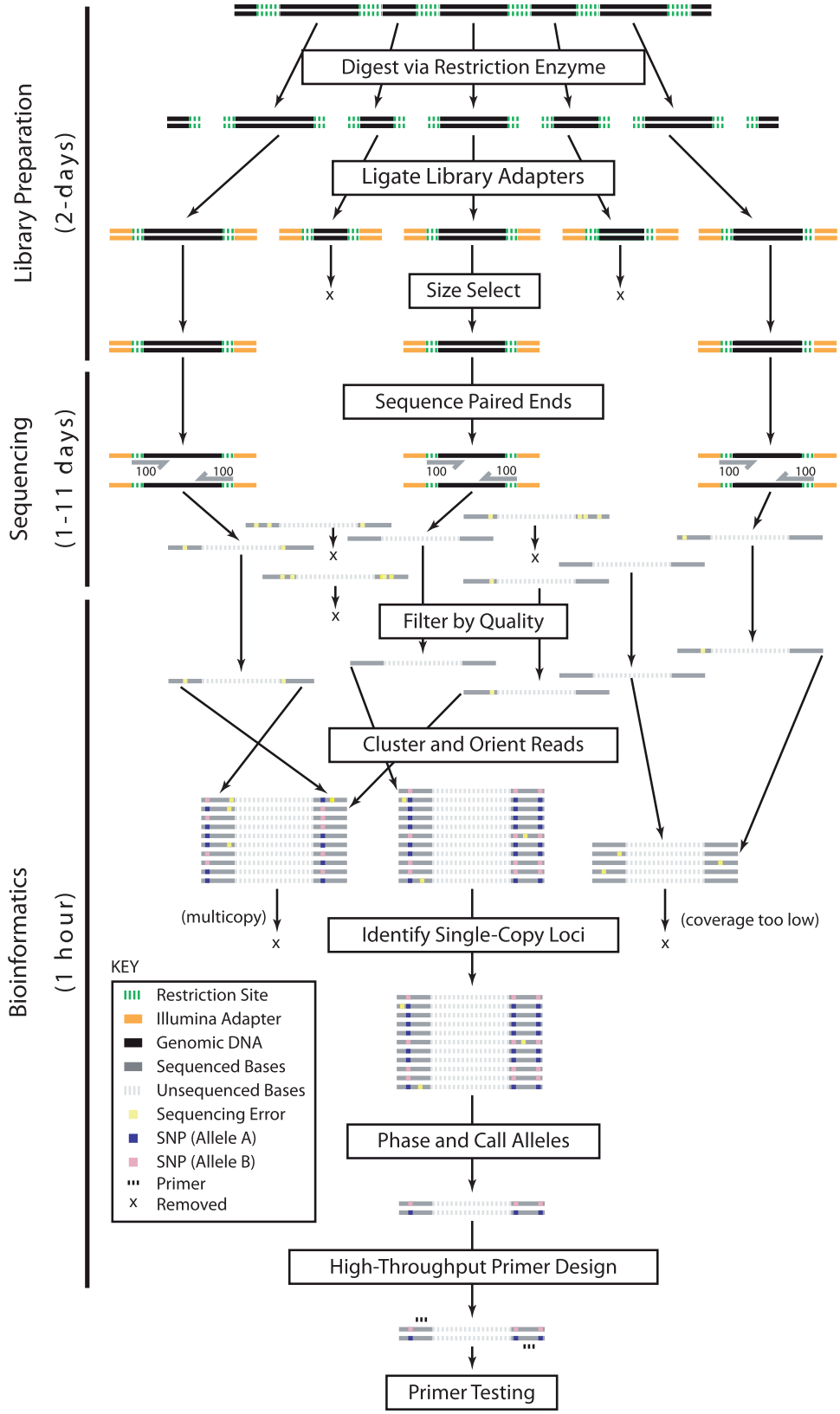

FIGURE 2. Overview of workflow. Thousands of single-copy nuclear loci can be identified in approximately three days. Library preparation includes digestion of genomic DNA using a restriction enzyme, ligation of sequencing adapters, and size selection. Paired-end sequencing produces sequence data for ~100 bp on each end of fragments (sequence between unknown). Bioinformatics consists of removing reads with low-quality base calls, clustering reads based on sequenced ends, orienting reads to account for bi-directional sequencing, identifying single-copy loci with expected allelic proportions after accounting for sequencing error, phasing/calling alleles using read pairs, and high-throughput primer design using Primer3 (Rozen and Skaletsky 2000). The workflow can be adjusted to allow pooling of libraries from different restriction enzymes and/or individuals.

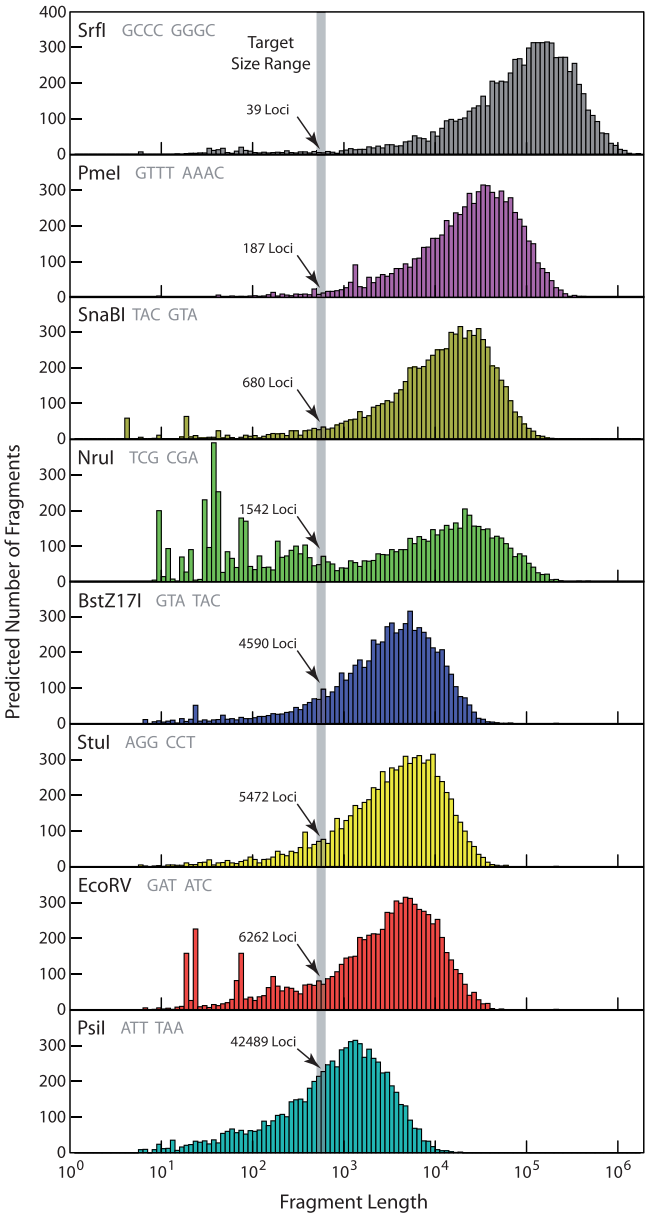

FIGURE 3. *In silico* predictions of fragment length distributions. Restriction sites corresponding to eight different restriction enzymes were identified in the *Xenopus tropicalis* genome. The distances separating adjacent restriction sites for a particular enzyme predict the fragment lengths that would be obtained if genomic DNA were completely digested. The predicted distributions are given assuming each enzyme was used independently. Given a particular distribution, the number of fragments within a target size range can be used to estimate the number of loci that would be sequenced if this range were used during size selection. Predicted numbers of loci are given for each restriction enzyme. More details are given in Supplementary Tables 2–4.

fragment range of 600–720 bp (86 bp of this length was composed of adapter oligos), and performing gel extraction. The three libraries were analyzed on an Agilent DNA 1000 bioanalysis chip to check for proper size range and concentration and diluted to a 10-nM stock for sequencing. Paired 100 bp sequencing was performed on three lanes (one library per lane) on an Illumina HiSeq 2000 sequencing system with v2 chemistry.

Bioinformatics

Prior to assembly, reads were screened for quality and sorted by restriction site. Quality filtering involved removing all pairs containing one or more bases with a corresponding quality score less than 20 (<99% accuracy) in the first 70 bp of either read of the pair. After quality filtering, read pairs were sorted according to the restriction enzyme that generated the sequenced fragment. Sorting was possible since restriction digestion with these particular enzymes produced fragments with half of the restriction site (3–4 bases) on the 5′-end of each sequencing read. Pairs with 5′-ends failing to match one of the eight expected restriction site sequences for both reads were removed from further analysis.

