Syst. Biol. 64(6):1032–1047, 2015

DOI:10.1093/sysbio/syv053
Advance Access publication July 29, 2015

Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies

ADAM D. LEACHÉ[1,2,*], BARBARA L. BANBURY[1], JOSEPH FELSENSTEIN[1,3], ADRIÁN NIETO-MONTES DE OCA[4], AND ALEXANDROS STAMATAKIS[5,6]

[1]*Department of Biology, University of Washington, Seattle, WA 98195, USA;* [2]*Burke Museum of Natural History and Culture, University of Washington, Seattle, WA 98195, USA;* [3]*Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA;* [4]*Departamento de Biología Evolutiva, Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad Universitaria, México 04510, Distrito Federal, México;* [5]*Exelixis Laboratory, Scientific Computing Group, Heidelberg Institute for Theoretical Studies (HITS gGmbH), Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany; and* [6]*Department of Informatics, Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Am Fasanengarten 5, 76131 Karlsruhe, Germany*
[*]*Correspondence to be sent to: Adam D. Leaché, Department of Biology, 24 Kincaid Hall, University of Washington, Seattle, WA 98195, USA; E-mail: leache@uw.edu.*

Received 8 January 2015; reviews returned 23 July 2015; accepted 24 July 2015
Associate Editor: Richard Glor

Abstract.—Single nucleotide polymorphisms (SNPs) are useful markers for phylogenetic studies owing in part to their ubiquity throughout the genome and ease of collection. Restriction site associated DNA sequencing (RADseq) methods are becoming increasingly popular for SNP data collection, but an assessment of the best practises for using these data in phylogenetics is lacking. We use computer simulations, and new double digest RADseq (ddRADseq) data for the lizard family Phrynosomatidae, to investigate the accuracy of RAD loci for phylogenetic inference. We compare the two primary ways RAD loci are used during phylogenetic analysis, including the analysis of full sequences (i.e., SNPs together with invariant sites), or the analysis of SNPs on their own after excluding invariant sites. We find that using full sequences rather than just SNPs is preferable from the perspectives of branch length and topological accuracy, but not of computational time. We introduce two new acquisition bias corrections for dealing with alignments composed exclusively of SNPs, a conditional likelihood method and a reconstituted DNA approach. The conditional likelihood method conditions on the presence of variable characters only (the number of invariant sites that are unsampled but known to exist is not considered), while the reconstituted DNA approach requires the user to specify the exact number of unsampled invariant sites prior to the analysis. Under simulation, branch length biases increase with the amount of missing data for both acquisition bias correction methods, but branch length accuracy is much improved in the reconstituted DNA approach compared to the conditional likelihood approach. Phylogenetic analyses of the empirical data using concatenation or a coalescent-based species tree approach provide strong support for many of the accepted relationships among phrynosomatid lizards, suggesting that RAD loci contain useful phylogenetic signal across a range of divergence times despite the presence of missing data. Phylogenetic analysis of RAD loci requires careful attention to model assumptions, especially if downstream analyses depend on branch lengths. [Conditional likelihood; ddRADseq; maximum likelihood; *Phrynosoma*; Phrynosomatidae; reconstituted DNA; SVDquartets]

Restriction site associated DNA sequencing (RADseq) has become a popular method for generating single nucleotide polymorphism (SNP) data sets in non-model organisms, because the method requires little to no prior knowledge of the genome (Baird et al. 2008; Seeb et al. 2011; Peterson et al. 2012). While the number of RAD loci obtained can be large (Davey et al. 2011), they are typically short (e.g., 50–300 base pairs) with each locus having the potential to contain just one or a few SNPs depending on the evolutionary distances separating the samples. Individually, RAD loci are incapable of resolving large gene trees since each one contains a limited number of SNPs. This problem can be circumvented by concatenating RAD loci into a large supermatrix, either using the SNPs alone (Emerson et al. 2010; Yoder et al. 2013) or using the entire RAD locus (Wagner et al. 2013). Deciding whether to include or exclude invariant sites from RAD loci has ramifications for phylogenetic inference. Acquisition bias is the result of nonrandom character sampling, and in the context of SNP-based phylogenetics it is caused by the omission of constant characters (i.e., invariant sites) from the data matrix. Acquisition bias is problematic for phylogenetic inference, because analyzing only variable characters without correction can lead to branch length overestimation and biases in phylogeny inference (Lewis 2001).

Likelihood methods for phylogenies using restriction sites (Felsenstein 1992), later adapted for SNPs (Kuhner et al. 2000) and discrete morphological data (Lewis 2001), provide solutions for analyzing full sequences that are reduced down to SNPs. To correct for the omission of invariant sites, a conditional likelihood is computed that conditions on the exclusive presence of variable sites in the data. Lewis (2001) illustrated the importance of correcting for acquisition bias using computer simulations on a four-taxon tree by demonstrating that branch lengths are overestimated dramatically when this conditional likelihood correction is not utilized. As a consequence of overestimating branch lengths, the probability of reconstructing the correct topology decreases (Lewis 2001). A recent simulation study demonstrated these problems with SNP data, namely, that the exclusion of invariant sites introduces systematic branch length biases and phylogeny reconstruction errors (Bertels et al. 2014). In this study, we investigate how acquisition bias correction models applied to SNP data perform in relation to analyses of full sequences

