Syst. Biol. 61(5):727–744, 2012

DOI:10.1093/sysbio/sys049
Advance Access publication on May 17, 2012

Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics

Alan R. Lemmon[1,*], Sandra A. Emme[2], and Emily Moriarty Lemmon[2]

[1]*Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, FL 32306-4102, USA; and* [2]*Department of Biological Science, Florida State University, 319 Stadium Drive, PO 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; E-mail: alemmon@fsu.edu.

Received 1 November 2011; reviews returned 19 January 2012; accepted 7 May 2012
Associate Editor: Bryan Carstens

Abstract.—The field of phylogenetics is on the cusp of a major revolution, enabled by new methods of data collection that leverage both genomic resources and recent advances in DNA sequencing. Previous phylogenetic work has required labor-intensive marker development coupled with single-locus polymerase chain reaction and DNA sequencing on clade-by-clade and locus-by-locus basis. Here, we present a new, cost-efficient, and rapid approach to obtaining data from hundreds of loci for potentially hundreds of individuals for deep and shallow phylogenetic studies. Specifically, we designed probes for target enrichment of >500 loci in highly conserved anchor regions of vertebrate genomes (flanked by less conserved regions) from five model species and tested enrichment efficiency in nonmodel species up to 508 million years divergent from the nearest model. We found that hybrid enrichment using conserved probes (anchored enrichment) can recover a large number of unlinked loci that are useful at a diversity of phylogenetic timescales. This new approach has the potential not only to expedite resolution of deep-scale portions of the Tree of Life but also to greatly accelerate resolution of the large number of shallow clades that remain unresolved. The combination of low cost (∼1% of the cost of traditional Sanger sequencing and ∼3.5% of the cost of high-throughput amplicon sequencing for projects on the scale of 500 loci × 100 individuals) and rapid data collection (∼2 weeks of laboratory time) are expected to make this approach tractable even for researchers working on systems with limited or nonexistent genomic resources. [Anchor regions, anchored enrichment, anchored phylogenomics, highly conserved regions, hybrid enrichment, phylogenetics, phylogeography, sequence capture, ultraconserved elements.]

The field of phylogenetics has achieved substantial progress toward resolving the Tree of Life (TOL), particularly through the coordinated efforts of projects such as the “Assembling the Tree of Life” program (AToL; http://www.phylo.org/atol/) and “Cyber Infrastructure for Phylogenetic Research” (CIPRES; http://www.phylo.org/; Cracraft and Donoghue 2004; Donoghue 2004; Lutzoni et al. 2004; Pace 2009; Parfrey et al. 2010; Soltis et al. 2010; Thomson and Shaffer 2010). Yet, researchers still face several major challenges in the remaining stages of TOL assembly: confirming phylogenetic estimates based on relatively small numbers of markers, filling in the gaps in taxon sampling, resolving more difficult branches in the tree, and combining the results of dozens of independent research groups into a single well-resolved phylogeny (Palmer et al. 2004; Keeling et al. 2005; Bader et al. 2006; Rokas and Carroll 2006; Lane and Archibold 2008; Soltis et al. 2010; Thomson and Shaffer 2010). This final phase represents perhaps the greatest challenge in the process of TOL assembly because it will push the limits of our ability to collect and analyze molecular data in a comprehensive manner. Recent developments in genome research—both in terms of increasing availability of genomic resources and advances in genomic technology—may now provide the bridge spanning the chasm created by these challenges that will ultimately allow the field of phylogenetics to reach the final goal.

Completing the Tree of Life will require a deliberate effort to collect data sets that have greater power to resolve species relationships in difficult biological scenarios. One way to increase power is to sample more loci. Recent studies have demonstrated that tens or even hundreds of nuclear loci may be required to resolve species relationships as a consequence of coalescent stochasticity (e.g., Leaché and Rannala 2011). This phenomenon is especially problematic when branch lengths are short relative to population size, such as in the case of rapid radiations and recent divergences (Maddison and Knowles 2006; Edwards et al. 2007; Huang et al. 2010; Leaché and Rannala 2011; Liu and Yu 2011). The results of these studies suggest, therefore, that clarifying currently unresolved branches may require more intense sampling of loci than is currently available for many clades. This work also suggests that current phylogenetic estimates based on few or no nuclear genes may need to be confirmed with additional data sets containing an adequate number of nuclear genes. A second way to increase power is to increase the number of taxa sampled. Theoretical work has shown that increasing taxon sampling can help to break up long branches, thus improving phylogenetic accuracy (e.g., Zwickl and Hillis 2002). Increasing taxon sampling within species has also been shown to help resolve gene tree discordance due to coalescent stochasticity (Huang et al. 2010). A third way to increase power is to include loci informative at appropriate time scales (Townsend 2007; Townsend et al. 2008; Townsend and Lopez-Giraldez 2010; Townsend and Leuenberger 2011). This aim can be accomplished by either including a large number of loci evolving across a range of different rates or by selecting specific genes evolving at rates that maximize the likelihood of resolving branches at a particular time scale (such as in the case of an adaptive radiation). All three of these strategies for increasing phylogenetic power require access to larger data sets than are currently available for most taxonomic groups.

Here, we present a novel approach for rapidly capturing hundreds of loci that are useful for shallow- and deep-level phylogenetic studies. The new approach leverages existing genomic sequences, modern target enrichment techniques, and high-throughput sequencing to enable very efficient large-scale sequencing for nonmodel species without the need for additional primer development or testing. More specifically, we use enrichment probes in highly conserved anchor regions of vertebrate genomes to capture more rapidly evolving adjacent regions and sequence the resulting fragments using high-throughput sequencing. With this new method that we term *anchored enrichment*, a researcher can conceivably generate data on the order of 500 loci for more than 100 individuals in less than two weeks, from extracted DNA to raw sequencing reads. Furthermore, a project of this scale can be performed at ~1% of the cost of a standard polymerase chain reaction (PCR) and Sanger sequencing-based project of the same magnitude. A highly appealing feature of this new approach is the fact that a simple change in the laboratory protocol (i.e., selection of a different size band during size selection) can produce loci containing the degree of sequence variation appropriate to the taxonomic scale of interest (with longer fragments producing more variable loci). After demonstrating the utility of this new approach for vertebrate phylogenetics, we discuss the potential for its application to nonvertebrate groups.

A similar method to the anchored phylogenomics approach we develop here was also independently developed by McCormack et al. (2012) and Faircloth et al. (2012). Since our manuscript was submitted before publication of these two studies, we reserve detailed comparison of the studies for future work. We note here, however, that our approach is novel in that we target: (1) a broader taxonomic scale (vertebrates as opposed to amniotes), (2) fewer loci (~500 as opposed to ~5000) in order to allow higher-throughput sample processing, and (3) less-conserved anchor regions (highly-conserved regions as opposed to ultraconserved elements) in order to potentially allow greater sequence variation at shallow time-scales. Moreover, our approach may tolerate greater sequence variation in probe regions because we utilize probes representing several lineages and a more densely-tiled probe design.

Methods

The purpose of this study was to design a set of target enrichment probes that could be used on a broad taxonomic scale to capture approximately 500 loci that are informative at a diversity of timescales within the clade of vertebrates. We chose 500 loci because simulation studies indicate this should be a sufficient number to resolve the most difficult nodes (e.g., Leaché and Rannala 2011), yet still allow hundreds of individuals to be pooled in a single sequencing lane with sufficient coverage per locus. Hybrid enrichment (a.k.a. sequence capture) uses oligonucleotides as baits that hybridize to genomic fragments containing the target sequence and allows target loci to be separated from nontarget regions of the genome (Albert et al. 2007; Gnirke et al. 2009). Since capture efficiency depends critically on probe–target similarity and we desired to use the same probe set across divergent species, we maximized the chance of success by targeting highly conserved regions of vertebrate genomes (e.g., Bejerano et al. 2004; Siepel et al. 2005; Stephen et al. 2008) using long, 120-bp capture probes, with dense tiling (Archer et al. 2010). These genomic regions show relatively low variation within the vertebrate clade. Because we desired to develop loci that are also variable at shallow time scales, we aimed to identify regions of high conservation that are immediately adjacent to regions of low conservation. Thus, captured genomic fragments would contain both conserved regions and less conserved regions. In short, we desired probe regions that were 1) highly conserved, 2) flanked by less-conserved sites, 3) highly unique within the genome (i.e., single copy), and 4) widely distributed throughout the genome.

Probe Design

For probe design, we chose five species with sequenced genomes to represent most of the major vertebrate lineages: Human (*Homo sapiens*; Mammals), Chicken (*Gallus gallus domesticus*; Birds), Green anole (*Anolis carolinensis*; Squamates), Western clawed frog (*Xenopus tropicalis* [*Silurana tropicalis*]; Amphibians), and Zebrafish (*Danio rerio*; Fish). We refer to these species as the five model species and often refer to them by their genus name only. Note that although some genome sequences (e.g., turtles, salamanders) could not be included because they did not exist when the initial probes were designed, they could easily be used to increase lineage representation in future refinements to the probe design (see Discussion). We downloaded genomes of the model species from the UCSC Genome Browser website (genome.ucsc.edu, Fujita et al. 2011): *Homo* (hg19, February 2009, Human Genome Consortium, 2001), *Gallus* (galGal3, May 2006), *Anolis* (anoCar1, February 2007), *Xenopus* (xenTro2, August 2005), and *Danio* (danRer6, December 2008). We also downloaded the multiz46way vertebrate alignment from the UCSC Genome Browser website (hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/), which used the aforementioned genome assemblies (in addition to other vertebrate genomes). Methods used to generate the alignment are given at (http://genome.ucsc.edu/cgi-bin/hgTrack

Ui?db=hg19&g=cons46way). Unless otherwise stated, downstream bioinformatic analyses were conducted using scripts written in Java or R by ARL. Scripts can be downloaded from Dryad (http://datadryad.org, doi:10.5061/dryad.r606d128) and at http://www.anchoredphylogeny.com.

A total of 512 target loci were chosen on the basis of two properties: conservation and uniqueness. Using the multiz46way vertebrate whole-genome alignment subsampled to include only the five model species, we employed a sliding window approach and computed metrics relating to sequence conservation and uniqueness across the genomes of the five model species. Loci chosen through this filtering process have three following properties: 1) the 240-bp center of each locus (the probe region) is unique in the genomes of the five model species, 2) sites within 700 bp of the locus center do not contain highly repetitive elements, 3) the incidence of indels in the probe region is low, 4) the sites in the probe regions are highly conserved (although some sequence divergence does exist), and 5) the sites in at least one flank of each probe region are not highly conserved. Specific details are given in the Supplementary Materials (http://datadryad.org, doi:10.5061/dryad.r606d128.). Note that greater than 512 loci could have been obtained with relaxed filtering stringency.

Probes were tiled across each of the probe regions. For each of the five model species, a new 120-bp probe began every 5 bp, producing a maximum of 25 probes per species per locus (due to indels, some species had fewer than 25 probes in a given locus). Probe sequences were combined across species to produce a single combined probe set. A total of 56,664 probes were included in the final probe design. In total, the probes occupy approximately 122,800 bp of the genome of each species.