Assembly consisted of clustering read pairs into groups based on the first 70 bp of sequence for the forward and reverse reads. We took advantage of the fact that Illumina reads have a very low incidence of indel error (<0.01%; Glenn 2011) and that complete restriction digestion should produce a set of fragments for each locus that are identical in sequence (excepting for true polymorphisms). The clustering involved two steps. (i) We used a hashing approach (with a kmer size of 70) to cluster read pairs that shared a 70 bp sequence on the 5′-end of one or both of the reads. Since the reads were paired, sequencing errors in one read within a pair would not necessarily result in reads from one locus to be split into two clusters, whereas sequencing errors (or true polymorphisms) in both reads within a pair would lead to splitting (but see Basic Local Alignment Search Tool (BLAST) procedure below). (ii) Pairs were oriented (sequences for read 1 and read 2 swapped) to maximize agreement across read pairs, correcting for the fact that sequencing could have occurred from either direction of a double stranded fragment. Finally, reads within a cluster were aligned by position to produce a raw assembly. Clusters with fewer than 10 reads were removed since a minimum of 10 are necessary to distinguish sequencing error from true polymorphism (Fig. 2). Only 140 bp per read pair were used because preliminary analyses suggested that use of the full 200 bp per pair resulted in SNP calling bias. This bias results from increased estimates of sequence variation as a function of position in the locus (corresponds to position in read, see Results, Sequence Divergence).

The raw assemblies for each cluster were evaluated for evidence that reads originated from more than one locus. For each locus and each site, we computed the likelihood that the observed site pattern resulted from each of six models (two single-locus and four two-locus models): AA, AB, AA/AC, AA/BC, AB/AC, AB/CD, where AA/BC, for example, indicates a two-locus model with the first locus containing one allele and the second locus containing two additional alleles. Figure 4 shows the expected pattern for each model.

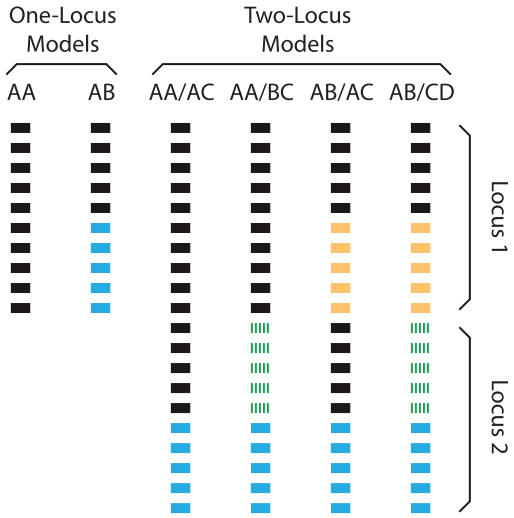

FIGURE 4. Models used to verify the quality of sequence assemblies and identify single-copy loci. Each site of each assembly was evaluated to determine whether the reads overlapping with that site were derived from a single locus or multiple loci. The six models producing distinct distributions of characters are shown, with each bar indicating a different read and each shade representing a different character state (A, T, C, and G). For simplicity, examples include 10 reads per locus, which is the minimum necessary to distinguish among the models. We computed the probability that the observed distribution of characters at a given site was derived from each of these models, allowing for sequencing error at the rate of 0.01 (see main text). Assemblies in which an Akaike information criterion test supported a one-locus model at all sites were identified as single-copy. Equations used to compute the probabilities are given in the pyton script provided on Dryad (http://datadryad.org, doi:10.5061/dryad.vh151q1c).

Our calculations assumed that sequencing error was present at 0.01 (corresponding to the minimum phred quality score allowed, 20, see above) and that allele- and locus-specific biases in coverage were absent. We used these likelihoods to remove clusters for which the best model for any site involved more than one locus, as determined by an AIC test (Akaike 1974). We then removed clusters for which the reads could not be collapsed to less than three alleles based on the set of site-specific models chosen (i.e., site patterns were not consistent across polymorphic sites after accounting for sequencing error). A subset of randomly selected clusters was evaluated by eye to verify that this approach produced reasonable results.

BLAST was used to perform final clustering of alleles into loci. More specifically, we used BLAST v2.2.23+ (Camacho et al. 2009) to perform an all-by-all comparison of alleles both within and between individuals (max e-value = 0.001). The within-individual comparisons allowed divergent alleles to be grouped into the same locus, whereas the comparisons across individuals allowed orthology to be established. All alleles with a significant BLAST match to more than one additional allele within individuals were removed from all subsequent analyses (including across-species analyses). This screen reduced the alleles to those belonging to single-copy loci.

Orthologous alleles were aligned: (i) within each individual, (ii) across all pairwise combinations of individuals, and (iii) across all three individuals in Multiple Sequence Comparison by Log-Expectation (MUSCLE; Edgar 2004) using default parameters. Two measures were taken to ensure accurate estimates of within- and among-species sequence divergence. First, we removed all loci with corresponding alignments that contained one or more gaps (−). Second, we removed all loci for which the two alleles contained any sequence differences in the first or last 3 bp of the corresponding alignment. This was done because preliminary analysis indicated that MUSCLE may occasionally fail to introduce gaps near the ends of alignments, and we did not want to interpret these misaligned sites as polymorphisms. These two stringent measures resulted in unambiguous alignments. Finally, uncorrected pairwise sequence divergence was then computed across alleles within each individual and across individuals for each locus. All alignments with estimates of sequence divergence >3% were visually inspected, and those that showed evidence of error (e.g., due to short inversions or homopolymer stretches) were removed. A Python script performing the read sorting, read clustering, copy-number analysis, allele identification, and allele phasing is available from Dryad (http://datadryad.org, doi:10.5061/dryad.vh151q1c).

Paired primer sequences were selected for each locus using a high-throughput approach. We began by taking the consensus sequence from the within- and across-species alignments for each locus. The stand-alone version of Primer3 (Rozen and Skaletsky 2000) was used to identify the top 10,000 primer pairs for each locus and taxonomic level, requiring each pair to include one primer from each 70-bp side of the alignment (corresponding to the fragment ends). Primers containing ambiguous bases (N or other IUPAC ambiguities) resulting from polymorphism were not allowed. Default parameters were used except for the following adjustments: PRIMER_MIN_SIZE=18, PRIMER_OPT_SIZE=20, and PRIMER_MAX_SIZE=24. From the list of 10,000 identified by Primer3, we chose the primer pair that would produce the longest amplicon, with ties being broken by choosing the primer pair that was listed first in the Primer3 output file.

Because our aim was to identify nuclear loci, we confirmed that the final loci were not derived from mitochondrial DNA. Specifically, we used BLAST (v2.2.23+; Camacho et al. 2009) to compare the final allele sequences with the mitochondrial genome of *H. japonica* (Igawa et al. 2008), the closest species for which a mitochondrial genome was available (divergence time of *Hyla* from *Pseudacris* ~44.7 Ma; Smith et al. 2007). Default BLAST parameters were utilized (max e-value = 10).

Testing Candidate Loci Within Species

To test the utility of the candidate loci within species, we used PCR to amplify 96 of the *P. feriarum* loci using the

primers described above. Loci were binned into groups based on primer melting temperature and tested using an annealing temperature 5°C lower than the lowest melting temperature of any primer in the group. The groups included the following annealing temperatures: 45°, 47°, 48°, 49°, and 51°C. Each locus was tested at a single annealing temperature during the course of this study without any optimization. PCR was performed on one representative from each of the two clades of *P. feriarum* (ECM7145 and ECM7522; Supplementary Table 5). A locus was classified as successful if it amplified in these two individuals and a single band was present in the expected size range. Of the set of successful loci, 20 were Sanger sequenced to verify estimates of sequence divergence across each locus (see below for details of PCR and sequencing reactions; Supplementaary Table 6). Two individuals were Sanger sequenced for these loci—these were the two *P. feriarum* samples from which the RRLs were derived (ECM7144 and ECM7210; Supplementaary Table 5).

Testing Candidate Loci Across Species

To test the utility of the candidate loci across species, we used PCR to amplify 187 loci. For this set of loci, PCR testing involved a multi-round approach. First, PCR was performed on only the two species from which the loci were derived (ECM2731 *P. maculata* and ECM7145 *P. feriarum*; Supplementary Table 5). Second, loci that successfully amplified in the first round were tested in five additional species within *Pseudacris* (ECM2695 *P. regilla* [*P. sierra* sensu Recuero et al. 2006a,b], MHP8258 *P. streckeri*, ECM5914 *P. crucifer*, ECM7199 *P. brachyphona*, and ECM5055 *P. nigrita*). Third, for loci that passed the second round, all 10 remaining species of *Pseudacris* were tested (ECM0151 *P. cadaverina*, ECM0140 *P. regilla* [*P. hypochondriaca* sensu Recuero et al. 2006a,b], ECM4375 *P. illinoensis*, ECM5956 *P. ornata*, ECM7001 *P. ocularis*, ECM0080 *P. brimleyi*, ECM1144 *P. clarkii*, ECM2293 *P. fouquettei*, ECM1064 *P. kalmi*, and ECM7221 *P. triseriata*). Finally, for loci that passed the third round, two individuals from each of two hylid outgroups were tested (LNB326 and ECM3184 *Hyla cinerea*, ECM5835 and ECM5938 *Acris gryllus*; Supplementary Table 5). A positive control (ECM7145 *P. feriarum*) from round 1 was included in rounds 2, 3, and 4 testing.