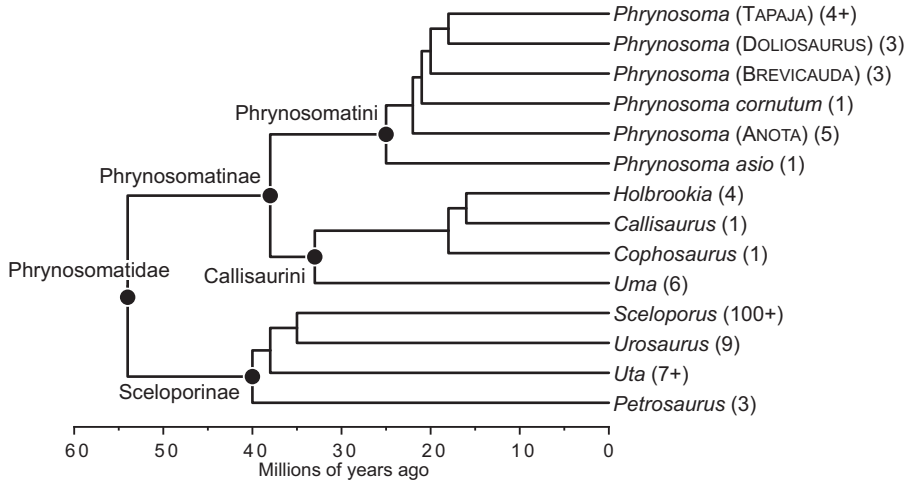

FIGURE 1. Phylogeny for Phrynosomatidae. Topology and clade names follow Leaché and McGuire (2006); Wiens et al. (2010, 2013); Nieto-Montes de Oca et al. (2014); Leaché et al. (2015); Leaché and Linkem (2015). Species numbers are shown in parentheses.

(i.e., SNPs and their flanking invariant sites), and in the presence of allelic dropout (ADO), which can be extensive with RAD loci (Arnold et al. 2013).

We implemented two new acquisition bias correction models in RAxML v.8 (Stamatakis 2014) that are intended for analyses of DNA sequences composed exclusively of SNPs. The first is a conditional likelihood method that is equivalent to the Lewis (2001) Mkv model for binary data, which we extend for the DNA alphabet. For DNA sequence data, it differs from the one-parameter Mkv model, because the variable sites can evolve according to a GTR matrix with five free rate parameters. The conditional likelihood method does not consider the known number of invariant sites that are missing from the data matrix, even if they have been purposefully removed. The second method is a reconstituted DNA approach (Kuhner et al. 2000; McGill et al. 2013) that explicitly specifies the number of invariable sites that are known to be missing from the alignment. The number of invariant sites can be specified for each base separately (i.e., A vs. C vs. G vs. T), or as the total count of all four types of invariant sites. The reconstituted DNA method is straightforward to apply to RAD loci, since the number of invariant sites at each locus is observed and easy to calculate, but high-throughput SNP genotyping methods that only interrogate prescreened SNPs (Thomson 2014) provide no information on invariant sites, which makes the conditional likelihood method a useful alternative. The new acquisition bias correction models that we developed in RAxML are provided in more detail in the Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.t9r3g. Apart from describing the equations we also provide some implementation details such that they can easily be integrated into other likelihood-based (maximum likelihood; ML and Bayesian) tools.

We evaluate the accuracy of SNP-based measures of topology, branch lengths, and support measures using simulations and empirical double digest RADseq (ddRADseq) data for lizards in the family Phrynosomatidae. These new data are used to investigate the ability of RAD loci to resolve clades across a wide range of timescales, from recently diverged species within horned lizards (genus *Phrynosoma*) to deeper evolutionary relationships within the family that approach 40–60 Ma (Wiens et al. 2013). Many aspects of the phrynosomatid phylogeny are essentially "known" based on concordance across previous phylogenetic analyses of morphology and molecular data (Wiens et al. 2010; Leaché et al. 2015; Leaché and Linkem 2015; Fig. 1), which makes this a useful clade for exploring the performance of empirical ddRADseq data. However, the phylogenetic relationships at the base of *Phrynosoma* and the Sceloporinae have proven difficult to resolve with smaller multilocus data sets (Leaché and McGuire 2006; Wiens et al. 2010; Nieto-Montes de Oca et al. 2014). In addition to comparing topologies estimated with full sequences or SNPs under various acquisition correction models, we characterize the branch length and bootstrap support biases produced by data assemblies containing varying levels of missing data. The majority of our comparisons are focused on the implementation of new acquisition bias corrections that are intended for concatenated data, but we also conduct species tree analyses using a coalescent approach.

MATERIALS AND METHODS

Data Simulation

We used computer simulations to investigate the effects of acquisition bias and ADO on phylogenetic inference with SNPs using the new acquisition bias corrections implemented in RAxML. We simulated gene trees and RAD loci using the MCCOAL program (Rannala and Yang 2003). We simulated 50 data sets containing 10 species and 1000 unlinked loci (Fig. 2). The mutation rates were allowed to vary across loci according to

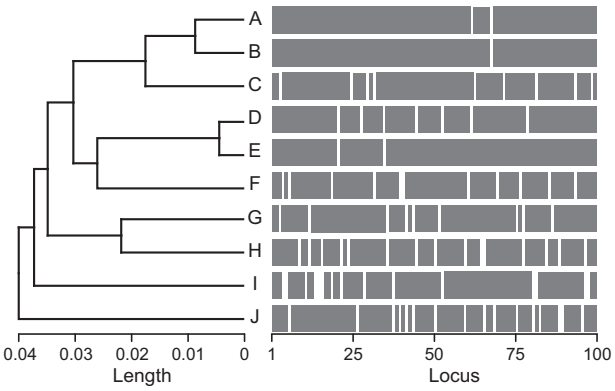

FIGURE 2. Species tree topology used for the simulation of RAD loci with (ADO). The pattern of ADO is illustrated for the first 100 of 1000 loci (black=present, white=ADO). For this example, the data assembly required there to be at least 8 out of 10 sequences per locus (min. ind. = 8) for the locus to be included in the alignment (i.e., the maximum amount of missing data at any locus is 20%).