Sampling

Ten species were included in the anchored enrichment component of the study, including the five model organisms described previously and five nonmodel species from the same five vertebrate clades. Divergence times between each pair of model and nonmodel species spanned the range from 94 million years ago (Ma) to 254 Ma (www.timetree.org, Hedges et al. 2006). This paired design permitted assessment of the efficiency of capture in nonmodel species across an array of divergence times from the nearest model species. Along with the model species listed previously, the following non-model species were also included: House mouse (*Mus musculus*; Mammals), Lance-tailed manakin (*Chiroxiphia lanceolata*; Birds), Eastern diamondback rattlesnake (*Crotalus adamanteus*; Squamates), Upland chorus frog (*Pseudacris feriarum*; Amphibians), and Least killifish (*Heterandria formosa*; Fish; Supplementary Table 1). Although mouse is technically a model species, the genome was not used for probe design and is therefore included in the "nonmodel" category here.

Library Preparation, Enrichment, and Sequencing

Three indexed libraries corresponding to three different insert sizes were prepared for each species using a protocol modified from Meyer and Kircher (2010). Indexes were included as part of the Illumina adapters and were sequenced in a separate indexing read. Details are given in the Supplementary Materials. Libraries of each insert size were pooled across species, resulting in three separate pools (one consisting of 375-bp, a second of 575-bp, and a third of 775-bp insert sizes). Each pool was enriched using an Agilent Custom SureSelect kit containing a single pool of all probes for all species (see Supplementary Materials for details). To compare the efficiency and quality of data produced by current Illumina instruments to assess their utility for future anchored enrichment experiments, high-throughput sequencing was performed on both the Illumina MiSeq Personal Sequencing System (375-bp insert size multispecies pool only) and the Illumina HiSeq 2000 sequencing system (all three multispecies pools were combined in a single lane). Paired-end 150-bp reads were produced on the MiSeq, whereas paired-end 100-bp reads were produced on the HiSeq 2000. Sequencing included an 8-bp indexing read (index sequences are given in Supplementary Table 2). Note that although we performed Illumina sequencing here, other current or emerging high-throughput sequencing technologies could be used as well.

Read Processing and Assembly

Raw sequencing reads were processed in the following four ways to ensure quality of downstream results. 1) Low-quality reads were removed. 2) Reads with corresponding 8-bp index sequences not matching exactly with one of the 30 expected indexes were removed. Recall that each index sequence differed by a minimum of two base pairs from all other index sequences used. Two-fold degeneracy of indexes is necessary to avoid an excessive number of misidentified reads (Meyer and Kircher 2010). 3) Reads with evidence of overlapping sequence were merged into a single read. 4) PCR duplicates were removed. Details of these analyses are given in the Supplementary Materials.

Reads from four of the species (*Homo*, *Gallus*, *Danio*, and *Mus*) were mapped to their respective reference genome using Bowtie (version 0.12.7; Langmead et al. 2009). *Xenopus* and *Anolis* were not mapped due to lower quality nature of their reference genomes. Bowtie indexes were either obtained from the Bowtie website (i.e., hg19, mm9, galGal3; http://bowtie-bio.sourceforge.net/index.shtml) or built from the genome sequence obtained from the UCSC Genome Browser website (i.e., danRer6; genome.ucsc.edu, Fujita et al. 2011). Default parameters were used except that only uniquely mapped reads were retained (m = 1). Mapping results were used to confirm enrichment of the target regions and compute genome-wide and within-locus coverage distributions for each of the four species.

These mapping assemblies also allowed verification of the assemblies produced by the quasi-de novo approach described later.

Reads corresponding to each of the ten sequenced species were assembled separately using a quasi-de novo assembly approach that we refer to as the low-sensitivity approach. In short, reads were first mapped to probe region sequences from the five model species using SeqMan NGen 3 (DNASTAR, Inc.). Reads that did not map but were paired with reads that did map were then included in the assembly for each locus. Finally, read positions within the assembly of each locus were adjusted to maximize the agreement across reads. After preliminary analysis, we found this quasi-de novo approach to be better than a strict de novo approach (as implemented in NGen) because it requires at least one read in each pair to map to the single-copy probe region and thus reduces the number of assembly errors caused by repetitive elements that may exist outside of the probe regions. Additional details of the quasi-de novo approach are given in the Supplementary Materials.

The number of captured loci was estimated using the following high-sensitivity analysis that tolerates high levels of sequence divergence: 1) slide each read sequence past the probe region forward and reverse-complement sequences for each locus, 2) at each position determine the number of matches between the read and probe region sequences, 3) determine the maximum number of matches observed for the read, and 4) consider each locus captured if more than 10 reads had greater than 55 matches (maximum matches possible equaled the read length of 100). The two thresholds were chosen based on preliminary analysis with the goal of minimizing false positives without substantially compromising sensitivity (see Supplementary Fig. 1). The number of captured loci estimated using this high-sensitivity approach was compared with the number estimated using the low-sensitivity approach.

Recognizing that all downstream analyses rely on accurate assemblies, we subjected all quasi-*de novo* assemblies to the following four quality control measures (details given in the Supplementary Materials). 1) We identified a locus for a particular species as successfully captured only if the corresponding assembly contained more than 60 total reads. This measure was taken to avoid difficulties arising from having a nontrivial proportion of mis-indexed reads (this problem occurs when sequencing errors cause one of the index sequences used to be converted to one of the other index sequences used in the study), a situation that may occur when coverage for a locus varies substantially across samples. For example, suppose that the number of captured fragments originating from a particular locus for two different samples was 10,000 and 10. Given that mis-indexed reads occur at a frequency of 0.0003 when 8-bp indexes differ by two sites (Kircher 2011), the second sample would contain 13 reads, 3 of which actually originated from the first sample but were incorrectly identified as originating from the second sample (mis-indexed). 2) Assemblies were treated with an automated assembly refinement (cleanup) step designed to remove obvious sequencing errors and aid in efficient manual inspection. 3) All assemblies were inspected by eye using Geneious Pro (v5.5.1; Drummond et al. 2010) and classified according to assembly quality (0: excellent, no manual adjustment required; 1: very good, minor manual adjustment required; 2: Good, moderate manual adjustment required, and 3: Poor, probably not possible to resolve errors). Numbers of loci assigned to each category are given for each species in Table 1. 4) Assemblies corresponding to loci used in phylogenetic analyses (see later) were manually adjusted in Geneious Pro to ensure the quality of the final consensus sequences. These adjusted assemblies are referred to as the final assemblies.

Consensus sequences were then obtained from each of the final assemblies, with each heterozygous character being given the appropriate ambiguous base code. We did not attempt to phase alleles in sequences with heterozygous sites for two reasons: the occurrence of heterozygous characters in the final trimmed alignments was low (<0.24%, see Supplementary Tables 3 and 4) and the set of taxa are deeply divergent. We do expect, however, that the paired nature of the sequencing strategy

TABLE 1. Summary of assembly quality scores across loci. Each locus was assembled separately for each species using a quasi-*de novo* approach (Low Sensitivity, see text for details), manually inspected, and scored for quality. Assemblies thought to require no, minor, substantial, and insurmountable manipulation/adjustment were given scores of 0 (Excellent), 1 (Very Good), 2 (Good), and 3 (Poor), respectively. Loci deemed to be not captured in the low-sensitivity analysis (LS) were given a score of 4. The table presents for each species the number of loci (out of 512) given each score. The number of loci captured was also estimated using a second, high-sensitivity approach (HS; see text for details)

Analysis	Quality	Description	*Anolis*	*Gallus*	*Homo*	*Xenopus*	*Danio*	*Crotalus*	*Mus*	*Chiroxiphia*	*Pseudacris*	*Heterandria*	Total, *N* (%)
Low-Sensitivity (quasi-de novo)	0	Excellent	304	278	362	349	320	163	206	293	55	27	2357 (46.0)
	1	Very good	169	176	116	137	142	80	106	128	35	14	1103 (21.5)
	2	Good	34	52	29	24	47	16	34	32	18	22	308 (6.0)
	3	Poor	3	6	1	2	1	1	5	1	2	6	28 (0.5)
	4	Not captured	2	0	4	0	2	252	161	58	402	443	1324 (25.9)
	0–3	Captured LS	510	512	508	512	510	260	351	454	110	69	3796 (74.1)
High-Sensitivity	N/A	Captured HS	511	512	509	512	510	464	481	508	337	294	4638 (90.6)

will greatly facilitate the phasing of alleles when this approach is applied at shallow time scales and allele phasing is critical. Consensus sequences were deposited in Dryad (http://datadryad.org, doi:10.5061/dryad.r606d128).

Alignment of Loci

We performed phylogenetic analyses on two different data sets, taking a conservative approach by analyzing only the loci with the best quality assemblies. The first data set includes 32 loci for the eight tetrapod species (*Xenopus*, *Pseudacris*, *Homo*, *Mus*, *Gallus*, *Chiroxiphia*, *Anolis*, and *Crotalus*) plus *Danio* as an outgroup. The second data set includes 123 loci for the six amniote species (*Homo*, *Mus*, *Gallus*, *Chiroxiphia*, *Anolis*, and *Crotalus*) plus *Xenopus* as an out-group. For each of the two data sets, the relevant consensus sequences for each locus were aligned using MUSCLE (Edgar 2004), with default parameters as implemented in Geneious Pro (v5.5.1; Drummond et al. 2010). Alignments were manually inspected, and all ambiguous regions or missing sites (i.e., "N" or "-") were denoted as character sets using MacClade 4.08 (Maddison and Maddison 2005) and excluded from phylogenetic analyses, thus the alignments contained 0% missing data. The loci were designed to be conserved in the probe region and less conserved outside the probe region, on average. Thus, generally the entire probe region and as much of the adjacent regions on both sides as could be reliably aligned were included in a single character block and used in the downstream phylogenetic analyses.

Phylogenetic Analyses

We performed two types of phylogenetic analyses on each data set, a Bayesian concordance analysis and a species tree analysis. We began by identifying the most appropriate model of sequence evolution for each of the loci in the two data sets using an AIC test implemented in MrModelTest (version 2; Nylander 2004), which is based on ModelTest (Posada and Crandall 2001). Assuming the chosen model for each gene, we estimated the Bayesian posterior distribution of trees for each gene separately using MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003), with a *samplefreq* = 100, *stopval* = 0.01, *diagnfreq* = 10,000, *nrun* = 4, and *nchain* = 2. Using the resulting posterior estimates, we performed a Bayesian concordance analysis using BUCKy (Ané et al. 2007). To test the sensitivity of the concordance estimates due to the prior on discordance among loci, we performed the analyses assuming $\alpha =$ 0.5, 1, 2, 4, 8, 16, 32, 64, and 128. All other parameters were set at default values. The results from the MrBayes and BUCKy analyses and corresponding alignments were deposited in Dryad (http://datadryad.org, doi:10.5061/dryad.r606d128) and TreeBase (http://purl.org/phylo/treebase/phylows/study/TB2:S12669).