PCR reactions consisted of 1X Go Taq®Reaction Buffer (Promega), 0.08 mM dNTPs, 0.4 U Go Taq®DNA Polymerase, 0.2 μM each primer, and either 12 ng template DNA in a total volume of 10 μL for screening or 30 ng template DNA in a total volume of 25 μL for generating PCR product for sequencing. Amplification was performed on a Bio-Rad DNA Engine Tetrad® 2 thermal cycler using the following program: 1 cycle of 2 min at 95°C, 35 cycles of 30 s at 95°C, 30 s at the primer specific annealing temp (45, 47, 48, 49 or 51°C), then 1 min at 72°C, 1 cycle of 5 min at 72°C, and then held at 4°C. PCR products were electrophoresed in 1% agarose in 1× TAE for 20 min at 120 V and visualized with a Ultra-Violet Products (UVP) transilluminator. As described for the within-species tests, a locus was classified as successful for the given set of individuals if it amplified a single band in the expected size range. The successfully amplified loci will be sequenced as part of a future study. Primers sequences for all loci tested are available on the Dryad data depository (http://datadryad.org, doi:10.5061/dryad.vh151q1c).

Estimating Levels of Sequence Divergence

Final loci were assessed for levels of sequence divergence at three taxonomic levels: within-individuals, across conspecific individuals, and across species. The Illumina sequencing approach used above produced loci sequenced for 70 bp of each end (with ~430 bp of unknown sequence in between). We used the degree of sequence divergence observed for the sequenced ends of each locus (140 bp per locus) as an estimate of the sequence divergence of the entire locus (~570 bp). The accuracy of this predictor may be compromised by two factors. First, loci obtained across individuals or species using the RRL approach may be a biased sample of all loci in the genome since the sequence of the restriction sites and the distance between restriction sites must be conserved across individuals for the loci to be obtained for all individuals. Since this factor may negatively bias estimates of sequence divergence, we expect our results pertaining to the number of loci with high sequence variation to be conservative. The second factor is rate heterogeneity. The degree of bias produced by this factor is a function of the scale at which rates vary, with the greatest degree of bias expected when rates vary on the scale of the locus length, such that rates in the 70 bp ends of 570-bp loci are substantially different than the 430-bp center. Under these conditions, estimates of rates based on the ends are expected to be positively biased with respect to the degree of variation across loci.

To test for this bias and establish a correction factor, we sequenced the entire length of a subset of the loci (20) for two *P. feriarum* individuals (Coastal clade: ECM7144, Inland clade: ECM7210). Since these are the same individuals used in the Illumina sequencing, we could make a direct comparison between estimates of sequence variation based on 140 bp (the ends sequenced by Illumina) and the full ~570 bp (sequenced by Sanger). We plotted the Illumina estimates against corresponding Sanger estimates, tested for a significant correlation, and used the slope of the best fit linear regression to correct the estimates. Sanger-sequencing involved the following: (i) 20 loci were chosen from across the range of sequence variation, (ii) PCR products for these loci were purified using a QIAquick® PCR Purification Kit and eluted with 50 μL molecular grade H_20, (iii) the concentrations of the cleaned PCR products were adjusted to 10 ng/μL using a NanoDrop® ND1000 spectrophotometer, and (iv) sequencing was performed

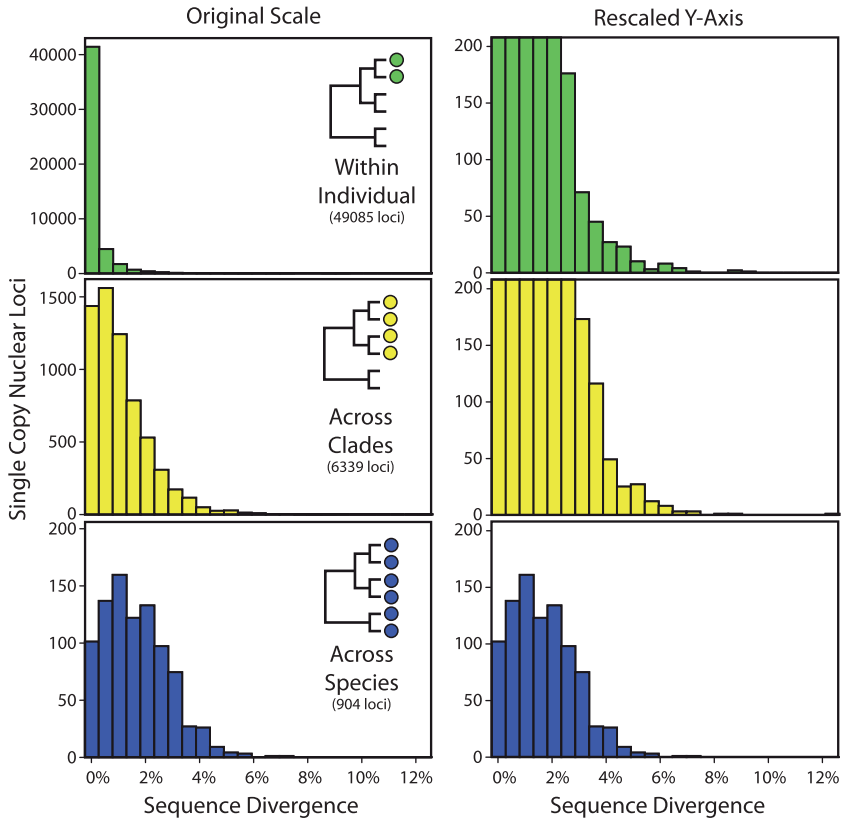

FIGURE 5. Estimates of sequence divergence at three time scales for thousands of loci. Pairwise sequence divergence is shown across alleles taken from: (i) within an individual (*P. feriarum* from the Coastal clade, top row), (ii) across individuals from Coastal and Inland clades of *P. feriarum* (center row), and (iii) across individuals from different species (*P. feriarum* and *P. maculata*; Supplementary Table 5). Note that the sequence divergence estimates for the nuclear loci were not corrected for multiple substitutions due to their short length and are thus conservative. For sequence divergence calculations involving more than two alleles, values were averaged across all pairwise combinations of alleles that spanned the node of interest. Combined results for all eight restriction enzymes are shown here. Distributions for within-individual comparisons of sequence divergence are similar for the two *P. feriarum* and for the *P. maculata* sequenced. Plotting the histograms on a rescaled *y* axis indicates that a large number of loci with substantial sequence variation can be obtained at shallow time scales (upper two rows).

at the Florida State University DNA Sequencing Facility on an Applied Biosystems 3730 Genetic Analyzer using Big-Dye v. 3.1 terminator chemistry. All Sanger sequences were deposited in Dryad (http://datadryad.org, doi:10.5061/dryad.vh151q1c).

RESULTS

Sequence Divergence

More than 100,000 single-copy nuclear loci were identified using the RRL sequencing approach. A large number of loci show high sequence divergence at shallow taxonomic levels. Distributions of sequence divergence across loci (Fig. 5) demonstrate that although average sequence divergence is modest (0.1% within individual, 1.1% across clades within *P. feriarum*, and 1.8% across species), a large number of loci are estimated to have greater than 3% sequence divergence at each of the three levels. We obtained 195, 419, and 146 loci with greater than 3% estimated sequence divergence for within-individual (ECM7144), across-clades, and across-species comparisons, respectively. We found more highly divergent loci within species than we did across species (419 vs. 146), probably due to the fact that we were able to obtain over seven times as many loci for both individuals within a species compared with the three individuals that span two species (6339 vs. 904; Table 1). Recall that the number loci that can be obtained for two individuals is expected to decrease with evolutionary distance between the individuals because restriction sites flanking loci, as well as the number of characters separating them, are more likely to be different among distantly related individuals (see below). Results from comparison of RRL-derived consensus sequences to the *H. japonica* mitochondrial genome demonstrate that none the loci we obtained were of mitochondrial origin.