a gamma distribution with $\alpha = 1$. This is a realistic assumption given that genes evolve at different rates, and that ADO should be more prevalent at loci that evolve more quickly (Huang and Knowles 2014). We also conducted simulations without rate variation to determine whether or not rate variation produces branch length differences between analyses with and without acquisition bias corrections. The gene trees were used to simulate DNA sequences along the branches of the genealogies using the Jukes–Cantor (JC) substitution model (Jukes and Cantor 1969), which is currently the only model available in `MCCOAL`. The species tree population sizes (θ) were set to 0.01 and the root height (τ) parameter was set to 0.04. These values were chosen to approximate the levels of sequence divergence observed in our empirical ddRADseq data, which are equivalent to 1% sequence divergence within populations, and 4% sequence divergence from the root of the tree to any random tip, or a maximum of 8% sequence divergence from tip to tip via the root. The average sequence divergence (uncorrected p-distances) among the 10 species in the simulation was approximately 4%.

We simulated ADO by treating the first 12 bp of each 51 bp locus as a restriction site, and removed any allele that contained a mutation in that region in comparison to the ancestral sequence simulated at the root of the tree. Alleles containing mutations at the restriction site were removed from the data set and replaced with *N* characters to represent missing data (i.e., "*N*" characters for the entire locus). We chose 12 bp to correspond to the combined length of the two restriction enzyme recognition sequences used in our empirical ddRADseq study (see below). The remaining 39-bp loci were concatenated and used for phylogenetic inference using four approaches. First, we ran `RAxML` on the full sequences (i.e., variable and invariant sites included; referred to as "full sequences") under the JC model. The inclusion of invariant sites in the full sequence analysis made the use of an acquisition bias correction obsolete. It is important to note that even loci that lack any variable sites should remain in the analysis of full sequences, since removing those loci would introduce acquisition bias. Second, we removed the invariant sites and analyzed the concatenated SNP data without acquisition bias correction (= uncorrected model; `-m GTRGAMMA --JC69`). The concatenated SNP data included all variable sites (and not just a single randomly chosen SNP from each RAD locus) and no invariant sites (referred to as "SNP data"). Third, we analyzed the SNP data with acquisition bias correction using the conditional likelihood method (`-m ASC_GTRGAMMA --JC69 --asc-corr=lewis`). Finally, we analyzed the SNP data using the reconstituted DNA corrections after obtaining a tally of the number of each of the four invariant sites (`-m ASC_GTRGAMMA --JC69 --asc-corr=stamatakis`), or with the total count of all invariant sites (`-m ASC_GTRGAMMA --JC69 --asc-corr=felsenstein`). For analyses of the simulated data sets that lacked rate variation, we used the JC model and disabled among site rate heterogeneity using the `-V` command in conjunction with `-m GTRCAT --JC69`, which invokes inferences under a simple JC model.

To explore the consequences of ADO on phylogeny estimation using different acquisition bias corrections, we estimated phylogenies for three different assemblies that varied the levels of missing data. The amount of missing data was adjusted by specifying the minimum number of individuals (min. ind.) that were required to have data present at a locus for that locus to be included in the final matrix. For example, a min. ind. of 8 excludes any locus containing <8 individuals with data (Fig. 2). We included a total of 10 species in our simulations, and produced 3 data sets with the min. ind. parameter set to either 4, 6, or 8. This parameter introduces a trade-off with respect to the number of SNPs in the alignment and the amount of missing data; maximizing the number of SNPs also increases the amount of missing data (Wagner et al. 2013).

All post-processing of simulation files was done using scripts that are available on github (https://github.com/bbanbury/phrynomics-data). We created a package in the R environment, `Phrynomics`, that manipulates RAD loci by removing invariant sites, evaluates missing data, and exports files for `RAxML`, `MrBayes`, and `SNAPP`. These tools are also accessible on a web-based graphical user interface (https://rstudio.stat.washington.edu/shiny/phrynomics/).

Taxon Sampling

The lizard family Phrynosomatidae is a diverse group containing nine genera and approximately 148 species. This family originated in the New World and has a broad distribution across North and Central America from southern Canada to Panama, with most diversity centered in the arid regions of Mexico and the

TABLE 1. Species included in the study and a summary of the empirical ddRADseq data