We also estimated the species tree for each multilocus data set using BEST (Edwards et al. 2007; Liu 2008). Each of the two data sets was partitioned by locus. Topologies, branch lengths, and other model parameters were unlinked across loci. Each analysis included two chains for each of the two runs. Sample frequencies of 2000 and 5000 were used for the amniote and tetrapod analyses, respectively. Chains were run until gene tree topologies and species tree topologies converged, as assessed by the comparison of likelihood scores, *LnJointGenePr*, and other model parameters across the four runs in R (R Development Core Team 2011). The amniote and tetrapod data sets required 20 million and 15 million generations to reach convergence, respectively. Samples collected before convergence were discarded as burnin. Species tree estimates for the amniote and tetrapod data sets were derived from 30 and 15 million post-burnin generations, respectively. Species trees and corresponding alignments were deposited in TreeBASE (http://purl.org/phylo/treebase/phylows/study/TB2:S12669).

Results

Phylogenetic Utility

Phylogenetic utility of the anchored enrichment approach is very good on intermediate to deep timescales: we were able to capture a substantial number of loci for both the model and nonmodel species (>90% of all targeted loci were captured), and subsets of the captured loci were sufficiently informative to produce well-resolved phylogenies (Fig. 1). The number of captured orthologs shared by all sampled amniotes, tetrapods, and vertebrates decreased, as the evolutionary distances among the species increased. A total of 440, 321, and 235 orthologous loci were shared by amniotes, tetrapods, and vertebrates, respectively. Higher numbers of orthologous loci were shared among less divergent groups (Fig. 1a). As expected, the number of captured loci of a species is a function of the divergence time between that species and the nearest model relative (Fig. 1b). Note that 511 of the 512 *Homo* probe regions had orthologs in *Mus* (identified using the lift over tool at the UCSC genome browser website), confirming that failure to capture loci was not a result of gene loss in *Mus*, but instead was likely due to sequence divergence at those loci.

The number of loci captured was sufficient to produce highly resolved species trees. The tetrapod phylogeny estimated using 32 genes in BEST was well supported on all branches, despite evidence of gene tree discordance (Fig. 1c, Supplementary Figs. 5 and 6). The branch at the base of the Aves/Squamata clade, for example, received 100% species tree support, despite a sample-wide concordance factor of 0.68. The amniote phylogeny, based on 123 genes, was also highly supported for all branches (Fig. 1d). Branch lengths and support values were comparable across the two

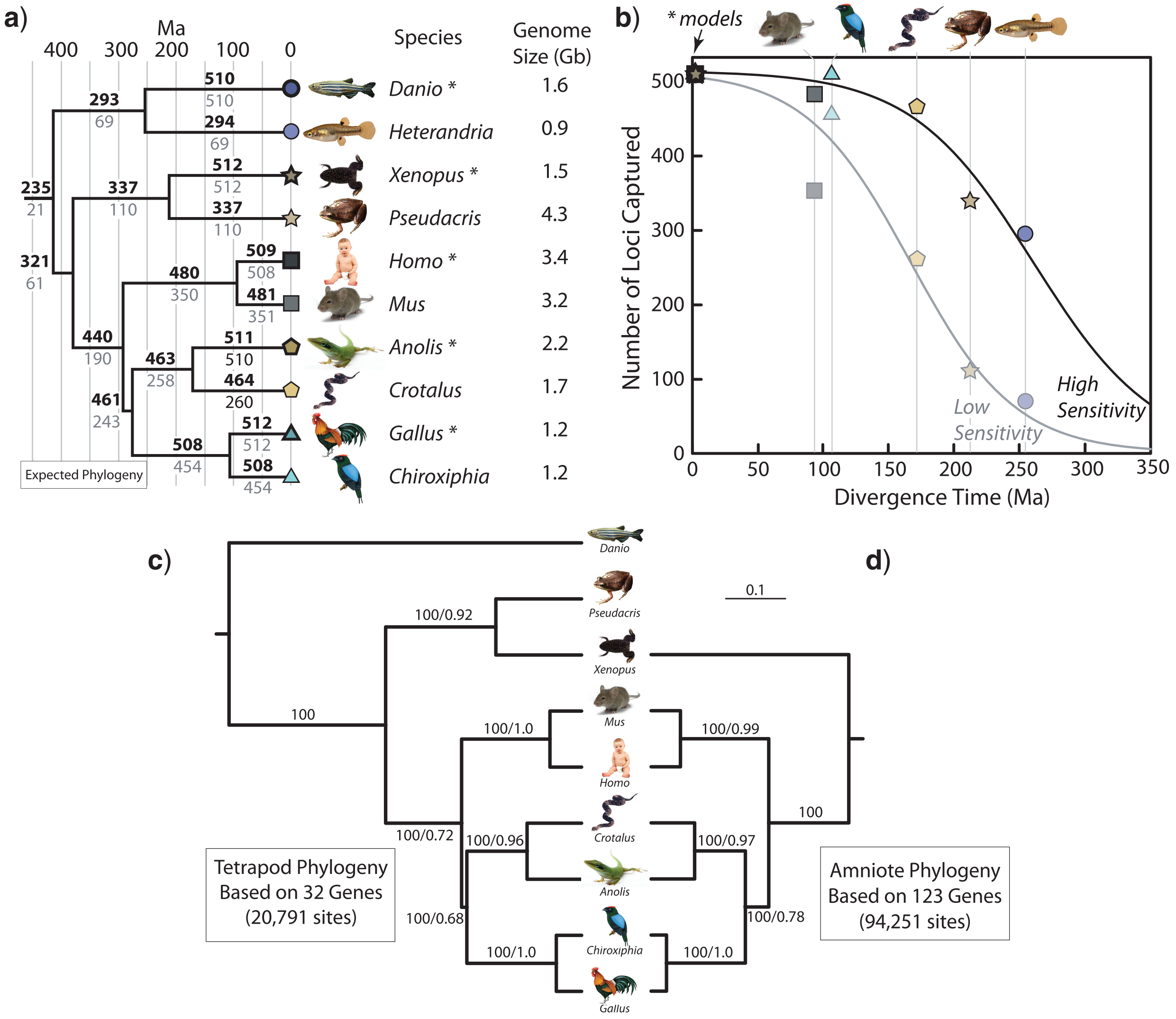

FIGURE 1. Deep-scale phylogenetic utility. Enrichment probes were designed using genome sequences from five model species (indicated by asterisk). (a) The number of loci captured is given for each taxon and the number of orthologous loci shared among taxa is given for each clade on the expected phylogeny (topology based on Tree of Life Project, http://tolweb.org/tree/; branch lengths based on Time Tree divergence times, http://www.timetree.org; Hedges et al. 2006). Numbers above and below branches are estimates from the high and low-sensitivity analyses, respectively. Estimates of genome size were obtained for the most closely related species to each of the target taxa from genomesize.com (Gregory et al. 2006). (b) The number of loci captured for a species is correlated with the divergence time between the species and the closest model species. Best-fit logistic regression lines are based on the ten species sequenced: High Sensitivity, $P < 0.00001$, $y = 512/[1 + e\hat{}[-5.5972317 + x*0.02151495)]$; Low Sensitivity, $P < 0.0001$, $y = 512/[1 + e\hat{}(-4.1491567 + x*0.02491859)]$. Species trees were estimated by BEST for c) a tetrapod data set containing eight ingroup and one outgroup species for 32 loci (20,791 sites total), and (d) an amniote data set containing six ingroup and one outgroup species for 123 loci (94,251 sites total). Note that both data sets had 0% missing data because sites with indels (-) and unknown bases (N) were excluded before analysis. Support values on each branch (c and d) indicate the bipartition posterior probability from the BEST species tree (left value on each branch) and the sample-wide Bayesian concordance factor from BUCKy. Branch lengths are proportional to divergence times, but the overall scale is arbitrary.

phylogenies. Despite low taxon sampling, both species trees were congruent with expected topologies (e.g., http://tolweb.org/tree/). Note that *Heterandria* was not included in the vertebrate phylogeny because only one outgroup is allowed in BEST, and we did not sample taxa outside of the vertebrate clade.

Phylogenetic utility of the anchored enrichment approach is also expected to be good at shallow timescales. Analysis of the genomic regions containing each of the 512 target loci indicate that probe regions tend to be in conserved regions (predominantly in exons), whereas adjacent regions lie in less conserved regions (e.g., introns). Figure 2 presents relative coverage, conservation scores, and genomic annotation across the positions of the loci. Plotting the relative coverage for different insert sizes (Fig. 2a) produced the expected pattern: longer insert sizes allows longer loci to be obtained. Loci ~900 bp in length can be obtained using a 375-bp insert size whereas loci ~1500 bp can be obtained using a 775-bp insert size. Based on previous