The RRL approach can be used to obtain nuclear loci with levels of sequence divergence useful for phylogeography. To put the observed levels of sequence divergence for the nuclear loci into perspective, we computed percent sequence divergence for three mitochondrial regions commonly utilized for shallow scale phylogenetics: 12S/16S, COI, and ND2. Sequence

TABLE 1. Number of single-copy orthologous nuclear loci obtained from the three RRL libraries and number of overlapping loci derived from sets of these libraries

	Number of orthologs								
	SrfI	PmeI	SnaBI	NruI	BstZ17I	StuI	EcoRV	PsiI	Total
Individuals in set	GGGC	AAAC	GTA	CGA	TAC	CCT	ATC	TAA	
ECM7144	0 (0)	107 (102)	102 (96)	5 (5)	10,689 (9730)	6481 (5963)	6080 (5626)	30,348 (27,568)	53,812 (49,090)
ECM7210	0 (0)	136 (124)	65 (61)	4 (3)	10,598 (9504)	9374 (6731)	7920 (7168)	31,505 (27,977)	59,602 (51,568)
ECM7278	0 (0)	118 (111)	145 (137)	5 (5)	10,079 (9129)	7056 (6405)	6762 (6145)	30,763 (27,597)	54,928 (49,529)
ECM7144 ECM7210	0 (0)	17 (10)	4 (3)	0 (0)	2001 (1153)	1380 (849)	1393 (866)	6115 (3458)	10,910 (6339)
ECM7144 ECM7278	0 (0)	10 (2)	9 (6)	0 (0)	1272 (413)	889 (360)	817 (341)	3842 (1295)	6839 (2417)
ECM7210 ECM7278	0 (0)	14 (6)	2 (0)	0 (0)	1069 (345)	870 (370)	893 (387)	3543 (1241)	6391 (2349)
ECM7144 ECM7210 ECM7278	0 (0)	3 (0)	1 (0)	0 (0)	402 (124)	344 (146)	325 (135)	1358 (499)	2433 (904)

Notes: Different numbers of loci were obtained from each of the eight restriction enzymes (names of enzymes shown in second row and restriction sites indicated in third row). The number in parentheses in each cell indicates the number of loci for which all alleles could be unambiguously aligned.

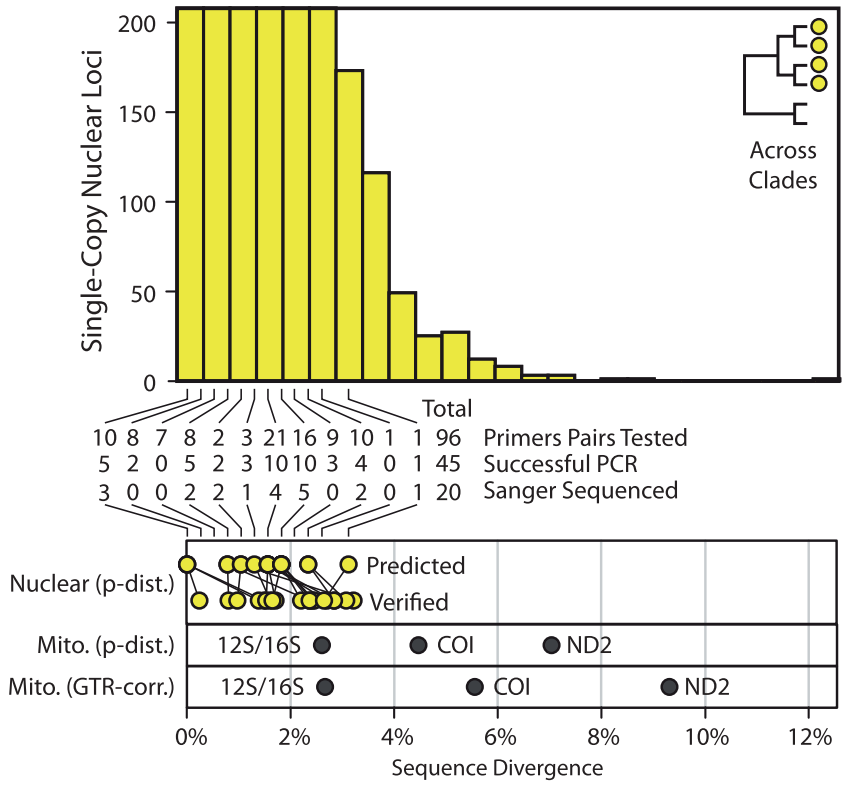

FIGURE 6. Comparison of sequence divergence between nuclear and mitochondrial loci. Estimates of pairwise sequence divergence across alleles from Coastal and Inland *P. feriarum* using Illumina sequencing are presented in the histogram, which is identical to the center right panel of Fig. 5. The number of loci tested and the PCR success for those loci are shown as a table that includes results for loci grouped by initial estimates of sequence divergence. Note that conserved primers could not be found for many of the loci with the highest estimated levels of sequence divergence. Twenty loci were Sanger sequenced to verify estimates of sequence divergence. Relationships between the predicted and verified sequence divergence are shown as lines connecting the points. Three mitochondrial regions commonly used for shallow taxonomic scales (12S/16S, COI, and ND2) were also Sanger sequenced for the same individuals (note that the 715 bp partial 12S/16S sequence was included). Comparison of sequence divergences for the mitochondrial and nuclear loci indicates that many nuclear loci will have phylogenetic utility at shallow taxonomic scales. Note that the sequence divergence estimates for the nuclear loci were not corrected for multiple substitutions due to their short length and are thus conservative. Both corrected and uncorrected estimates of sequence divergence are given for the mitochondrial loci, which are much longer (substitution models assumed and numerical results are given in Supplementary Table 6).

divergence for these loci and all of the nuclear loci sequenced by Sanger sequencing is given in Figure 6 and Supplementary Table 6. A large number of nuclear loci with levels of sequence divergence greater than the mitochondrial regions were identified (Fig. 6). Of the 6339 unverified loci obtained for the two individuals from the different *P. feriarum* clades, we estimated 729, 91, and 6 loci with higher sequence divergence than 12S/16S, COI, and ND2, respectively. For many of the loci estimated to have the highest levels of sequence

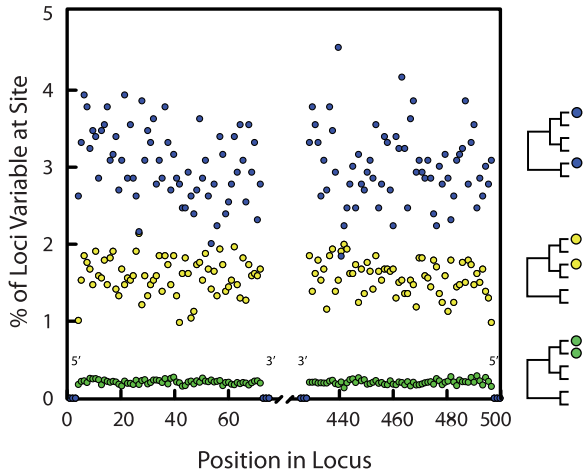

FIGURE 7. Estimates of sequence divergence are not biased by Illumina sequencing errors. Recall that Illumina sequencing error is most pronounced at the 3′-ends of reads (Supplementary Fig. 1). If sequencing error has biased sequence divergence estimates, then sites toward the center of assembled loci (3′-ends of Illumina reads) should have inflated estimates of sequence divergence relative to sites at the ends of loci (corresponding to 5′-ends of Illumina reads). No such pattern exists at any of the three time scales. Instead, the percent of variable loci is consistent across sites. Moreover, the fact that the sequence variation scales with taxonomic depth further confirms that estimates of sequence variation are not an artifact of sequencing error. All loci are invariant at the 12 sites corresponding to 5′- and 3′-ends of reads because loci with variation at these sites were removed to ensure alignment accuracy (see Methods section). Note that the scale on the *x*-axis, which assumes that all loci are 500 bp in length, is used for convenience only. Accounting for unknown variation in locus length would affect the *x*-axis labels but would not change the pattern observed in the plotted points.

variation, however, primers either could not be designed for the locus or did not amplify in both *P. feriarum* individuals (note that degenerate primers were not considered). Of the 20 loci verified by Sanger sequencing, we obtained six with greater sequence divergence than 12S/16S. No Sanger-sequenced loci had greater divergence than COI or ND2 (note that the amount of sequence divergence varies among mitochondrial loci only as a result of among-site rate variation since they are inherited as a single nonrecombining block). Given that the number of loci obtained for the two clades is estimated to comprise only 0.8% of the single-copy *Pseudacris* genome, an estimated 89,112, 11,124, and 733 single-copy loci occur in the nuclear genome that have higher sequence divergence than 12S/16S, COI, and ND2, respectively. Note that these results were generated using estimates of pairwise sequence divergence (not corrected for multiple substitutions).

Estimates of sequence divergence obtained by sequencing the ends of fragments can be used to facilitate the discovery of nuclear loci with ideal levels of sequence variation. We observed a significant correlation between estimates of sequence variation based on 140 bp from the paired-end reads and those based on Sanger sequences of the ~570 bp full length loci ($r = 0.655$, $P = 0.001709$). Linear regression of the Sanger estimates on the Illumina estimates produced a slope of 0.72 when the y-intercept was set to zero. The fact that slope of the regression is <1 suggests that estimates of sequence variation based on the 140 bp are somewhat positively biased, as expected (see Methods section). We have corrected for this bias in all results presented by multiplying the initial Illumina estimates by 0.72 (see Methods section), thereby ensuring that the slope of the corrected estimates is 1.0. All estimates of sequence divergence (Sanger, uncorrected Illumina, and corrected Illumina) for the 20 verified loci are given in Supplementary Table 6.