Species	Samples	Inclusive Clade[a]	Loci[b]	Retained[c]	Coverage[d]	Polymorphic[e] (%)
Phrynosoma asio	4	Phrynosomatini	30,833	9819	45.2	0.46
P. blainvillii	4	Anota	54,009	10,641	35.7	0.58
P. braconnieri	4	Brevicauda	34,139	10,349	40.5	0.50
P. cerroense	6	Anota	36,394	11,567	38.4	0.68
P. cornutum	4	Phrynosomatini	40,513	9542	29.7	0.50
P. coronatum	1	Anota	35,443	7596	17.5	0.67
P. ditmarsi	2	Tapaja	28,589	8944	40.0	0.37
P. douglasii	2	Tapaja	38,999	12,822	41.6	0.58
P. goodei	4	Doliosaurus	43,014	9013	38.7	0.62
P. hernandesi	5	Tapaja	30,337	9886	44.3	0.51
P. mcallii	4	Anota	43,656	11,572	35.5	0.54
P. modestum	2	Doliosaurus	43,762	9413	33.5	0.48
P. orbiculare	4	Tapaja	29,807	9041	33.6	0.55
P. platyrhinos	3	Doliosaurus	31,023	9167	52.2	0.53
P. sherbrookei	4	Brevicauda	38,365	7909	25.5	0.38
P. solare	3	Anota	31,030	9270	37.2	0.48
P. taurus	4	Brevicauda	28,361	8753	47.1	0.45
Callisaurus draconoides	2	Callisaurini	33,939	10,217	37.0	0.52
Cophosaurus texanus	1	Callisaurini	31,462	9801	38.1	0.47
Holbrookia maculata	1	Callisaurini	22,325	5241	25.4	0.62
Uma notata	1	Callisaurini	20,029	4483	37.0	0.67
Petrosaurus thalassinus	1	Sceloporinae	43,010	13,735	49.5	0.32
Sceloporus angustus	1	Sceloporinae	42,690	7194	32.1	0.34
S. gadoviae	1	Sceloporinae	23,373	7293	61.3	0.38
S. magister	1	Sceloporinae	29,427	8734	44.3	0.46
S. occidentalis	1	Sceloporinae	23,999	7697	48.0	0.25
Urosaurus bicarinatus	1	Sceloporinae	36,577	7875	65.8	0.58
U. ornatus	1	Sceloporinae	22,563	8958	106.2	0.59
Uta stansburiana	1	Sceloporinae	35,302	11,125	111.3	0.38
Gambelia wislizenii	1	Crotaphytidae	52,045	18,593	75.8	0.58

Note: Values shown for species with multiple samples are averages.
[a]Clade names follow Leaché and McGuire (2006); Wiens et al. (2010); Nieto-Montes de Oca et al. (2014).
[b]Number of loci after clustering reads with a 94% clustering threshold.
[c]Loci retained after passing coverage and paralog filters.
[d]Mean depth of clusters.
[e]Frequency of polymorphic sites.

southwestern United states. The most recent common ancestor of phrynosomatids is dated at approximately 55 Ma based on recent phylogenetic analyses (Wiens et al. 2013), which places the clade within the time frame of where simulation studies suggest that RAD loci should provide accurate phylogenetic relationships (Rubin et al. 2012; Cariou et al. 2013). We sampled a total of 29 species, including one sample each for *Cophosaurus*, *Holbrookia*, *Petrosaurus*, *Uma*, and *Uta*, two samples for *Urosaurus* and *Callisaurus*, four samples for *Sceloporus*, and 60 samples for *Phrynosoma* (representing all 17 species and multiple samples for most species; Table 1; Supplementary Material). Our sampling is focused on *Phrynosoma*, and we consider this the ingroup of our study. The remaining phrynosomatid species are outgroups, and we included *Gambelia wislizenii* (a non-phrynosomatid lizard) to root the trees.

Data Collection

We collected ddRADseq data using the Peterson et al. (2012) protocol. For each individual, we extracted high-molecular weight genomic DNA from liver or muscle tissue using Qiagen DNeasy extraction kits (Qiagen Inc.), checked the quality on agarose gels, and measured concentration using a Qubit fluorometer. We digested 0.5 µg of genomic DNA with 20 units each of a rare cutter SbfI (restriction site 5′-CCTGCAGG-3′) and a common cutter MspI (restriction site 5′-CCGG-3′) in a single reaction with the manufacturer recommended buffer (New England Biolabs) for 4 h at 37 °C. Neither SbfI nor MspI are sensitive to methylation, and thus efficiently digest genomic DNA. Fragments were purified with Agencourt AMPure beads before ligation of barcoded Illumina adaptors onto the fragments. The custom oligonucleotide sequences used for barcoding and adding Illumina indexes during library preparation are provided in the Supplementary Material. Using a combination of eight unique barcodes and 10 Illumina indexes allowed us to multiplex all samples into a single sequencing lane. We produced 10 pools differing by their Illumina index, each containing equimolar amounts of up to eight uniquely barcoded samples. The barcodes differed by at least two base pairs to reduce the chance of errors caused by inaccurate barcode assignment. The 10 pooled libraries were size-selected (between 415 and 515 bp after accounting

for adapter length) on a Blue Pippin Prep automatic size fractionator (Sage Science). Precise size selection is critical with ddRADseq, because it minimizes variation in fragment size-based locus selection among libraries (Puritz et al. 2014). The final library amplification used proofreading Taq and Illumina's indexed primers. The fragment size distribution and concentration of each pool was determined on an Agilent 2200 TapeStation, and qPCR was performed to determine sequenceable library concentrations before multiplexing equimolar amounts of all 10 pools for sequencing on a single Illumina HiSeq 2000 lane (50 bp, single-end run) at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley.

Data Assembly

We processed raw Illumina reads using the program pyRAD v.2.1.7 (Eaton and Ree 2013). An advantage of pyRAD is that it can assemble RAD loci for divergent species using global alignment clustering, which includes indel variation. This is accomplished using the multiple sequence alignment program MUSCLE (Edgar 2004). Several studies have shown that other *de novo* data assemblers designed for processing population genetic data are prone to genotyping errors, and that new methods are needed to take these sources of error into account (Davey et al. 2013; Mastretta-Yanes et al. 2014). The potential sources of error inherent to pyRAD data processing have not been explored in detail, and although we did not conduct an examination of pyRAD data processing here, we note that changing the "threshold" parameters listed below do affect the final locus counts (Pante et al. 2014) as well as the phylogeny (Leaché et al. 2015).