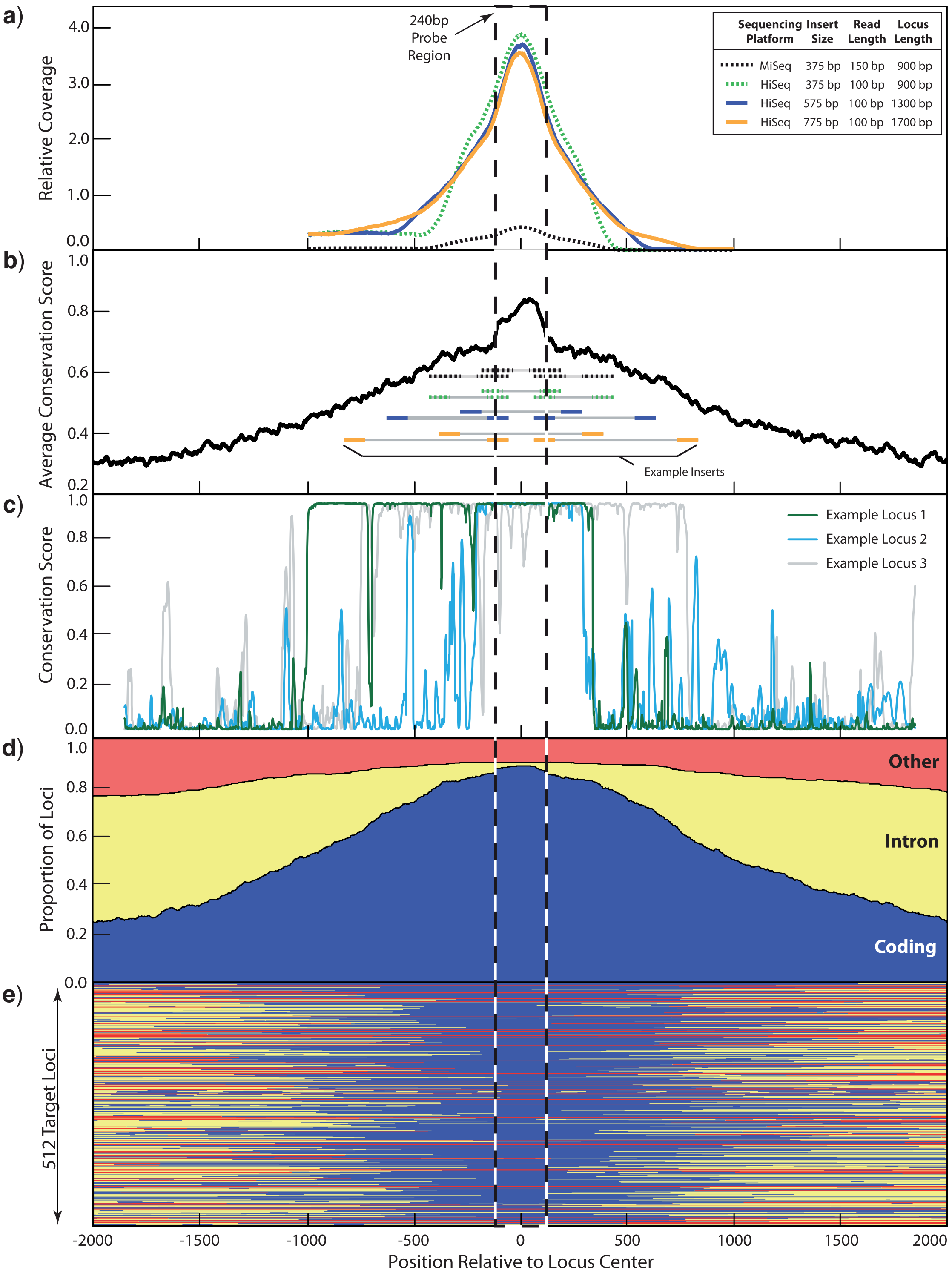

FIGURE 2. Shallow-scale phylogenetic utility. (a) Within locus coverage is affected by insert size chosen (results shown based on Bowtie mapping assembly of *Homo*). Relative coverage was computed by taking the average coverage for corresponding positions within a library (insert size), then scaling the coverage distributions such that the average coverage is 1 for each library (to remove effects of unequal pooling). The coverage distribution for the MiSeq was additionally scaled to reflect lower output relative to the HiSeq. Probe regions were chosen to maximize sequence conservation in the region and minimize conservation just outside the region. (b) Primate phastCons conservation scores (Siepel et al. 2005) were averaged across loci for each position within 2000 bp of the probe region center. Examples of fragments centered over the probe region and minimally overlapping with the probe region are shown below the conservation curve to indicate maximum locus size for each of the four read types. Note that longer fragments can be used to obtain the level of sequence variation needed for a study. (c) Examples of conservation scores for individual loci indicate sharp boundaries between conserved and less conserved regions for some loci. (d) About 90% of the probe regions contain coding sequence, but the proportion of loci with sites in coding regions decreases rapidly with increasing distance from the probe region. (e) UCSC Genes annotation is indicated for each locus as a horizontal line with color at each position indicating existence in coding, intron, or other genomic element (e.g., UTR), with shades corresponding to those in (d).

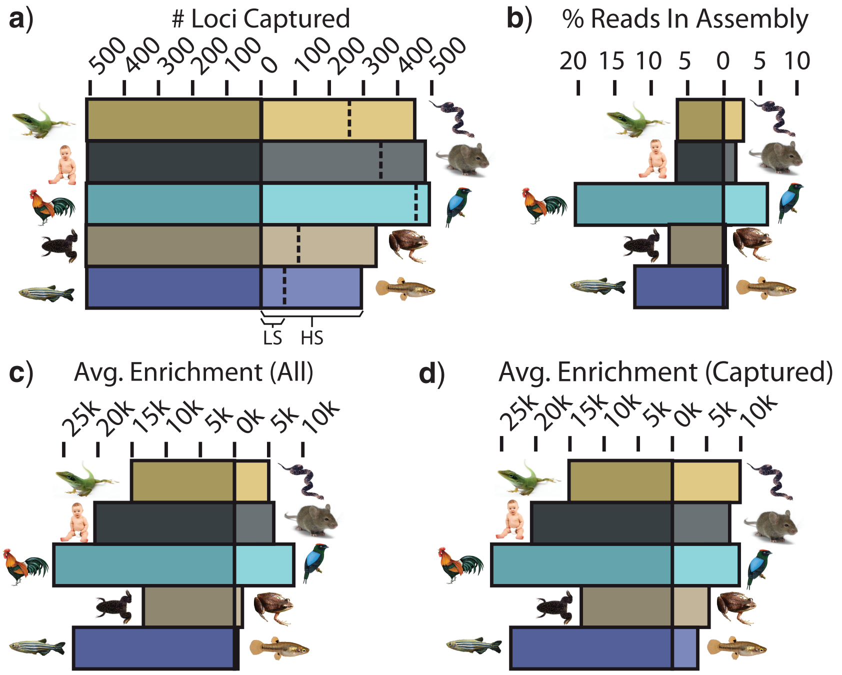

FIGURE 3. Variation in capture efficiency across species. Capture efficiency is represented as (a) the number of captured loci (high-sensitivity estimates shown with solid bars, low-sensitivity shown with dashed lines), (b) the percent of reads that were included in the final assemblies, (c) fold-enrichment (percent of reads included in final assemblies divided by the percent of reads that are expected to fall in target regions by chance, based on genome size), (d) fold-enrichment considering only captured loci. Results presented in (b–d) are based on the low-sensitivity (quasi-*de novo*) analyses and are therefore conservative. Data for the five model species are shown to the left of the vertical line and data for the five nonmodel species are shown to the right of this line in each figure. Note that even for taxa that had relatively low levels of enrichment (e.g., nonmodel frog *Pseudacris*), a substantial number of loci were still captured with enough coverage to be useful for phylogenetic studies.

comparisons of primate reference genome sequence (Siepel et al. 2005), the average degree of sequence conservation (Fig. 2b) is highest in the probe region (at 80% average conservation across primates) and declines steadily with increasing distance from the probe region (to about 30% average sequence conservation across primates at 2000 bp from the probe region center). Fragments from libraries of all insert sizes extend into more variable regions, though the larger insert-sized libraries reach the most variable regions. The longest insert size used in this study, 775 bp, extended to sites below 50% average sequence conservation. Inspection of conservation scores for individual loci reveal sharp boundaries between conserved and less conserved regions for many loci (three examples are given in Fig. 2c). The majority of loci contain a mixture of coding, intron, and/or other sequence (Fig. 2d). The annotation for each locus (by position) is given in Figure 2e. This indicates that many of the loci captured contain rgions with high degrees of sequence variation, especially if long inserts are used.

Enrichment Efficiency

Essentially all loci were successfully captured for the five model species (average 99.8%), but not all were successfully captured for the five nonmodel species (avg. 88.4%, Fig. 3a). The numbers of loci successfully captured (of 512 possible) for the model species were 511 (99.8% *Anolis*), 512 (100% *Gallus*), 509 (99.4% *Homo*), 512 (100% *Xenopus*), and 510 (99.6% *Danio*). The numbers of loci captured for the nonmodel species were 464 (90.6% *Crotalus*), 508 (99.2% *Chiroxiphia*), 481 (93.9% *Mus*), 337 (65.8% *Pseudacris*), and 294 (57.4% *Heterandria*). The percent of reads included in final, quasi-*de novo* assemblies also varied by species (Fig. 3b), ranging from 0.27% to 21.67%. These statistics were computed as the percent of all quality-filtered sequencing reads that remained in the assemblies after all of the quasi-*de novo* assembly and adjustment steps were performed. These percentages may be somewhat conservative, however, because they do not factor out reads that may not map anywhere in the genome (e.g., due to excessive sequencing error) and do not account for captured fragments for which neither read extended into the probe region (see Methods section). We computed enrichment as the percentage of filtered reads that assembled to a locus divided by the percentage that would have been obtained if fragments were uniformly distributed across the genome. This metric provides a measure of the factor by which the cost of sequencing the target loci is reduced by application of the anchored enrichment approach as opposed to sequencing a standard genomic library. Enrichment for model species ranged from 13,342 to 26,451 (averaged across loci within each species). Mixed

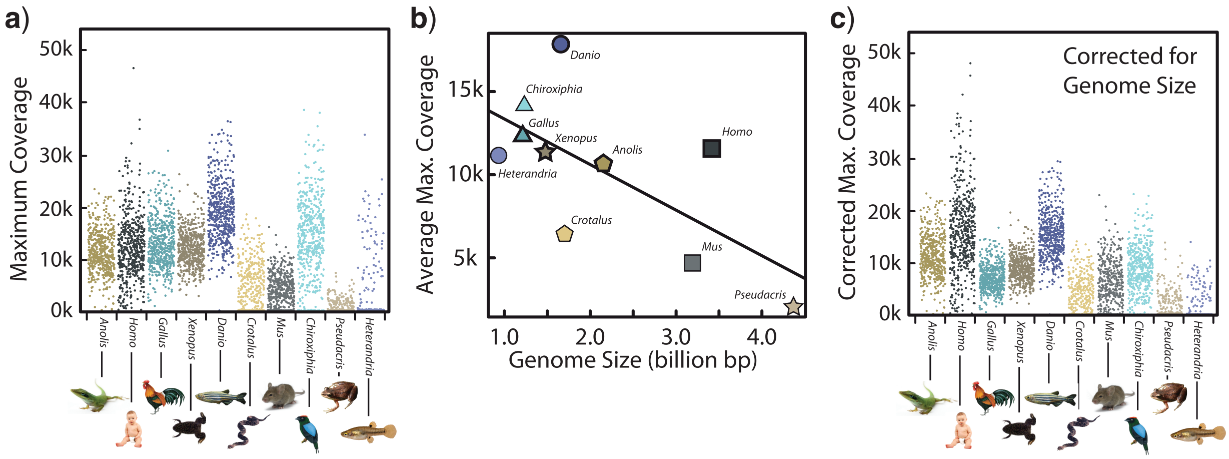

FIGURE 4. Variation in capture efficiency across loci. (a) Coverage was computed as the maximum number of reads overlapping a site within an assembled locus. (b) Since libraries from different species were pooled in equal concentrations regardless of genome size, coverage in target regions is depressed for species with larger genomes (given equal enrichment). (c) When coverage is corrected for genome size (r^2 = 0.4, $P = 0.02955$, $y = 16\,759 \pm 2974x$), coverage is more consistent across species for nonmodel species, suggesting that pooling in concentrations proportional to genome size may improve coverage equality across species. Results are based on the low-sensitivity analyses and are therefore conservative.

results were observed for the nonmodel taxa for which enrichment levels ranged from 500 to 7900 (Fig. 3c). Note that correcting the enrichment levels by removing loci not identified as captured in the low-sensitivity, quasi-*de novo* assembly yields a range in the nonmodel species from 3711 to 10,294 (Fig. 3d).

Capture efficiency varied substantially across loci and across species (Fig. 4). High coverage was observed for the five model species (used in the probe design) for nearly all loci. Coverage varied for the nonmodel species: High coverage was obtained for *Chiroxiphia* at most loci and for *Heterandria* at some loci; Reduced coverage was obtained for *Crotalus* and *Mus* at most loci; and modest coverage was obtained for *Pseudacris* at a moderate number of loci (Fig. 4a). Some of the among-species variation in coverage could be attributed to variation in genome size ($r^2 = 0.4$, $P = 0.02955$, $y = 16\,759 \pm 2974x$, Fig. 4b). Correcting coverage for genome size (to estimate the coverage that would have been obtained if we had pooled in concentrations proportional to the genome size rather than in equal concentrations) has mixed effects (Fig. 4c). Among-species variation in coverage increased for the model species but decreased for the nonmodel species when the correction was applied.