Estimates of sequence divergence were not biased by sequencing error. Recall that the positions in the locus correspond to the positions in the reads (Supplementary Fig. 1). Because average sequencing error increases from 5′ to 3′ in the reads (Supplementary Fig. 1), estimates of polymorphism are expected to be higher toward the center of the loci if biased by sequencing error. No such trend is observed when the proportion of loci containing a SNP is plotted as a function of locus position (Fig. 7). Instead, the degree of polymorphism is uniform across the positions in the loci, with the exception of the positions corresponding to the 3 bp at the beginning and end of the reads, which show zero polymorphisms. This pattern is expected because we removed loci containing polymorphisms at these positions to ensure alignment accuracy (see Methods section). When we attempted to assemble reads and estimate divergence using all 100 bp of the reads, we did observe some bias. For this reason we based all of our results on the paired 70 bp analyses.

Efficiency of RRL Approach Across Species

RRL sequencing is an efficient method for producing large numbers of loci both within and across species. The number of useable loci decreases steadily with increasing taxonomic depth, as indicated by the results of the 187-locus primer screening across multiple species (Fig. 8). Nonetheless, we were able to obtain 30 loci (16%) for which PCR was successful for all 17 *Pseudacris* species plus at least one outgroup species (*Acris gryllus* or *Hyla cinerea*). The origin of the clade containing these species occurred >30 Ma. A larger number of loci, 34 (18%), are available if only members of *Pseudacris* are required (the origin of *Pseudacris* is estimated to be 27 Ma; Lemmon et al. 2007). Of 96 loci tested within *P. feriarum*, 51 amplified successfully (53%). The number of useable loci corresponding to other clades is given in Figure 8c.

The quantity of PCR screening required to obtain the final 30 loci was not trivial. Using the four-round screening approach (see Methods section), we performed 1675 PCR tests, 1339 of which were successful (80%). The multi-round approach allowed us to avoid 57% of the 3927 possible tests (187 loci × 21 individuals). The overall efficiency of the approach was 32%, which was computed by dividing the number of PCR products that could be used to build a complete data matrix with one outgroup (30 loci × 18 individuals = 540) by the total number of PCR tests performed (1675, see above). If we had tested all 3927 possible tests, the efficiency would

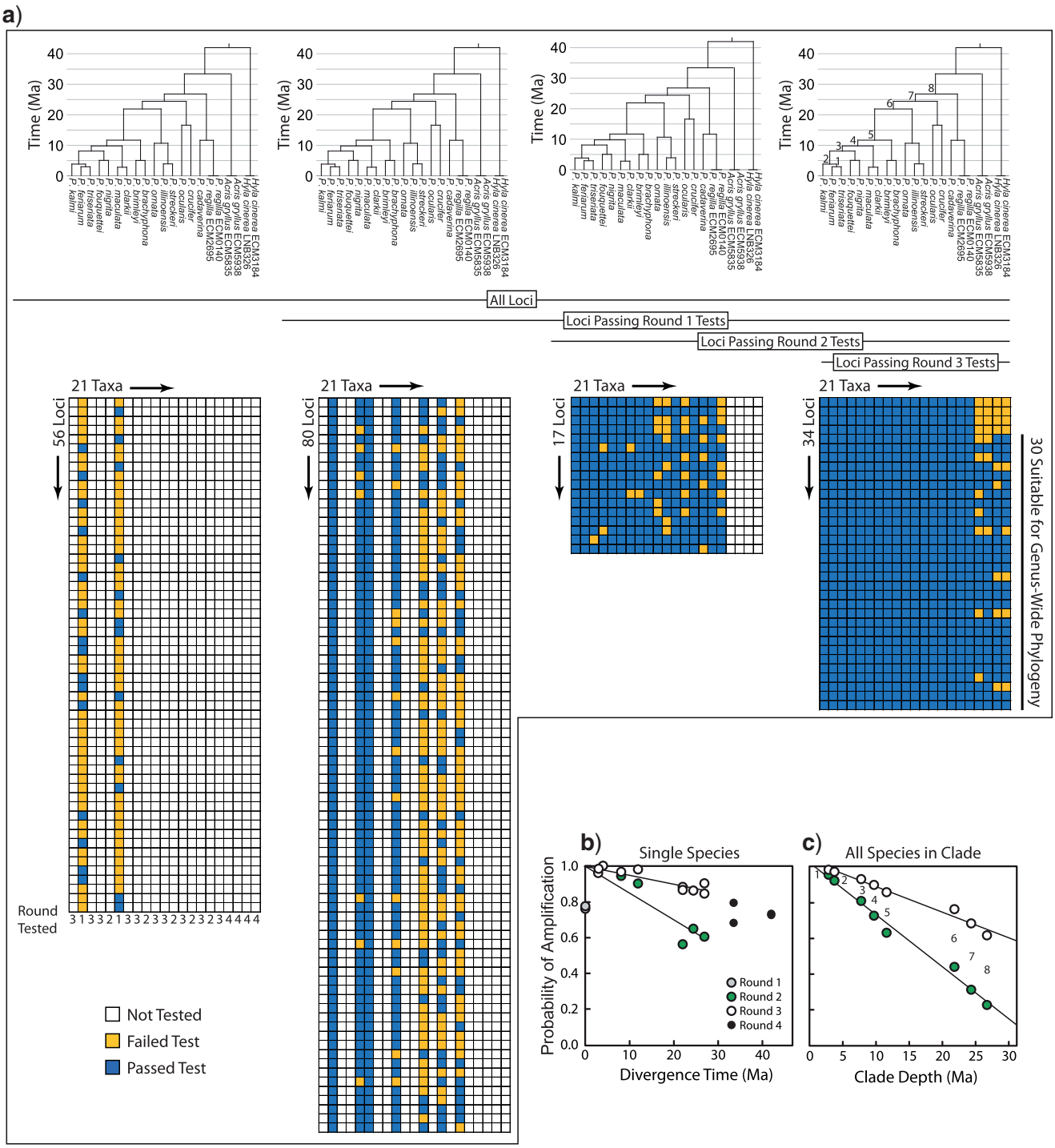

FIGURE 8. Utility of RRL-generated loci with increasing taxonomic depth. a) Primer pairs designed for 187 loci were tested for across-species utility using four rounds, with 2, 5, 10, and 4 species being tested in the four rounds, respectively (see Methods section). Results from these tests are shown for each locus-by-species combination. Primers sequences for all loci tested are available on the Dryad data depository (http://datadryad.org, doi:10.5061/dryad.vh151q1c). The chronogram shown is adapted from Lemmon et al. (2007). b) For tests in each stage, we computed the probability of a successful PCR test for each species (number of successful amplifications/number of loci tested for that species) and plotted the result as a function of the divergence time between that species and *P. feriarum* or *P. maculata* (whichever was closer). c) We also computed for several nested clades within *Pseudacris* (clade numbers on far right chronogram in [a] also correspond to numbered points in [c]) the probability that all species of a given clade would amplify successfully. For each branch within a clade, the probability that amplification would be successful along the branch was computed using the branch length and the relationship obtained in b). These probabilities were multiplied across all branches within the given clade to obtain the overall probability that all species in the clade would amplify successfully.

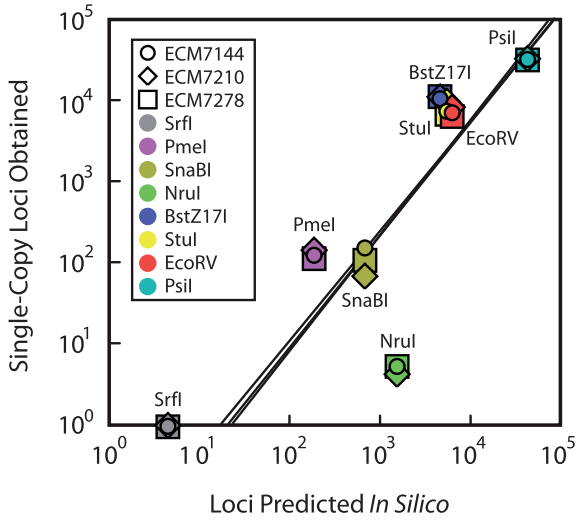

FIGURE 9. Accuracy of *in silico* predictions. The number of single-copy loci obtained is plotted against the number of loci predicted for each individual and restriction enzyme. Predictions were accurate for all restriction enzymes, with the exception of NruI. Note the consistency across individuals with respect to the number of loci obtained. The best-fit regression line, presented for each individual, was computed using log-transformed values for all restriction enzymes except SrfI, for which zero loci were obtained. SrfI is plotted at 1 locus obtained (10^0) for visual purposes only.

have been only 14%. This indicates that the multi-round approach we used was more than twice as efficient as an exhaustive testing approach in which all loci are tested for all species. An alternative approach would have been to test for amplification in two of the most distant related species first (e.g., *Pseudacris feriarum* and *Hyla cinerea*), then to verify amplification in the remaining species. It is difficult to determine the relative efficiency of this approach given the data we collected.