We demultiplexed samples using their unique barcode and adapter sequences, and sites with Phred quality scores under 99% (Phred score = 20) were changed into "*N*"'s, and reads with >10% "*N*"'s were discarded. These filtered reads for each sample were clustered using the program USEARCH v.6.0.307 (Edgar 2010) with a clustering threshold of 88%, and then aligned with MUSCLE. This clustering threshold was selected to reflect the average uncorrected sequence variation observed in phrynosomatid lizard nuclear genes (Leaché et al. 2015). Each locus was reduced from 50 to 39 bp after the removal of the 6-bp restriction site overhang (the restriction enzyme sequences are 12 bp, but only one 6-bp overhang is sequenced) and the 5-bp barcode. As an additional filtering step, consensus sequences were discarded that had either: (i) low coverage (<6 reads), (ii) excessive undetermined or heterozygous sites (>3), or (iii) too many haplotypes (>2 for diploids). The consensus sequences were clustered across samples using the same procedure and thresholds. Finally, each locus was aligned with MUSCLE, and loci with >10 samples sharing heterozygosity at a site were treated as paralogs and discarded. The justification for this paralog filtering is that if a site is heterozygous across some large number of species, then it is more likely to be a fixed difference among paralogs that all those samples share rather than a true heterozygous site that is shared among species.

TABLE 2. Summary of ddRADseq data matrices for phrynosomatid lizards

Matrix[a]	Loci	Variable sites[b]	Missing (%)[c]	Average data overlap (%)[d]	
				Phrynosoma	non-*Phrynosoma*
s70	6	37	5.7	95	76
s65	27	174	9.5	88	61
s60	73	471	14.1	81	47
s55	160	1035	19.5	73	38
s50	300	1915	24.9	65	28
s45	526	3341	30.3	57	21
s40	904	5652	36.2	49	15
s35	1447	8843	41.7	42	11
s30	2219	13,232	47.2	35	9
s25	3327	19,152	52.9	29	6
s20	4954	27,195	58.8	23	5
s15	7506	38,539	65.3	17	3
s10	12,701	59,111	73.6	11	2
s5	25,709	101,937	83.7	6	1

[a]Minimum number of individuals (min. ind.) needed to retain a locus (total number of samples is 74).
[b]Includes all SNPs within a locus, which are presumed to be linked.
[c]Percentage of total data matrix cells with missing data.
[d]Average percentage of SNPs shared across taxa for all loci.

We exported 14 data sets that contained varying levels of missing data, which were adjusted using the min. ind. parameter. The empirical data included 74 individuals, and we set the strictest limit on missing data to 70 (matrix s70). Only six loci met this requirement, and the total amount of missing data in these loci was low (5.7%; Table 2). At the opposite extreme, matrix s5 required that only five individuals had data at a locus, and this matrix included over 25,000 loci and 83% missing data. A summary of the final data sets is provided in Table 2.

Maximum Likelihood Phylogenies

We inferred ML phylogenies using RAxML. We used the K80 model of nucleotide substitution without rate heterogeneity for all 14 data sets (`-m GTRCAT -V --K80`). This model of sequence evolution was the best fit for the concatenated ddRADseq data (using matrix s50), determined using jModelTest (Darriba et al. 2012). For each data set, we conducted ML analyses using either full sequences or with SNP data. The SNP data were analyzed using three models: (i) uncorrected (`-m GTRCAT -V --K80`), (ii) conditional likelihood (`-m ASC_GTRCAT -V --K80 --asc -corr=lewis`), and (iii) reconstituted DNA (`-m ASC_GTRCAT -V --K80 --asc-corr =stamatakis`). For each analysis, branch support was estimated using the automatic bootstrap convergence function that calculates a stopping rule to determine

when enough replicates have been generated (Pattengale et al. 2010).

Comparisons of Branch Lengths, Topologies, and Support Values

We extracted measures of total tree length from results files (RAxML info files). For each phylogeny, we compared individual branch lengths and support values for taxon bipartitions that were shared (i.e., no additional or missing tips) between the full sequences and SNP analyses. We use the results from the analysis of the full sequences as the basis for all of our tree metric comparisons (as the true tree and branch lengths are not known for the empirical data). We extracted branch lengths and bootstrap values directly from RAxML bipartition tree files (the tree file containing the best scoring ML tree with branch lengths and bootstrap values without branch labels). We also calculated branch length differences in terms of error (or relative bias) between the full sequences and SNP analyses. For example, for each shared branch we compared the relative bias of the uncorrected analysis in relation to the full sequences analysis by dividing the difference in branch lengths by the branch length from the full sequences: [(uncorrected–full sequences)/full sequences].

We quantified topological differences between the full sequences and SNP analyses (uncorrected, conditional likelihood, and reconstituted DNA) using Robinson–Foulds (RF) distances (Robinson and Foulds 1981) calculated in the Phangorn v.1.99-7 package in R (Schliep 2011). We used symmetric RF distances, which exclude branch lengths, and thus focus solely on topological comparisons. The RF distances were divided by the total number of branches to obtain relative RF values. We scripted all post-analysis comparisons using custom R scripts to ensure reproducibility and to avoid errors. These scripts and the associated functions are available on GitHub (https://github.com/bbanbury/phrynomics-data).

Species Tree Estimation

We used the program SVDquartets v.1.0 (Chifman and Kubatko 2014) to estimate a species tree using the RAD loci. The method infers the relationships among four species at a time (=quartets) under a coalescent model. The reduction of the species tree inference problem down to quartets makes the method well suited to RAD loci that contain high levels of missing data with few shared loci among all species. Operationally, the method works in two steps. First, quartets are randomly sampled from the data matrix, and for each quartet the singular value decomposition (SVD) score is calculated to evaluate the optimal or "true" relationship for the sampled quartet (Chifman and Kubatko 2014). Second, a quartet reconstruction program is used to infer the species tree. Uncertainty in relationships is quantified using nonparametric bootstrapping (Chifman and Kubatko 2014).