Captured loci were distributed broadly across the genome and corresponded with target regions, as confirmed by results from the genome-wide mapping of the reads (Fig. 5; Supplementary Figs. 2–4). Note that probe regions were chosen without consideration of their genomic location and thus were not expected to be broadly distributed a priori. We considered a site to be within a target region if it exists within 1000 bp of the center of a probe region (a range just greater than the maximum expected for the 775-bp insert size libraries). In the four species for which Bowtie mapping was performed (*Homo*, *Gallus*, *Danio*, and *Mus*), 60–82% of reads mapped uniquely to the genome, and 6–26% of the uniquely mapped reads mapped to the target regions. Coverage within target regions was peaked in shape at the probe region (Fig. 5, inset) as expected. A small number of nontarget regions were substantially enriched. Although distributed broadly, target regions were clustered in some areas of the genome. In some cases, in fact, probe regions were within 10,000 bp of the neighboring probe region.

Read Properties and Assembly Results

Details of the read quantity and quality, as well as the details of the assemblies are given in the Supplementary Materials. In brief, the Illumina HiSeq and MiSeq sequencers produced the expected number of reads (more than 300 and 13 million, respectively) and had comparable quality scores (averaging >24, Supplementary Figs. 5 and 6). The 8-bp indexes allowed the vast majority of reads to be identified (>98%), the occurrence of PCR duplicates was modest ($\leqslant$20%, Supplementary Fig. 7), and the number of overlapping reads requiring merging was moderate (~20%). Manual inspection of the assemblies indicated that of the captured loci (74% by low-sensitivity estimation), a large portion of the corresponding assemblies (91%) belonged to the "excellent" or "very good" categories (Table 1). The remaining loci showed evidence of co-assembled gene duplicates or other assembly errors. Metadata for each of the final alignments are given in Supplementary Tables 3 and 4.

Throughput Using Anchored Enrichment

The length of loci recovered using the anchored enrichment approach depends on the shape of the fragment length distribution generated during

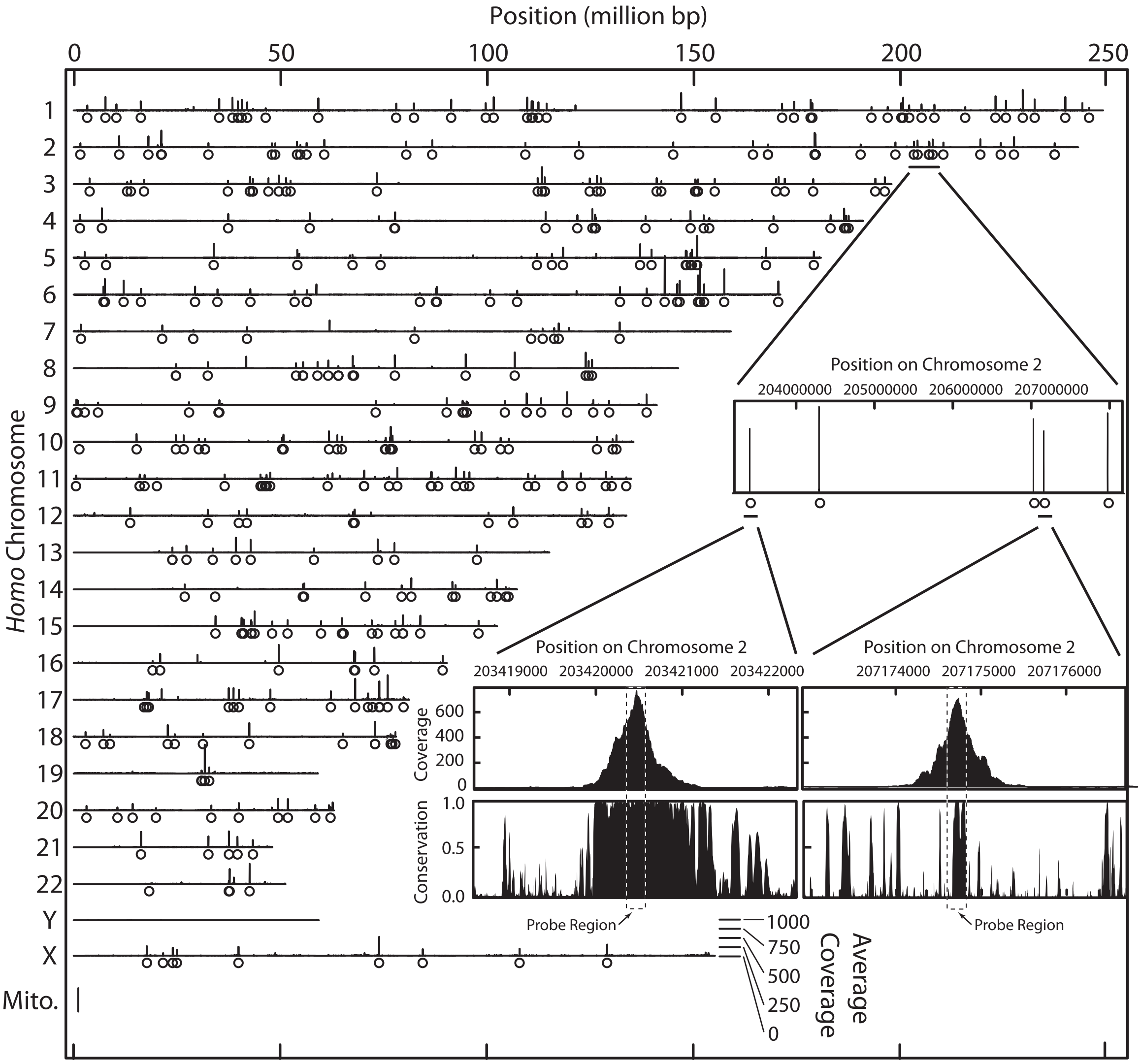

FIGURE 5. Genome-wide capture efficiency. The coverage profile for each chromosome (indicated by height of vertical lines above each chromosome), computed as the average coverage for 10,000 bp segments, was obtained from the results of a Bowtie mapping assembly for *Homo*. Probe regions, indicated by circles under corresponding genome positions, are well-dispersed across the human genome. Finer-scale coverage plots reveal that coverage for enriched target regions is peaked at the probe region (here coverage is defined as the number of reads overlapping with a given base). Conservation scores shown are derived from phastCons data set for primates and represent the posterior probability that the given site is conserved (Siepel et al. 2005). Analogous plots for *Gallus*, *Danio*, and *Mus* are given in Supplementary Figures 2–4.

library preparation. Four expected fragment length distributions and the corresponding expected coverage distributions are given in Figure 6a and b, respectively. In particular, we compare 1) sonication followed by gel-based size selection of 775-bp insert sizes, 2) Covaris sonication with 450-bp protocol (Covaris, Inc., http://www.covarisinc.com), 3) Covaris sonication with 1000-bp protocol, and 4) Nextera sample preparation kits (Illumina, Inc. http://www.illumina.com/products/nextera_dna_sample_prep_kit.ilmn). In general, size distributions with a higher mean are expected to produce longer loci and vice versa. Moreover, size distributions with higher variance are expected to produce coverage distributions that decrease more gradually from the center of the locus. All coverage distributions are expected to be peaked in the probe region because reads with different orientations may overlap there.

The number of enriched samples that can be sequenced in one sequencing lane depends on the expected coverage distribution and the species-specific enrichment efficiency (which is a function of genome size and evolutionary distance from model). Using the expected coverage distributions shown in Figure 6b and the maximum coverage values presented in Figure 4a, we estimated the tradeoff between the number of loci captured and the median length of loci for different combinations of fragment size distribution and number of individuals pooled in one sequencing lane. As seen in Figure 7 (and Supplementary Figs. 8–10), when genome

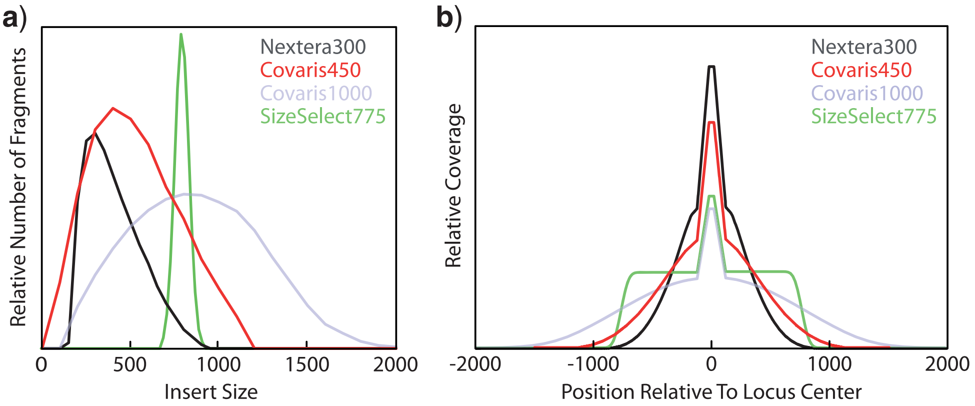

FIGURE 6. The length of captured loci can be adjusted by changing the distribution of insert sizes produced during library preparation. (a) Expected insert size distributions are shown for four different methods (Nextera producing a mean of 300 bp; Covaris sonication to 450 bp; Covaris sonication to 1000 bp; sonication followed by size selection of 775-bp fragments). (b) Each fragment distribution is expected to produce a different distribution of sequencing coverage (results for 100-bp reads shown). Coverage peaks sharply at the locus center because reads extending in different directions may overlap only in the probe region. Locus length obtained depends on overall coverage.