Efficiency of Experimental Design

In silico predictions using the *Xenopus tropicalis* genome provided reasonable estimates of the number of loci for most restriction enzymes. We observed a strong relationship between the number of loci predicted *in silico* and the number of single-copy loci actually obtained (Fig. 9). Restriction enzymes predicted to yield a small number of loci yielded a small number of single-copy loci (e.g., SrfI). The reverse was also true (e.g., PsiI). The exception was NruI, which produced a substantially reduced number of loci (potentially because of methylation sensitivity). The total number of single-copy loci obtained for each individual was surprisingly close to the number predicted (61,261 predicted compared with 53,812 obtained from *P. feriarum* ECM7144, 59,602 from *P. feriarum* ECM7210, and 54,928 from *P. maculata* ECM7278).

Several factors resulted in a somewhat inefficient utilization of sequence reads. The first factor was read quality. Our stringent quality requirements (all bases in first 70 bp of each read were required to have a minimum quality score of 20) resulted in the removal of 45% of reads. Requiring 5′-ends of reads to match restriction enzyme sites also resulted in the removal of a substantial portion of the reads (39%). Together, these two factors resulted in combined loss of 65% of the reads. The other major factor reducing efficiency was the presence of multi-copy loci (e.g., duplicated loci) and repetitive elements (e.g., short interspersed elements [SINEs], long interspersed elements [LINEs]). Preliminary loci (clusters with >10-fold coverage that could represent single-copy loci, multi-copy loci, and repetitive elements) contained 98% of the reads passing quality filters whereas final single-copy loci contained only 31% of the reads passing quality filters. Accounting for all steps in the pipeline, only 11% of the original reads were used to call alleles for the final set of loci. Although these factors did not appear to substantially reduce the number of loci (relative to *in silico* predictions), they did contribute to a substantial reduction in the average per-locus coverage observed: coverage was predicted to be 979 reads per locus, but observed to be ~65 reads per locus (a 93% reduction from the expectation; Supplementary Tables 2–4). Restriction enzyme efficiency, calculated as the number of single-copy loci obtained per million quality-filtered read pairs, varied widely across restriction enzymes, from 0 (SrfI) to 2693 (PsiI). This level of variation suggests that use of some restriction enzymes may be more cost effective than others. Note, however, that this efficiency measure will vary with the nature of the genome and the size range selected during library preparation.

DISCUSSION

High-Throughput Discovery of Nuclear Loci for Phylogeography and Phylogenetics

In this paper, we present a new approach to identifying thousands of nuclear loci for shallow-scale phylogenetics and phylogeography. This approach allows researchers to overcome two major challenges in these fields (Avise 2009): to increase the number of informative single-copy nuclear loci and to efficiently phase alleles. The estimated distributions of sequence divergence suggest that many of these loci will be informative at shallow taxonomic scales (~7% of loci are >3% divergent within species). Moreover, PCR testing of loci at deeper taxonomic scales suggest that the number of loci amplifying across species decreases gradually with increasing taxonomic depth, with >16% amplifying across all species within the target genus. Therefore, we expect these loci to be useful for studies ranging from the within-genus to within-species level of phylogenetic analysis. Additional loci may be obtained by relaxing some of the stringent quality measures we utilized, such as only using loci with alignments that did not contain gaps. Moreover, use of longer reads (e.g., paired-end 250 bp reads expected on the Illumina platform in the third quarter of 2012) may enable primers to be designed for more

of the loci with high levels of sequence variation. Ideally, additional validation of the identified loci would involve demonstrating Mendelian inheritance of alleles, free recombination between loci, low incidence of recombination within loci, and selective neutrality of each locus.

The success of this approach for identifying loci useful for both within- and across-species studies is a result of the experimental design we used. The use of multiple individuals allowed us to: (i) identify conserved primer regions for across-species locus development and (ii) estimate levels of sequence variation to allow flexibility of locus selection for within-species locus identification. Moreover, the use of paired-end sequencing at high coverage allowed us to identify single-copy loci and phase alleles within loci. Although final sequencing of these loci would also utilize a paired-end, high coverage approach, enrichment could be performed using PCR and fragmentation as opposed to restriction digestion and size selection. Alternatively, one could apply applying the RRL-based enrichment approach discussed here to all samples and forego the locus screening and PCR testing procedure. The disadvantage to applying the RRL-based enrichment approach to all samples is the greatly reduced number of loci that could be sequenced with the same coverage as would be obtained with amplicon sequencing. The optimal approach would likely depend on the number of samples in the study, and whether or not a large number of highly polymorphic loci are required.

Utilization of reads shorter than the locus length may, in some circumstances, hinder a researcher's ability to establish allelic phase. Here, we are concerned with determining the sequences of two alleles possessed by a single heterozygous individual, utilizing only reads from that individual (as opposed to sequence data from a population as in PHASE; Stephens et al 2001). To establish allelic phase, two conditions must be met: (i) heterozygous sites must be identified and (ii) the relationships among character states across heterozygous sites must be established. Using the modeling scheme used in this study (and the assumed rate of sequencing error), a minimum of 10-fold coverage is necessary to identify a site as heterozygous. Satisfying the second condition requires that connections are formed between all heterozygous sites (either directly or indirectly) by reads that overlap two or more sites. In the case of single-end sequencing, at least one read must overlap with each pair of adjacent heterozygous sites (more may be desired to accommodate sequencing error). If the distance between any two adjacent heterozygous sites is greater than the length of the reads, then allelic phase cannot be determined. Establishing allelic phase for a locus with a high degree of sequence variation is easier since distances between heterozygous sites are expected to be smaller. When levels of sequence variation are low, longer reads are required so that larger distances between heterozygous sites can be accommodated. Use of information from paired-end sequencing reads, however, can circumvent the need

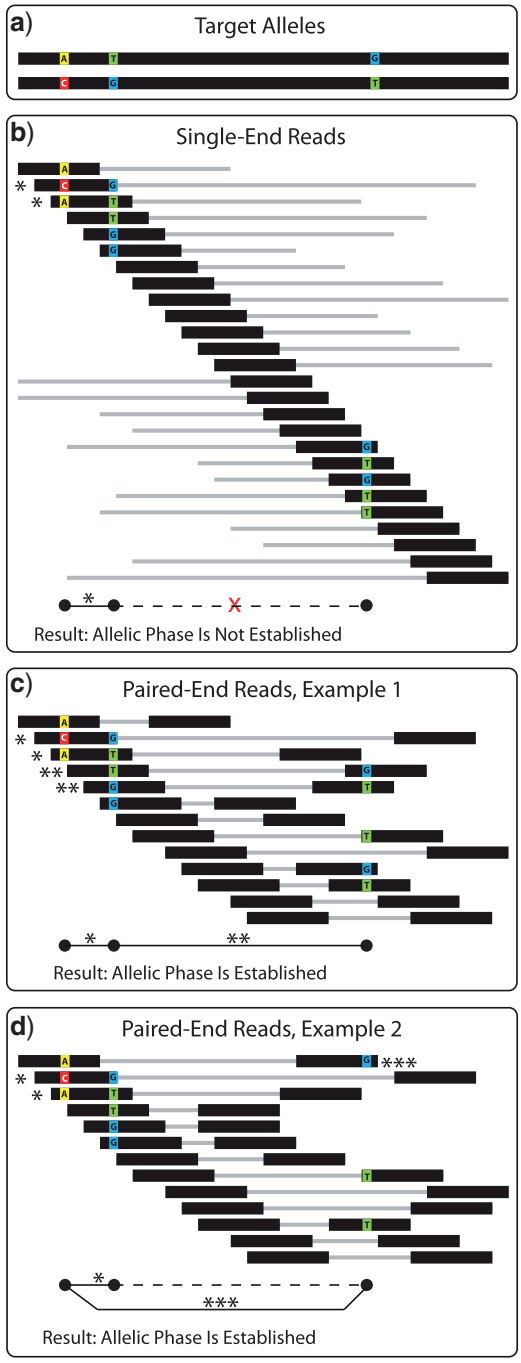

FIGURE 10. Use of paired-end sequencing reads facilitates the determination of allelic phase. In this example, a) an individual possesses two distinct alleles at a given 600 bp locus containing three heterozygous sites. With single-end reads b), allelic phase cannot be determined because the distance between the second and third heterozygous sites (300 bp) is greater than the read length (100 bp). With paired-end reads c, d), however, allelic phase can be determined because paired reads from long fragments can overlap with the distant heterozygous sites. Note that in the case of paired-end reads, it is not necessary for all pairs of adjacent heterozygous sites to be overlapped by paired reads d). Both the single and paired-end examples assume the same fragment length distribution, read length, and coverage. In b–d), black and gray lines indicate sequenced and unsequenced portions of fragments, respectively. Single, double, and triple asterisks indicate reads that enable phase of nearby, distant, and nonadjacent heterozygous sites, respectively.

for long reads (Fig. 10). For paired-end reads, the phase of characters at a pair of heterozygous sites can be established if they are overlapped by either one contiguous read or two different reads from the same pair. If the distance between any two adjacent heterozygous sites is greater than the length of the fragments from which the library was prepared, then allelic phase cannot be determined. In the case of paired-end reads, both read length and fragment length determine the likelihood of establishing allelic phase. In general, libraries prepared from fragments with a broad length distribution are expected to perform best. In Supplementary Figure 2, we present simulation results showing that paired-end reads outperform single-end reads, all else being equal.