We applied SVDquartets to three data matrices, s5, s25, and s50. For each data matrix, we randomly sampled 100,000 quartets from the 74 sampled individuals. The quartet program Quartet MaxCut v.2.1.0 (Snir and Rao 2012) was used to infer the species tree from the sampled quartets. We used nonparametric bootstrapping with 100 replicates to measure uncertainty in bipartitions. The bootstrap values were mapped to the species tree estimated from the original data matrix using SumTrees v.3.3.1 (Sukumaran and Holder 2010).

Results

Data Simulation

We examined the accuracy of acquisition bias corrections in response to different levels of missing data using simulated data containing 1000 loci (Fig. 3a). For the simulations with the least missing data (min. ind. = 8), the analysis of full sequences, and SNPs with acquisition bias correction (conditional likelihood or reconstituted DNA), provided tree length (TL) estimates that were close to the true TL (true TL = 0.264). For the same SNP alignment, the uncorrected model overestimates the TL over 4-fold (Fig. 3a). This result coincides with the original Lewis (2001) study; the uncorrected model overestimates branch lengths in comparison to the corrected model. We have extended this result to SNP data using a DNA-based model, and show that using full sequences or a corrected model (i.e., with acquisition correction) can provide similar TL estimates when there is little missing data. However, increasing the amount of missing data results in increasing TL overestimation for the uncorrected model and for the conditional likelihood method, although the effect is not as profound for the latter (Fig. 3a). The reconstituted DNA approach only slightly underestimated TL with increasing levels of missing data (Fig. 3a).

We examined the properties of simulated RAD loci containing different amounts of missing data to determine the contribution of loci that contain high levels of missing data towards combined phylogenetic analyses. For instance, how accurate are phylogenetic trees estimated with loci containing 60% missing data? Topological discordance, as measured by relative RF distances, increases with the amount of missing data at a locus (Fig. 4a). Similarly, the branch length estimation error also increases with the amount of missing data (Fig. 4b), and bootstrap values decrease as well (Fig. 4c). Loci with missing data do not necessarily contribute phylogenetic signal to particular depths of a phylogenetic tree. Instead, the simulated data suggest that missing data limits the ability of a locus to accurately estimate shorter branches (Fig. 4d). Concatenating loci with missing data together with complete loci, which is

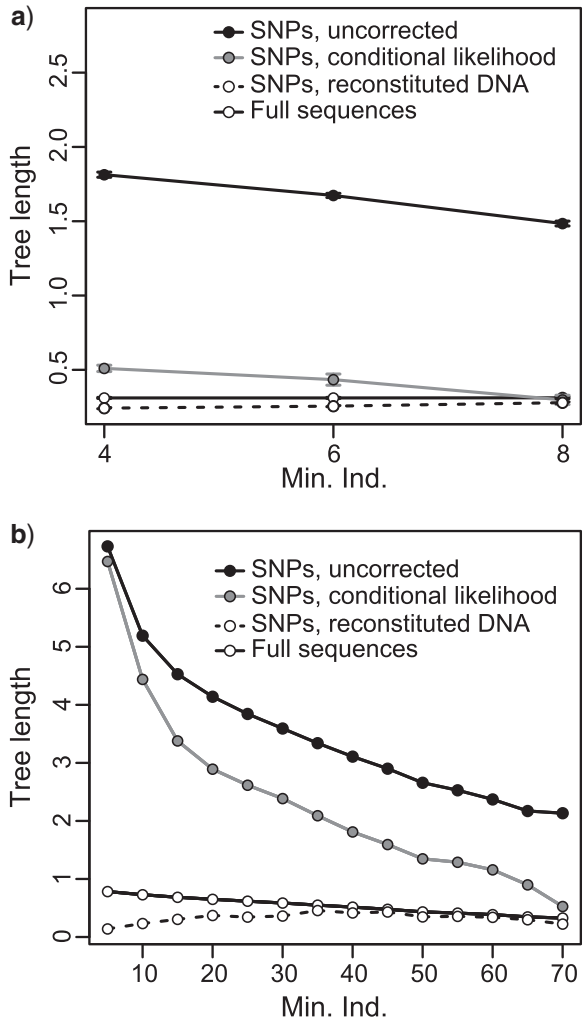

FIGURE 3. Tree lengths estimated using acquisition bias correction are sensitive to allelic dropout. Simulations (a) and empirical data for phrynosomatid lizards (b) show similar patterns. The tree length is overestimated by the conditional likelihood correction and underestimated by the reconstituted DNA correction. The true tree length for the simulation is 0.264. Simulations without locus rate variation (a) show similar patterns to those that include rate variation.

the common practise for RAD loci, does not appear to negatively impact any of the measures of phylogenetic accuracy that we examined.

Empirical ddRADseq Data

One lane of Illumina HiSeq2000 sequencing produced 46,173,267 reads that could be demultiplexed and assigned to the 74 samples included in our analysis. The samples each had >147,000 reads (mean = 712,301), and the number of loci retained after quality filtering exceeded 4483 (mean = 9542) and had high sequencing coverage >17.5 × (mean = 46×) (Table 1).