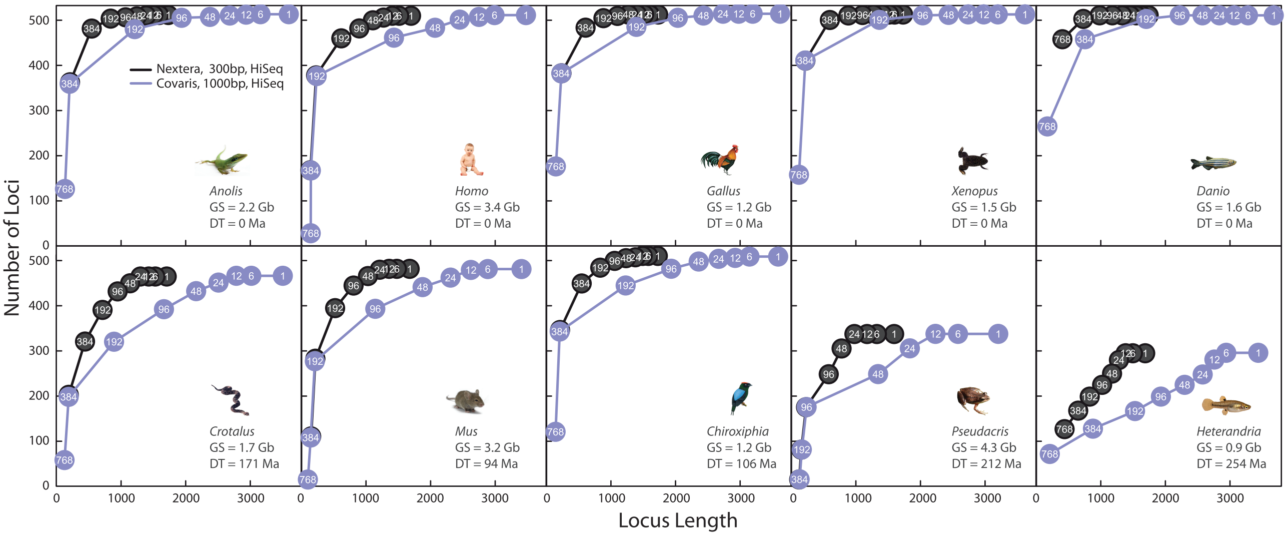

FIGURE 7. Estimated tradeoff between the number of captured loci, the median length of captured loci, and the number of individuals pooled in one sequencing lane (indicated as the number contained in each point). The nature of the tradeoff is a function of the fragment length distribution produced during library preparation (Nextera, median = 300 bp; Covaris, median=1000 bp; other examples shown in Supplementary Figs. 8–10), as well as the genome size (GS) and evolutionary distance to the nearest model species (DT). Results assume paired-end 100 bp sequencing is performed on the Illumina HiSeq 2000 sequencer with version 3 chemistry, a minimum of 10-fold coverage for inclusion of site in locus, and percent on-target efficiencies reflecting those observed in this study.

size and/or evolutionary distance to a model is small, large numbers of long loci can be obtained even when large numbers of individuals are pooled in a single Illumina HiSeq sequencing lane (but note that current Illumina sequencing protocols may preferentially read clusters generated by shorter fragments). For example, more than 192 manakin (*Chiroxiphia*) samples can be pooled if approximately 450 loci with median length ~1000 bp are desired. A smaller number of individuals can be pooled for species with large genome sizes and high degrees of divergence. For example, pooling 48 chorus frog (*Pseudacris*) individuals is expected to result in approximately 250 loci with median length ~1200 bp. In general, a larger number of longer loci can be obtained by using longer size fragments and pooling a smaller number of individuals. Conversely, large numbers of individuals can be pooled if smaller numbers of short loci are sufficient for a given project.

DISCUSSION

The anchored enrichment approach introduced here is a highly cost-effective, rapid, and massively multilocus method for obtaining the sequence data needed to produce high-resolution species trees. This anchored phylogenomic approach has the potential to accelerate the completion of the Tree of Life because it effectively eliminates marker development, one of the main bottlenecks in phylogenetic research. Once a capture probe set is fully developed for a broad taxonomic group (e.g., vertebrates), the same tool can be used to quickly obtain many orthologous loci from any collection of organisms within that group, regardless of the taxonomic depth. The anchored phylogenomic approach may also serve to unify systematic research because researchers working on different groups and on different taxonomic scales can easily use the same set of loci. Use of a common set of loci will aid in assembly of the Tree of Life because meta-analyses can be more easily conducted.

One novel feature of this approach is that it can be easily customized to fit the phylogenetic scale of interest. Sites with a greater degree of sequence variation can be accessed by simply increasing the insert size during the library preparation, allowing longer loci with more sequence variation to be used for shallow-scale applications. The extent to which locus length relates to shallow-scale phylogenetic utility will depend on genomic properties. In species with relatively reduced intron content, use of long fragments may simply extend some of the loci to include adjacent exons, which may have limited shallow-scale utility. One additional feature of the anchored enrichment approach is the number of individuals can be increased to some degree without sacrificing the number of loci, although the usable locus length may decrease.

Phylogenetic Utility

The vertebrate anchored enrichment tool we developed here shows great promise in terms of phylogenetic utility. Although designed from just five model species, we have successfully applied the probe set to five nonmodel species quite divergent from the models (divergence from nearest model ranging from 94 to 254 Ma). We captured over 500 orthologous loci from the model species and between 294 and 508 of the 512 target loci from each of the nonmodel species. Results from published simulation studies suggest that we captured a sufficient number of loci to resolve difficult species tree branches (Huang et al. 2010; Leaché and Rannala 2011). We used 32 loci, for example, to completely resolve a tetrapod phylogeny, despite low taxon sampling and substantial gene tree discordance for several branches.

Estimates of sequence conservation in primates suggest that the capture tool developed here could be used to resolve shallow-scale phylogenies (Fig 2b). We chose probe regions that were highly conserved but adjacent to less conserved regions. Given that average sequence conservation declines gradually with the distance from the probe region, we expect that sequence data informative at shallow taxonomic depth can be obtained through the use of larger insert sizes. Although our study was not designed to confirm the utility of this approach for phylogeography, we expect that a subset of our markers would be informative at that time scale since many intron and nontranscribed sequences were recovered. It is noteworthy that average sequence conservation continues to decline well beyond the point reached by our longest insert size (775-bp insert producing 1700-bp loci max). This suggests use of even larger insert sizes may increase the utility of the probe set to even shallower time scales, though there is undoubtedly a limit to the fragment size that can be efficiently sequenced. Use of large insert sizes to sequence long loci for deep-scale phylogenetics may not be particularly useful because much of the locus may not be alignable due to high sequence variation or lack of homology. Use of small insert sizes in this case will allow a greater number (potentially thousands) of individuals to be pooled in a single sequencing lane (see below).

Comparison to PCR-Based Approaches

The method developed here can be used in nonmodel organisms for a fraction of the cost and effort required by other approaches used to collect data in phylogenetics (Tables 2 and 3). We estimated the cost (reagents and sequencing only) of a project involving 100 taxa and up to 500 loci, for example, to be approximately $7085 (approximately $71 per individual). This cost is <1% of the cost of traditional PCR-based Sanger sequencing and 5% of the cost of PCR amplicon sequencing on the Illumina platform (Table 2). The cost can be reduced further using the Illumina MiSeq sequencer (approximately $2600 reduction from HiSeq cost for a paired-end 150-bp run), if shorter loci and/or smaller numbers of individuals are required. For smaller scale projects (e.g., 1 locus for 60–800 individuals or up to 10 loci for 20 individuals), Sanger sequencing of PCR products is still the more economical option. According to our calculations, there were no cases where high-throughput sequencing of PCR amplicons was the lowest cost option (Table 2).

Estimates of labor costs overwhelmingly favor anchored enrichment for any projects larger than 20 taxa × 10 loci (Table 3). We estimated the cost (labor) of a project including 100 taxa and 500 loci at 750 working days (approximately $142,500) for both PCR-based approaches but only 14 days (approximately $2705) for the anchored enrichment approach. This amount is <2% of the cost of both PCR-based Sanger sequencing and PCR amplicon sequencing on the Illumina platform. Collectively, the total cost (reagents, sequencing, and labor) for anchored enrichment of a 100 taxa × 500 loci project is 1–3.5% of these other methods. Furthermore, these estimates ignore the time and resources required for marker development in PCR-based approaches, which can substantially increase the cost for some systems.

One potential drawback to the anchored enrichment approach is the somewhat large initial investment in the system. Purchasing the indexed library adapters necessary for indexing 96 samples, for example, costs

TABLE 2. Comparison of costs of sequencing approaches for varying numbers of loci and individuals. The first table illustrates the cost of performing Sanger sequencing on PCR products. The second table shows the cost of PCR amplicon sequencing via paired-end sequencing on an Illumina HiSeq. The third table indicates the cost of performing anchored enrichment and Illumina HiSeq sequencing on libraries derived from genomic DNA. The fourth table presents the base costs from which the numbers in the top three tables are taken (an "—" indicates a cost that does not pertain to the sequencing method). Regions of parameter space where one method has a cost advantage over the others are indicated in bold (un-bolded numbers indicate a less cost-efficient method). Note that for any project larger than 20 individuals and 10 loci, anchored enrichment has a substantial advantage over other approaches. All cost estimates are in US dollars. Note that the costs below are current estimates only and can vary substantially based on reagent sources, NGS sequencing costs, and technological changes. These prices are based on actual costs incurred in this and other recent studies performed in our laboratory. Note that costs do not include labor, which is expected to be substantially higher in PCR-based approaches (Table 3)

PCR + Sanger sequencing

			Number of individuals		
Number of Loci	20	60	100	200	800
1	**292**	**875**	**1458**	**2916**	**11664**
10	**2916**	8748	14580	29160	116640
50	14580	43740	72900	145800	583200
100	29160	87480	145800	291600	1166400
500	145800	437400	729000	1458000	5832000

PCR + Illumina HiSeq sequencing

			Number of individuals		
Number of loci	20	60	100	200	800
1	4049	4946	5843	8086	21544
10	4513	6339	8165	12730	40120
50	6577	12531	18485	33370	122680
100	9157	20271	31385	59170	225880
500	29797	82191	134585	265570	1051480

Anchored enrichment + Illumina HiSeq sequencing

			Number of individuals		
Number of loci	20	60	100	200	800
1	4297	5691	7085	10570	31480
10	4297	**5691**	**7085**	**10570**	**31480**
50	**4297**	**5691**	**7085**	**10570**	**31480**
100	**4297**	**5691**	**7085**	**10570**	**31480**
500	**4297**	**5691**	**7085**	**10570**	**31480**

	PCR + Sanger	PCR + Illumina	Anchored enrichment + Illumina
PCR cost per locus per indiv.[a]	2.58	2.58	—
Sanger sequencing per locus per indiv.[b]	12	—	—
Library preparation per indiv.[c]	—	19.85	19.85
Anchored enrichment per indiv.[d]	—	—	15
Illumina sequencing cost per project[e]	—	3600	3600

[a] Includes cost of PCR reagents, gels, gel extraction via kit, tubes, but not pipette tips; also assumes no PCR failures.

[b] Assumes each locus sequenced via a forward and reverse reaction at $6.00/reaction (current price at Florida State University Core DNA Sequencing Facility); also does not include the cost of cloning different alleles if phasing is required.

[c] Assumes no library failures; these costs are derived from Meyer and Kircher (2010) protocol.

[d] Costs calculated from the Agilent SureSelect kit (120kb capture kit with 100 reaction scale) and assume 10 individuals are multiplexed per tube as done in present study.

[e] Assumes cost of one lane on an Illumina HiSeq per project given current costs (August 2011) at the Hudson-Alpha Institute for Biotechnology in Huntsville, AL.

approximately $5000 due to the need for HPLC purification of indexing oligos (following Meyer and Kircher 2010, the library preparation protocol used here). Note, however, that hundreds of projects involving 96 individuals can be performed with this reagent purchase. The cost of the hybrid enrichment reagents is also an investment. For this study, we used Agilent custom SureSelect probes together targeting 120,000 bp, a kit which includes 100 enrichments at a cost of approximately $15,000 (the minimum-priced kit for our target size in Oct 2010). Other companies produce probe kits as well at potentially lower costs, such as Illumina, Inc., Roche NimbleGen Inc., and MYcroarray, Inc. Multiple individuals, however, can be combined before enrichment (here we pooled 10 individuals per enrichment and used only 3% of the enrichment reagents), and this strategy can dramatically reduce the cost. Though these costs are figured into the estimates shown in Table 2, researchers desiring to conduct a small number of anchored enrichment studies would benefit from collaborating with other laboratory groups already invested in the system.