Recommendations for Future Locus Identification Using the RRL Approach

The experience gained from this study allows us to make a number of recommendations for researchers to optimize their efforts to develop loci using the RRL approach. An obvious limitation of this approach is that loci must be developed specifically for each study system. The goal of this section, however, is to use what we have learned here to minimize the cost and maximize the efficiency of the locus identification step.

Restriction digestion and in silico *predictions.*—Our experience suggests that fairly divergent reference genomes can be used to estimate number of loci to expect from a particular restriction enzyme and insert size combination (Fig. 3; Supplementary Tables 2–4). Note that nucleotide composition alone is a poor predictor of the number of loci that would be obtained (see Supplementary Figure 3). Our estimates, which were based on *in silico* digestion of *Xenopus tropicalis* (~200 million years divergent from *Pseudacris*), were reasonably accurate (Fig. 9; Supplementary Tables 2–4). Given this result, future studies using this approach can use a smaller number of restriction enzymes. The most efficient restriction enzymes we tested were EcoRV, PsiI, and SnaBI, each of which generated approximately 2000 single-copy nuclear loci per million quality-filtered reads, given the fragment size range we utilized. One restriction enzyme, NruI, produced substantially fewer loci than expected, resulting in only 10 loci per million quality filtered reads (Supplementary Tables 2–4). The low yield for NruI could have resulted from incomplete digestion because NruI is known to be sensitive to methylation. We recommend avoidance of this enzyme because we found it to be highly inefficient for locus identification. Researchers aiming to identify longer loci should confirm our findings with *in silico* predictions using the most closely related species with genomic resources because the properties of the genome may affect efficiency.

One important caveat is that use of a single restriction enzyme may yield low-quality sequence data, unless certain precautions are taken. The reason for this is the fact that current Illumina base-calling software calibrates base calls using the first few bases of the reads (Davey et al. 2011). Libraries with low sequence diversity for these bases (as would be obtained by using a single restriction enzyme) may thus yield low-quality sequence data. We successfully avoided this problem because we pooled libraries derived from eight restriction enzymes. Researchers intending to use a small number of restriction enzymes may circumvent this problem by mixing the RRL library with a library derived from whole genomic DNA. We have successfully sequenced a RRL library generated from a single restriction enzyme by mixing it with a whole genome library at a 75:25 ratio (RRL:genomic DNA; Lemmon and Lemmon, unpublished data).

Our *in silico* predictions of coverage for each locus were overly optimistic. We computed expected coverage by dividing the total number of read pairs expected by the total number of loci expected. For our experimental design, which included pooling fragments produced by eight restriction digests and selecting 600–720 bp fragments, we predicted 979-fold coverage per locus. The actual median coverage observed for single-copy loci was 67, only 7% of the expected coverage (Supplementary Tables 2–4). Three factors contributed to this loss of coverage. The largest factor was the presence of multi-copy loci and repetitive elements. Removal of loci from these two classes resulted in loss of 68% of the reads. This high rate of loss due to these elements is perhaps not too surprising since the genome size of *Xenopus tropicalis*, the model species used to predict the coverage, is only 39% of the size of the *Pseudacris feriarum* genome. Future *in silico* predictions of coverage could be improved through the inclusion of a factor that corrects for differences in genome size between the model and target species. The second largest factor leading to reduced coverage was the presence of low-quality reads. Since we desired accurate estimates of sequence divergence, we used a very stringent quality filter that resulted in the removal of 45% of reads (Supplementary Tables 2–4). Although researchers not interested in levels of sequence variation may be able to increase coverage by applying a less-stringent filter, it may be difficult to account for this factor because read quality may vary across libraries and sequencing runs. The third factor contributing substantially to decreased coverage is the presence of 5′ sequence not matching one of the eight restriction enzymes. Mismatches, which may have resulted from accidental breakage of fragments before library preparation, led to the loss of 39% of reads. Future improvements to the library preparation or use of higher-quality DNA may reduce the loss. Taken together, these three factors resulted in an 89% loss relative to the expected coverage. One final factor may also have contributed to inaccurate estimates of coverage is the fact that size selection produces a peaked distribution of coverage across fragment lengths, whereas our *in silico* analyses assumed a uniform distribution of coverage. This factor resulted in higher than expected variation in

coverage across loci. Despite the inefficient utilization of raw reads, we were able to develop thousands of single-copy loci due to the extremely high-throughput nature of the sequencing platform we utilized.

Library preparation and sequencing.—The library preparation protocol could be improved in at least two ways. First, increasing the width of the size range of the selected fragments may increase the number of orthologous loci obtained across individuals or species. The reason for this is that wider size ranges will tolerate greater variation in locus length (due to the presence of indels). Of course, more reads would be required to obtain the same level of coverage, unless the number of individuals or loci was reduced. Increasing the width of the size range may also have the beneficial effect of decreasing variation in coverage across loci. The second way to improve the library preparation protocol may be to size-select larger fragments, which would produce longer loci. In our laboratory, we have successfully sequenced the ends of fragments up to 900 bp in length using 100-bp paired-end Illumina sequencing without adjustments to the library preparation or sequencing protocols (Lemmon et al. 2012). Other researchers have sequenced fragments up to 2 kb in length (e.g., Li et al. 2009).

The cost of producing nuclear loci can be reduced considerably by making small changes to the sequencing and primer screening strategies we used. The most obvious modification is to use only one sequencing lane. Since the data for this study were collected, Illumina has released new reagents (v3) capable of generating 3× more data per lane. For those planning to use a similar experimental design and genome size, therefore, a single Illumina HiSeq lane should provide comparable coverage. Note, however, that libraries would need to be indexed or barcoded by individual before pooling. Indexing, in which a sample-specific oligo is designed into the adapter and sequenced in a separate sequencing read, would be preferred since use of a barcode (sample-specific oligo read as the 5′ N bases of a normal sequencing read) reduces the number of usable nucleotides for each read. The second modification to reduce costs is to only perform ∼70-bp paired-end sequencing. Although we performed 100-bp paired-end sequencing, we discarded 30 bp from the 3′-end of each read due to concerns about this lower quality region affecting our estimates of sequence variability. The third modification is to sequence fewer individuals. If loci only need to be identified for within-species use, then inclusion of individuals from multiple species may not be necessary. Moreover, if loci with random instead of enhanced levels of sequence variation are needed, then sequencing multiple individuals may be superfluous. Reducing the number of individuals sequenced would allow a larger number of loci to be developed with the same level of sequencing effort but would compromise PCR efficiency since conserved regions may not be easily identified. A sequencing platform producing a smaller number of reads, such as the Illumina MiSeq may also be utilized at a decreased cost per run. The current output of this platform produces approximately twelve million reads. Based on our estimates of restriction enzyme efficiency for EcoRV, for example, this translates into the development of approximately 20,000 loci using one individual and a size range slightly larger than the one we used here. To reduce PCR reagent and labor costs, one could develop multiplex PCR assays to allow amplification of multiple loci in the same reaction (e.g., Meuzelaar et al. 2007). Software and/or services for multiplex design can be obtained from companies such as PREMIER Biosoft (PrimerPlex software).

Interpreting Estimates of Sequence Divergence

One of the goals of our study was to estimate the distribution of sequence variation across nuclear loci. Some care must be taken, however, when interpreting these estimates. The reason is that sequence divergence for a given locus is a function of both the rate of sequence evolution for that locus and the stochastic nature of allele coalescence (Hudson 1991). Since we were working on a shallow time scale and a small number of alleles were used in each comparison, we expect that some portion of the variation observed is due to coalescent stochasticity. Selecting loci with high levels of sequence variation may bias loci toward those that happened to coalesce at deep times, since deeper coalescence will contribute to increased sequence variation. Two approaches can be taken to avoid this bias if the models that will be used to perform downstream analyses require unbiased loci. The first approach is obvious: select loci randomly with respect to sequence divergence. In this way, the loci are a more random representation of the genome, though probably not perfectly random due to the need for conserved primer regions. The second way to reduce these effects is to include more individuals during locus identification. Increasing the number of individuals will allow variation in estimates of sequence divergence that may more accurately reflect relative rates of evolution since a greater number of coalescent events are represented and stochastic effects can be mitigated.