The characteristics of the 14 data matrices produced for this study are outlined in Table 2. Despite the recovery of thousands of loci for each sample (Table 1), no shared loci were recovered across all 74 samples. More shared loci are recovered after reducing the min. ind. parameter, but decreasing this threshold introduces more missing data (Table 2). For example, matrix s70 (requiring 70 of the 74 samples to have data at each locus) contains only 5.7% missing data, but only contains six loci and 37 variable sites (Table 2). Conversely, matrix s5 (requiring only five of the 74 samples to have data at each locus) is composed of 83.7% missing data, but contains >25,700 loci and >101,900 variable sites (Table 2). The average number of SNPs shared among ingroup samples exceeds the number of SNPs shared among the outgroup samples (Table 2).

Acquisition Bias and Branch Lengths

We estimated phylogenetic trees for phrynosomatid lizards with 14 data matrices that contained varying levels of missing data (Table 2) using four approaches in RAxML. The effects of acquisition bias on TL are shown in figure 3b; the patterns are similar to the simulation results, although the errors are more pronounced. The uncorrected model produces the largest deviations in TL in comparison to the estimates from full sequences, and using the conditional likelihood method to correct for acquisition bias also overestimates TL. The effect is more pronounced for data sets that include more missing data (Fig. 3b). The TL estimates obtained using the reconstituted DNA correction are almost identical to those from full sequences up to a point, but the reconstituted DNA correction underestimates TL for the data sets with the largest amounts of missing data.

We compare individual branch length estimates from the full sequences and SNP analyses in figure 5. Branch lengths estimated with the uncorrected model and the conditional likelihood method are longer compared to the full sequences. The conditional likelihood method overestimates branches by 100%. The reconstituted DNA approach shows more variability in individual branch length estimates in comparison to the full sequences, and the general trend is to underestimate the longest branches.

Relative branch length differences are not distributed evenly across the phylogeny, and the degree of missing data plays a role in the location of the biases (Fig. 6). In comparison to branch lengths estimated using full sequences, SNP estimates of branch lengths are overestimated using the conditional likelihood correction, and overestimation is more severe in the ingroup. Once nearly 100K SNPs are present, and the alignment contains 84% missing data, overestimation is >500% for almost all branches (Fig. 6a). Branch length biases are not as severe for the reconstituted DNA correction (Fig. 6b). For example, most branches are within 25% of the full sequences branch lengths. However, the reconstituted DNA correction tends to underestimate branches, and the problem is more severe for the non-*Phrynosoma* taxa than within *Phrynosoma* (Fig. 6b).

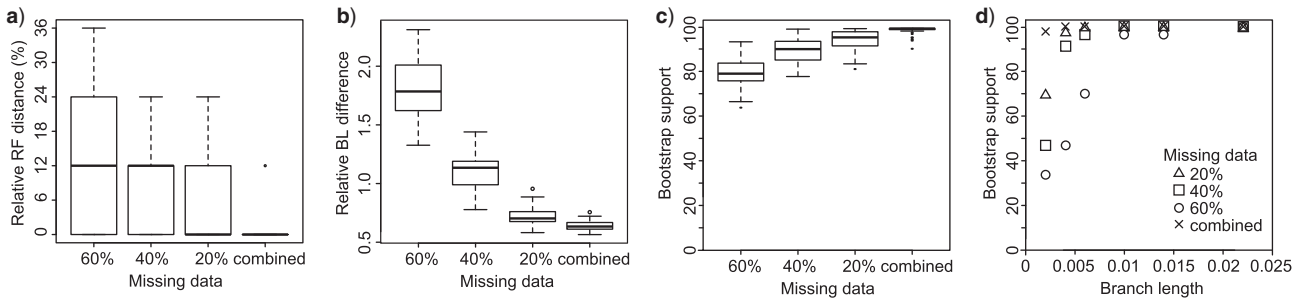

FIGURE 4. Properties of simulated RAD loci with different amounts of missing data. Loci that contain more missing data tend to result in discordant topologies (a), increased branch length errors (b), and lower bootstrap support (c). Loci that contain less missing data provide higher bootstrap support for shorter branches (d).

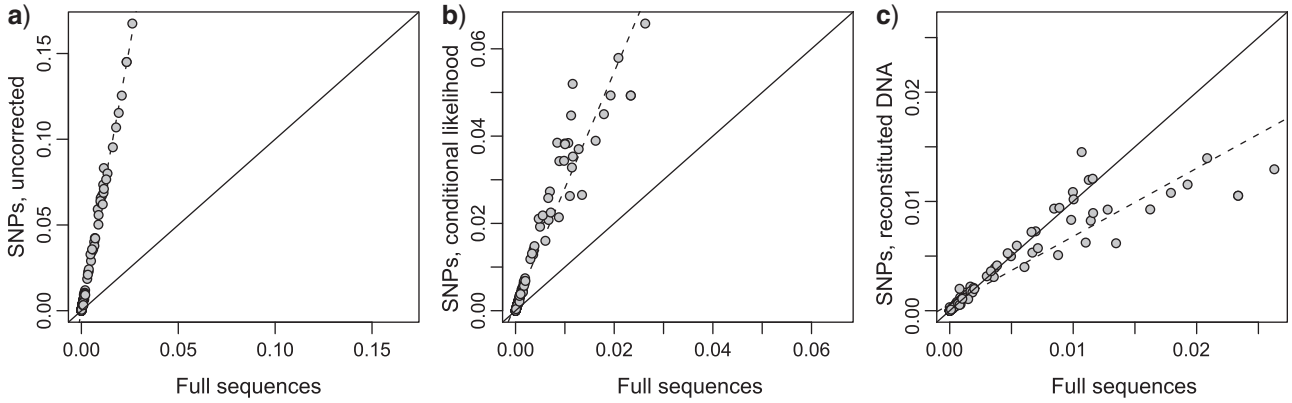

FIGURE 5. Comparisons of branch lengths estimated from the empirical phrynosomatid lizard data. In comparison to the analysis of full sequences (*x*-axis), branch lengths are overestimated when no acquisition bias correction is used (a), overestimated with the conditional likelihood correction (b), and underestimated with the reconstituted DNA correction (c). Results are shown for the s50 data matrix, which contains 1915 variable sites.