In addition to improved efficiency, the anchored enrichment approach also provides a more flexible approach to phylogenetic data collection for two reasons. First, locus length is not fixed as it is in PCR-based

TABLE 3. Comparison of labor required for different sequencing approaches, measured in number of days and salary costs, for varying numbers of loci and individuals. The first table illustrates the cost of performing Sanger sequencing on PCR products. These costs are expected to be greater for PCR amplicon sequencing via paired-end sequencing on an Illumina HiSeq due to the extra time required for library preparation (not shown). The second table indicates the cost of performing anchored enrichment and Illumina HiSeq sequencing on libraries derived from genomic DNA. The number of days of labor required for anchored enrichment scales with the number of individuals according to the equation: $y = 1.45 * x^{0.496}$ and is equivalent regardless of the number of loci targeted. Regions of parameter space where one method has a cost advantage over the others are indicated in bold (un-bolded numbers indicate a less cost-efficient method). Note that for any project larger than 20 individuals and 10 loci, anchored enrichment greatly exceeds the other two methods in cost efficiency. All cost estimates are in US dollars and assume a laboratory technician's salary of $190 per day (corresponding to a total annual cost of $49,400 including salary, fringe, health insurance, and life insurance benefits). Note that time estimates are based on our actual experience utilizing all three of these approaches in our laboratory.

PCR + Sanger sequencing (Number of days of laboratory work required)[a]

			Number of individuals		
Number of loci	20	60	100	200	800
1	**3**	**3**	**3**	**3**	**12**
10	**3**	15	15	30	120
50	15	75	75	150	600
100	30	150	150	300	1200
500	150	750	750	1500	6000

PCR + Sanger sequencing (Cost of project)

			Number of individuals		
Number of loci	20	60	100	200	800
1	**570**	**570**	**570**	**570**	**2280**
10	**570**	2850	2850	5700	22800
50	2850	14250	14250	28500	114000
100	5700	28500	28500	57000	228000
500	28500	142500	142500	285000	1140000

Anchored enrichment + Illumina HiSeq sequencing (Number of days of laboratory work required)

			Number of individuals		
Number of loci	20	60	100	200	800
	6	11	14	20	40
10	6	**11**	**14**	**20**	**40**
50	**6**	**11**	**14**	**20**	**40**
100	**6**	**11**	**14**	**20**	**40**
500	**6**	**11**	**14**	**20**	**40**

Anchored Enrichment + Illumina HiSeq sequencing (Cost of project)

			Number of individuals		
Number of loci	20	60	100	200	800
1	1217	2099	2705	3814	7587
10	1217	**2099**	**2705**	**3814**	**7587**
50	**1217**	**2099**	**2705**	**3814**	**7587**
100	**1217**	**2099**	**2705**	**3814**	**7587**
500	**1217**	**2099**	**2705**	**3814**	**7587**

[a] Assumes projects with 20–200 total reactions (loci x individuals) require three days of labor. Larger projects with ⩾60 individuals and ⩾10 loci are calculated on 96-well plate scale, assuming ~two plates can be prepared for sequencing every three days of labor (including PCR amplification, gel purification, and sequencing reactions).

methods. If longer loci are required, the insert size can simply be increased during the size selection step of the library preparation. PCR-based approaches require a reference sequence from which additional primers can be developed to lengthen loci or require developing larger gene regions de novo. The second advantageous feature is the fact that the anchored enrichment approach presented here uses probes long enough to allow for sequence variation. It is this feature that allows the probe set to be applied on broad taxonomic time scales. Traditional PCR primers, typically <30 bp, are not particularly robust to sequence variation, a feature that greatly reduces their applicability to different taxonomic groups. Moreover PCR primers must be paired, reducing the set of target regions to those flanked by conserved sequence. Capture probes, in contrast, can be designed from islands of conservation immediately adjacent to less conserved regions of any size.

Improving Enrichment Efficiency in Vertebrates

Aside from improving the enrichment protocol, the optimal approach to increasing enrichment efficiency in vertebrates is to include probes designed from additional

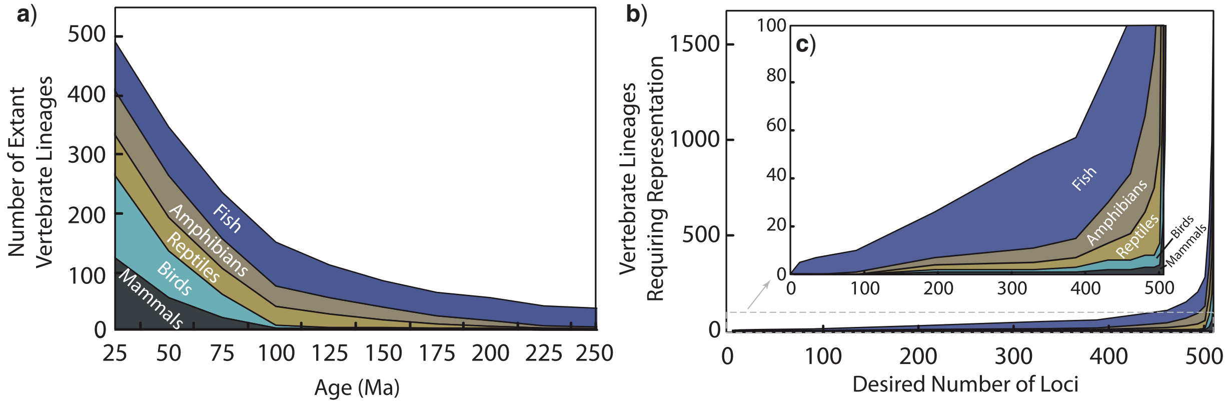

FIGURE 8. Prospects for increasing the number of captured loci across vertebrates. The number of evolutionary lineages varies across time and taxonomic group ([a], based on phylogeny of extant vertebrate families, Hedges et al. 2006). The number of loci captured for a species is a function of the evolutionary distance to the nearest model species included in the probe set design (Fig. 1b). To improve the utility of the probe set across vertebrates, the set should be supplemented with probes designed from additional species in various vertebrate groups. The total number of taxa from each respective clade required to achieve a desired number of orthologous loci is shown in (b) and (c). The number of required lineages was obtained by determining the number of lineages on the timetree.org phylogeny that existed at the time (depth) corresponding to the number of orthologs desired (based on the high-sensitivity relationship shown in Fig. 1b).

taxa. The initial probe design (used in this study) was based on only five representative model species across vertebrates. Additional probe sequences could be obtained from existing genomes or EST data as well as through low-coverage sequencing of new genomes. Since we found that the number of captured loci for a species is strongly correlated with the evolutionary distance to the nearest model species (Fig. 1b), we can estimate the number of species that should be represented in the probe set to capture a desired number of loci for all vertebrates. This estimate is given in Figure 8 for various numbers of loci. We used the TimeTree chronogram (www.timetree.org, October 20, 2012, Hedges et al. 2006) to estimate the number of vertebrate lineages that existed at different time points (Ma; Fig. 8a). This analysis suggests that the vertebrate groups most in need of additional genomic representation are (in order of priority) fish, amphibians, and reptiles. Capture efficiencies for mammals and birds would benefit much less from the inclusion of additional representatives.

Future Directions

The target enrichment approach demonstrated here could be applied to other major clades in the Tree of Life. Highly conserved regions of the genome have been discovered in insects, worms, and yeast as well as a diverse set of vertebrates (Glazov et al. 2005; Siepel et al. 2005; Stephen et al. 2008; Wang et al. 2009; Janes et al. 2010) and could easily be used in probe design as in the present study. Some challenges may arise in groups with large, highly repetitive genomes since repetitive elements are known to reduce capture efficiency (Bashiardes et al. 2005; Fu et al. 2010; Mamanova et al. 2010). The largest genome included in this study was the *Pseudacris* genome with haploid size equal to 4.27 billion base pairs (converted from pg C-value estimated by Goin et al. 1968). Since we pooled samples with equal concentration before sequencing, it is not surprising that we obtained the lowest coverage for the chorus frog sample; greater enrichment for that sample would be required to produce coverage on par with samples from species with smaller genomes. One additional hurdle to applying the anchored enrichment approach to nonvertebrates is the relatively small number of available genome sequences, although projects like the Genome 10K Project (vertebrates; Genome 10K Community of Scientists 2009; http://www.genome10k.org), the i5k Initiative (invertebrates; Robinson et al 2011; www.arthropodgenomes.org/wiki/i5K), and the 1000 Plant Genomes Project (www.onekp.com/) are likely to provide data very useful to the development of additional probe sets. Fully assembled genomes are not necessarily required for this effort; because probe regions were chosen based on conservation and uniqueness, low-coverage genome sequencing could potentially provide enough reads overlapping with probe regions to facilitate inclusion of underrepresented groups in the probe design. The application of anchored enrichment to phylogenetics of nonvertebrates, which is already underway, is likely to be fruitful area of future research.

The observed decline in efficiency with increasing taxonomic depth, though not surprising, suggests that this approach may be somewhat limited at very deep taxonomic depths (e.g., at the base of Tree of Life). The biological reality that a smaller number of orthologous loci exist at deeper taxonomic depths cannot be overcome. Nonetheless, probe sets may be designed to include a mixture of loci orthologous to members of the target clade (e.g., the 512 vertebrate-specific loci targeted in this study) and a smaller number of loci useful at deeper timescales (e.g., across the Tree of Life). This tiered approach would facilitate the integration of results from researchers working on different groups. Moreover, probes targeting loci commonly sequenced with PCR-based approaches could also be included to allow previously collected data to be incorporated into new studies. Regardless of the specific strategy, efforts to reconstruct the Tree of Life will greatly benefit from the adoption of a common set of loci by researchers. The integrated data collection enabled by the anchored enrichment approach will facilitate construction of a more robust phylogeny.

Our results suggest that target locus selection may best be performed at the taxonomic scale comparable to vertebrates, though the most appropriate scale is likely to change somewhat based on the genomic properties of the specific taxonomic group. Increasing the taxonomic depth beyond ~500 Ma would decrease the number of loci that could be targeted and thus decrease the utility of the locus set for shallow-scale studies requiring large numbers of loci. Decreasing the taxonomic depth, in contrast, would result in the need for an increasing number of researchers to develop independent target locus sets and may result in largely nonoverlapping target locus sets across groups.

The taxonomic scale at which probes can be designed, in contrast, is very flexible. A researcher wishing to study a particular family of Serpentes, for example, can use existing snake transcriptome or whole-genome resources to design a probe set representing the genetic diversity of different snake lineages within the family but that still targets the same vertebrate locus set developed in this study. In this way, the researcher can improve enrichment efficiency without decreasing the long-term value that the collected data would bring to vertebrate metastudies. The only downside to this approach is that the reagents purchased by the researcher would have somewhat limited use for species in vertebrate groups divergent from snakes. The alternative approach is to design probes that represent broad-scale species diversity, as we have in this study. As discussed previously, increasing the number of lineages represented in a probe set is one way to increase efficiency of the tool without decreasing the broader utility of the data collected with this method. The anchored enrichment approach developed in this study has the potential to transform phylogenetics, especially as the number of whole genome sequences and other genomic resources increase.