SUMMARY

This study demonstrates a rapid and low-cost RRL-based method for generating a large number of highly variable nuclear loci for shallow-level phylogenetic and phylogeographic studies of any diploid system. This approach is particularly useful to researchers working in nonmodel systems with few or no genomic resources. *In silico* digestion of the nearest reference genome prior to initiation of empirical work provided a reasonable estimate of the number of expected loci generated from this approach and facilitated estimation of the level of sequencing effort required to obtain adequate coverage. When primers for loci designed from the

target species were tested across multiple taxa, PCR amplification success was found to decline with the level of evolutionary divergence, as predicted. Despite this trend, a sizeable number of variable loci were identified that amplify across the entire target genus. This approach, therefore, is expected to be useful both to phylogeographers and to phylogeneticists requiring moderate to large numbers of nuclear loci.

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found at www.datadryad.org at doi:10.5061/dryad.vh151q1c.

Funding

This work was supported by Florida State University new faculty set-up funds to A.L. and E.M.L.

Acknowledgements

We are especially grateful to Lisa Barrow, Hannah Ralicki, and Sandra Emme for conducting the PCR screening and Sanger sequencing for this project. We thank Brian Caudle, Mallory Bedwell, and Sandra Emme for general laboratory support. We are grateful to David Cannatella and Travis LaDuc of the Texas Memorial Museum at the University of Texas, Austin, and to Curtis Schmidt and Joseph T. Collins of the Sternberg Museum of Natural History, Fort Hays State University, for tissue loans. We also thank Brooke Aden, Lisa Barrow, Joseph T. Collins, David Hall, Chris Hobson, Andrew Landis, D. Bruce Means, Moses Michelsohn, Cameron Siler, Courtney Swisher, and Mitch Tucker for assisting with tissue collection.

References

Akaike H. 1974. A new look at statistical model identification. IEEE Trans. Automat. Control 19:716–723.

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. 2009. Phylogeography: retrospect and prospect. J. Biogeogr. 36:3–15.

Baird N.A., Etter P.D., Atwood T.S., Currey M.C., Shiver A.L., Lewis Z.A., Selker E.U., Cresko W.A., Johnson E.A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376.

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.

Brito P.H., Edwards S.V. 2009. Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica 135:439–455.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. 2009. BLAST+: architecture and applications. BMC Bioinform. 10:421.

Davey J.W., Hohenlohe P.A., Etter P.D., Boone J.Q., Catchen J.M., Blaxter M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12:499–510.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Emerson K.J., Merz C.R., Catchen J.M., Hohenlohe P.A., Cresko W.A., Bradshaw W.E., Holzapfel C.M. 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proc. Natl. Acad. Sci. USA 107:16196–16200.

Filarczyk A., Nadachowska K., Hofman S., Litvinchuk S.N., Babik W., Stuglik M., Gollman G., Choleva L., Cogălniceanu D., Vukov T., Džukić G., Szymura J.M. 2011. Nuclear and mitochondrial phylogeography of the European fire-bellied toads *Bombina bombina* and *Bombina variegata* supports their independent histories. Mol. Ecol. 20:3381–3398.

Galtier N., Nabholz B., Glemin S., Hurst G.D.D. 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Mol. Ecol. 18:4541–4550.

Glenn T.C. 2011. Field guide to next-generation DNA sequencers. Mol. Ecol. 11:759–769.

Hare M.P. 2001. Prospects for nuclear gene phylogeography. Trends Ecol. Evol. 16:700–706.

Hudson R.R. 1991. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7:1–44.

Hurst G.D.D., Jiggins F.M. 2005. Problems with mitochondrial DNA as a marker in population, phylogeographic, and phylogenetic studies: the effects of inherited symbionts. Proc. R. Lond. B Biol. Sci. 272:1525–1534.

Igawa T., Kurabayashi A., Usuki C., Fujii T., Sumida M. 2008. Complete mitochondrial genomes of three neobatrachian anurans: a case study of divergence time estimation using different data and calibration settings. Gene 407:116–129.

Jennings W.B., Edward S.V. 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene tree. Evolution 59:2033–2047.

Kerstens H.H.D., Crooijmans R.P.M.A., Veenendaal A., Dibbits B.W., Chin-A-Woeng T.F.C., den Dunnen J.T., Groenen M.A.M. 2009. Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. BMC Genomics 10:479.

Lemmon A.R., Emme S., Lemmon E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61:747–761.

Lemmon A.R., Lemmon E.M. 2008. A likelihood framework for estimating phylogeographic history on a continuous landscape. Syst. Biol. 57:544–561.

Lemmon E.M., Lemmon A.R., Cannatella D.C. 2007. Geological and climatic forces driving speciation in the continentally distributed trilling chorus frogs (*Pseudacris*). Evolution 61:2086–2103.

Li R., Fan W., Tian G., Zhu H., He L., Cai J., Huang Q., Cai Q., Li B., Bai Y., Zhang Z., Zhang Y., Wang W., Li J., Wei F., Li H., Jian M., Nielsen R., Li D., Gu W., Yang Z., Xuan Z., Ryder O.A., Leung F.C., Zhou Y., Cao J., Sun X., Fu Y., Fang X., Guo X., Wang B., Hou R., Shen F., Mu B., Ni P., Lin R., Qian W., Wang G., Yu C., Nie W., Wang J., Wu Z., Liang H., Min J., Wu Q., Cheng S., Ruan J., Wang M., Shi Z., Wen M., Liu B., Ren X., Zheng H., Dong D., Cook K., Shan G., Zhang H., Kosiol C., Xie X., Lu Z., Li Y., Steiner C.C., Lam T.T., Lin S., Zhang Q., Li G., Tian J., Gong T., Liu H., Zhang D., Fang L., Ye C., Zhang J., Hu W., Xu A., Ren Y., Zhang G., Bruford M.W., Li Q., Ma L., Guo Y., An N., Hu Y., Zheng Y., Shi Y., Li Z., Liu Q., Chen Y., Zhao J., Qu N., Zhao S., Tian F., Wang X., Wang H., Xu L., Liu X., Vinar T., Wang Y., Lam T.W., Yiu S.M., et al. 2009. The sequence and de novo assembly of the giant panda genome. Nature 463:311–317.

Meuzelaar L.S., Lancaster O., Pasche J.P., Kopal G., Brookes A.J. 2007. MegaPlex PCR: a strategy for multiplex amplification. Nat. Methods. 4:835–837.

Meiklejohn C.D., Montooth K.L., Rand D.M. 2007. Positive and negative selection on the mitochondrial genome. Trends Genet. 23:259–263.

Neigel J.E., Avise J.C. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In: Nevo

E., Karlin S., editors. Evolutionary processes and theory. New York: Academic Press. p. 515–534.
Pyron R.A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Syst. Biol. 60:466–481.
Rozen S., Skaletsky H.J. 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S., Misener S., editors. Bioinformatics methods and protocols: methods in molecular biology. Totowa (NJ): Humana Press. p. 365–386.
Recuero E., Martínez-Solano Í., Parra-Olea G., García-París M. 2006a. Phylogeography of *Pseudacris regilla* (Anura: Hylidae) in western North America, with a proposal for a new taxonomic arrangement. Mol. Phylogenet. Evol. 39:293–304.
Recuero E., Martínez-Solano Í., Parra-Olea G., García-París M. 2006b. Corrigendum to "Phylogeography of *Pseudacris regilla* (Anura: Hylidae) in western North America, with a proposal for a new taxonomic rearrangement". Mol. Phylogenet. Evol. 41:511.
Smith S.A., Arif S., de Oca A.N., Wiens J.J. 2007. A phylogenetic hot spot for evolutionary novelty in Middle American treefrogs. Evolution 61:2075–2085.
Spinks P.Q., Thompson R.C., Shaffer H.B. 2010. Nuclear gene phylogeography reveals the historical legacy of an ancient inland sea on lineages of the western pond turtle, *Emys marmorata* in California. Mol. Ecol. 19:542–556.
Stephens M., Smith N.J., Donnely P. 2001. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68:978–989.
Thompson R.C., Shedlock A.M., Edwards S.V., Shaffer H.B. 2008. Developing markers for multilocus phylogenetics in non-model organisms: a test case with turtles. Mol. Phylogenet. Evol. 49:514–525.
Townsend T.M., Alegre R.E., Kelley S.T., Wiens J.J., Reeder T.W. 2008. Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: an example from squamate reptiles. Mol. Phylogenet. Evol. 47:129–142.
Van Tassell C.P., Smith T.P.L., Matukumalli L.K., Taylor J. F., Schnabel R.D., Lawley C.T., Haudenschild C.D., Moore S.S., Warren W.C., Sonstegard T.S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5:247–252.
Wiedmann R.T., Smith T.P.L., Nonneman D.J. 2008. SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genetics 9:81.
Zhang D-X., Hewitt G.M. 2003. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol. Ecol. 12:563–584.
Zink R.M., Barrowclough G.F. 2008. Mitochondrial DNA under siege in avian phylogeography. Mol. Ecol. 17:2107–2121.