Summary

We introduce a novel, hybrid enrichment-based approach for phylogenetic data collection that will enable researchers studying nonmodel organisms to rapidly access hundreds of loci, without time-consuming primer development, at a fraction of the cost of standard approaches. By leveraging existing genomic resources, recently developed genome technology, and high-throughput sequencing, the anchored enrichment approach has the potential to revolutionize the field of phylogenetics and accelerate final assembly of the Tree of Life. The approach demonstrated here in the vertebrate clade could easily be extended to any nonvertebrate system for which some genomic resources are available.

Resources

Resources developed during this study (e.g., probe sequences, bioinformatic scripts) can be obtained from the Dryad repository (http://datadryad.org, doi:10.5061/dryad.r606d128) and at http://www.anchoredphylogeny.com. The latter website will also contain updates to the probe set, laboratory protocol, and bioinformatic pipeline as they become available. Researchers interested in contributing to future development of the vertebrate (as well as nonvertebrate) anchored enrichment tools should visit http://www.anchoredphylogeny.com

Supplementary Material

Supplementary material, including supplemental methods, results, figures, and tables, can be found can be found in the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.r606d128.

Funding

This work was funded by Florida State University new faculty setup funds to Alan Lemmon and Emily Moriarty Lemmon.

Acknowledgements

We are grateful to Brian Caudle for contributing to the early stages of protocol development and to Lisa Barrow, Jay Hogan, Hannah Ralicki, John Malone, and Alisha Rossi for assisting with data screening and analysis. We thank John McCormack for a useful discussion about the UCE method at the 2011 Evolution Meetings. We thank John Malone, Ron Debry, Bryan Carstens, Dave Weisrock, and one anonymous reviewer for valuable comments that improved the quality of this paper. Finally, we are very grateful to Scott Taylor at Illumina for arranging free HiSeq and MiSeq sequencing runs for this study.

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., Weinstock G.M., Gibbs R.A. 2007. Direct selection of human genomic loci by microarray hybridization. Nat. Methods. 4:903–905.

Ané C., Larget B., Baum D.A., Smith S.D., Rokas A. 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24:412–426.

Archer M.J., Long N., Lin B. 2010. Effect of probe characteristics on the subtractive hybridization efficiency of human genomic DNA. BMC Res. Notes 3:109.

Bader D., Roshan U., Stamatakis A. 2006. Advances in computers. Computational grand challenges in assembling the tree of life: problems and solutions. Amsterdam, The Netherlands: Elsevier.

Bashiardes S., Veile R., Helms C., Mardis E.R., Bowcock A.M., Lovett M. 2005. Direct genomic selection. Nat. Methods. 2:63–69.

Bejerano G., Pheasant M., Makunin I., Stephen S., Kent W.J., Mattick J.S., Haussler D. 2004. Ultraconserved elements in the human genome. Science 304:1321–3125.

Cracraft J., Donoghue M.J. 2004. Assembling the tree of life. New York: Oxford University Press.

Donoghue M.J. 2004. Immeasurable progress on the tree of life. In: Cracraft J. and Donoghue M.J., editors. Assembling the tree of life. New York: Oxford University Press. p. 548–552.

Drummond A.J., Ashton B., Buxton S., Cheung M., Cooper A., Heled J., Kearse M., Moir R., Stones-Havas S., Sturrock S., Thierer T., Wilson A. 2010. Geneious v5.5.1. Available from http://www.geneious.com.

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

Edwards S.V., Liu L., Pearl D. K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA. 104: 5936–5941.

Faircloth B.C., McCormack J.E., Crawford N.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers for target enrichment spanning multiple evolutionary timescales. Syst. Biol. 61:717–726.

Fu Y., Springer N.M., Gerhardt D.J., Ying K., Yeh C-T., Wu W., Swanson-Wagner R., D'Ascenzo M., Millard T., Freeberg L., Aoyama N., Kitzman J., Burgess D., Richmond T., Albert T.J., Barbazuk W.B., Jeddeloh J.A., Schnable P.S. 2010. Repeat subtraction-mediated sequence capture from a complex genome. Plant J. 62:898–909.

Fujita P.A., Rhead B., Zweig A.S., Hinrichs A.S., Karolchik D., Cline M.S., Goldman M., Barber1 G.P., Clawson H., Coelho A., Diekhans M., Dreszer T.R., Giardine B.M., Harte R.A., Hillman-Jackson J., Hsu F., Kirkup V., Kuhn R.M., Learned K., Li C.H., Meyer L.R., Pohl A., Raney B.J., Rosenbloom K.R., Smith K.E., Haussler D., Kent W.J. 2011. The UCSC Genome Browser database: update 2011. Nucleic Acids Res. 39:D876–D882.

Genome 10K Community of Scientists. 2009. A proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered. 100:659–674.

Glazov E.A., Pheasant M., McGraw E.A., Bejerano G., Mattick J. S. 2005. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the constrol of homothorax mRNA splicing. Genome Res. 15:800–808.

Gnirke A., Melnikov A., Maguire J., Rogov P., LeProust E.M., Brockman W., Fennell T., Giannoukos G., Fisher S., Russ C., Gabriel S., Jaffe D.B., Lander E.S., Nusbaum C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27:182–189.

Goin O.B., Goin C.J., Bachmann K. 1968. DNA and amphibian life history. Copeia. 1968:532–540.

Gregory T.R., Nicol J.A., Tamm H., Kullman B., Kullman K., Leitch I.J., Murray B.G., Kapraun D.F., Greilhuber J., Bennett M.D. 2006. Eukaryotic genome size databases. Nucleic Acids Res. doi:10.1093/nar/gkl828.

Hedges S.B., Dudley J. Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22:2971–2972.

Huang H., He Q., Kubatko L.S., Knowles L.L. 2010. Sources of error inherent in species-tree estimation: impact of muational and coalescent effects on accuracy and implications for choosing among different methods. Syst. Biol. 59:573–583.

Janes D.E., Chapus C., Gondo Y., Clayton D.F., Sinha S., Blatti C.A., Organ C.L., Fujita M.K., Balakrishnan C.N., Edwards S.V. 2010. Reptiles and mammals have differentially retained long conserved noncoding sequences from the amniote ancestor. Genome Biol. Evol. 3:102–113.

Keeling P.J., Burger G., Durnford D.G., Lang B.F., Lee R.W., Pearlman R.E., Roger A.J., Gray M.W. 2005. The tree of eukaryotes. Trends Ecol. Evol. 20:670–676.

Kircher M., Sawyer S., Meyer M. 2011. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 2011:1–8.

Lane C.E., Archibald J.M. 2008. The eukaryotic tree of life: endosymbiosis takes its TOL. Trends Ecol. Evol. 23:268–275.

Langmead B., Trapnell C., Pop M., Salzberg S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.

Leaché A.D., Rannala B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Syst. Biol. 60:126–137.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics. 24:2542–2543.

Liu L., Yu L. 2011. Estimating species trees from unrooted gene trees. Syst. Biol. 60:661–667.

Lutzoni F., Kauff F., Cox C.J., McLaughlin D., Celio G., Dentinger B., Padamsee M., Hibbett D., James T.Y., Baloch E., Grube M., Reeb V., Hofstetter V., Schoch C., Arnold A.E., Miadlikowska J., Spatafora J., Johnson D., Hambleton S., Crockett M., Shoemaker R., Hambleton S., Crockett M., Shoemaker R., Sung G.H., Lucking R., Lumbsch T., O'Donnell K., Binder M., Diederich P., Ertz D., Gueidan C., Hansen K., Harris R.C., Hosaka K., Lim Y.W., Matheny B., Nishida H., Pfister D., Rogers J., Rossman A., Schmitt I., Sipman H., Stone J., Sugiyama J., Yahr R., Vilgalys R. 2004. Assembling the fungal tree of life: progress, classification and evolution of subcellular traits. Amer. J. Bot. 91:1446–1480.

Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

Maddison D.R., Maddison W.P. 2005. MacClade 4: Analysis of phylogeny and character evolution. Version 4.08a. http://macclade.org.

Mamanova L., Coffey A. J., Scott C. E., Kozarewa I., Turner E. H., Kumar A., Howard E., Shendure J., Turner D.J. 2010. Target-enrichment strategies for next-generation sequencing. Nat. Methods. 7:111–118.

McCormack J.E., Faircloth B.C., Crawford N.G., Gowaty P.A., Brumfield B.T., Glenn T.C. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis. Genome Res. 22:746–754.

Meyer M. and Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. 2010. doi:10.1101/pdb.prot5448.

Nylander J.A.A. 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Pace N.R. 2009. Mapping the tree of life: progress and prospects. Microbiol. Mol. Biol. Rev. 73:565–576.

Palmer J.D., Soltis D.E., Chase M.W. 2004. The plant tree of life: an overview and some points of view. Amer. J. Bot. 91: 1437–1445.

Parfrey L.W., Grant J., Tekle Y.I., Lasek-Nesselquist E., Morrison H.G., Sogin M.L., Patterson D.J., Katz L.A. 2010. Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. Syst. Biol. 59:518–533.

Posada D., Crandall K.A. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50:580–601.

R Development Core Team. 2011. R: a language and environment for statistical computing. [Internet]. Vienna (Austria): R Foundation for Statistical Computing. Available from: URL http://www.R-project.org.

Robinson G.E., Hackett K.J., Purcell-Miramontes M., Brown S.J., Evans J.D., Goldsmith M.R., Lawson D., Okamuro J., Robertson H.M., and Schneider D.J. 2011. Creating a buzz about insect genomes. Science. 331:1386.

Rokas A., and Carroll S.B. 2006. Bushes in the tree of life. PLoS Biol. 4:e352.

Ronquist F., Huelsenbeck J. P. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19: 1572–1574.

Siepel A., Bejerano G., Pedersen J.S., Hinrichs A.S., Hou M., Rosenbloom K., Clawson H., Spieth J., Hillier L.W., Richards S., Weinstock G.M., Wilson R.K., Gibbs R.A., Kent W.J., Miller W., Haussler D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.

Soltis D.E., Moore M.J., Burleigh J.G., Bell C.D., Soltis P.S. 2010. Assembling the angiosperm tree of life: progress and future prospects. Ann. Missouri Bot. Gard. 97:514–526.

Stephen S., Pheasant M., Makunin I.V., Mattick J.S. 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol. Biol. Evol. 25:402–408.

Thomson R.C., Shaffer H.B. 2010. Rapid progress on the vertebrate tree of life. BMC Biol. 8:19.

Townsend J.P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.

Townsend J.P., Leuenberger C. 2011. Taxon sampling and the optimal rates of evolution for phylogenetic inference. Syst. Biol. 60:358–365.

Townsend J.P., López-Giráldez F. 2010. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Syst. Biol. 59:446–457.

Townsend J.P., López-Giráldez F., Friedman R. 2008. The phylogenetic informativeness of nucleotide and amino acid sequences for reconstructing the vertebrate tree. J. Mol. Evol. 67:437–447.

Wang, J., Lee A.P., Kodzius R., Brenner S., Venkatesh B. 2009. Large number of ultraconserved elements were already present in the jawed vertebrate ancestor. Mol. Biol. Evol. 26:487–490.

Zwickl D.J., Hillis D.M. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.