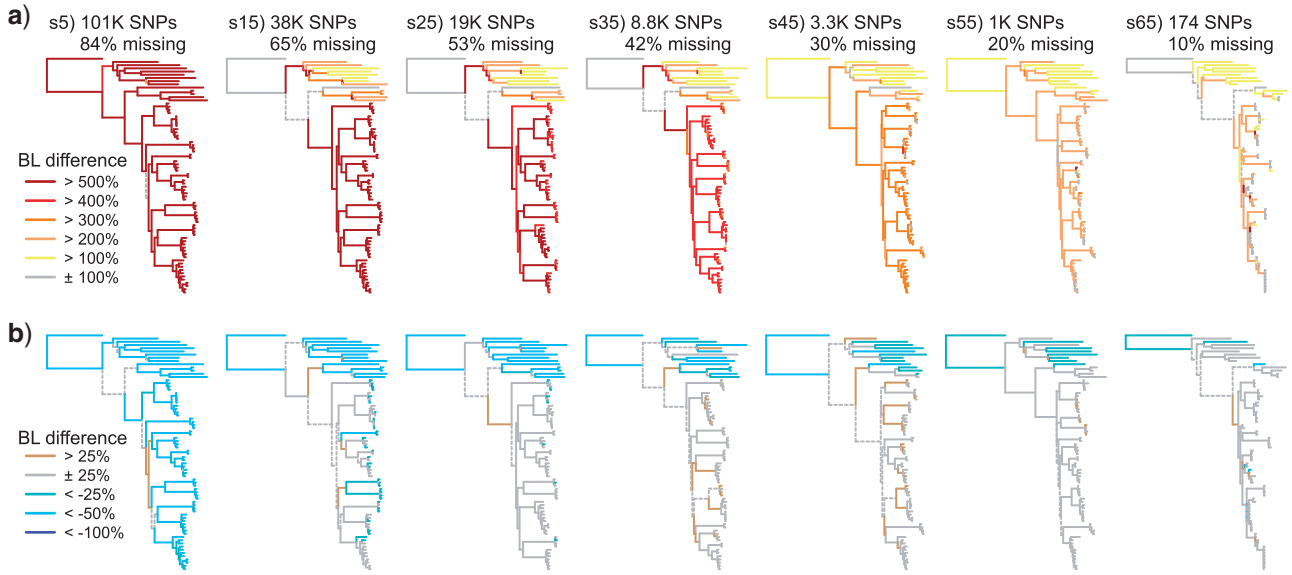

FIGURE 6. Biases in branch lengths (BLs) on phylogenies for phrynosomatid lizards increase as the size of the data matrix increases. Branch colors reflect the relative BL difference between the analysis of full sequences and the conditional likelihood correction (a), and the reconstituted DNA correction (b). Positive values indicate longer branches under the acquisition correction model, and negative values indicate shorter branches under the acquisition correction model. Branches with dashed lines indicate discordant bipartitions.

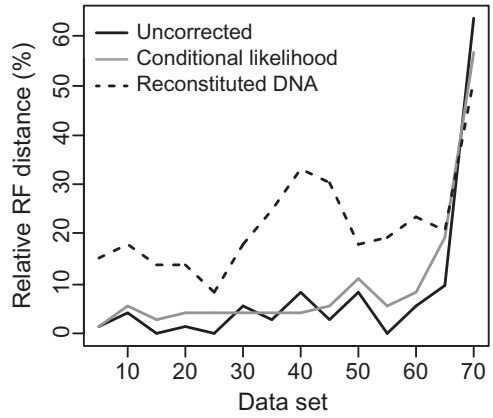

FIGURE 7. Relative RF distances between phrynosomatid lizard topologies estimated with full sequences versus topologies estimated with SNPs with no acquisition bias correction (=uncorrected), the conditional likelihood correction, and the reconstituted DNA correction.

Acquisition Bias and Topologies

The relative RF distances between topologies estimated using full sequences and SNPs are shown in Figure 7. Topologies estimated with full sequences and SNPs were more similar to one another (<10%) when analyzing SNPs without a correction or with the conditional likelihood correction compared to the reconstituted DNA correction. The SNP topologies estimated with the reconstituted DNA correction were typically >10% different from the topologies estimated using full sequences, and the variability in the relative RF distances was high across the data sets (Fig. 7). The topologies estimated with the smallest data set (s70) produced the most discordant topologies (Fig. 7).

Acquisition Bias and Bootstrap Support

We compared the support values for all shared bipartitions between the full sequences and SNP analyses (Fig. 8). We should expect to see an unbiased relationship between the support values that these models provide for shared bipartitions. However, the uncorrected model and the conditional likelihood method both tend to provide higher bootstrap support compared to full sequences. Some extreme outliers are present in both comparisons. For example, a bipartition that received 100% bootstrap in the uncorrected analysis received 10% support from full sequences (Fig. 8a). The reconstituted DNA method only slightly overestimates bootstrap support for shared bipartitions (Fig. 8c).

Bootstrap support values for the major relationships within phrynosomatid lizards are shown in Table 3. The patterns of bootstrap support can be separated into three categories. First, three clades receive strong support (i.e., >90%) across all analyses of all alignments, including Phrynosomatini, Brevicauda, and Tapaja. Second, three of the clades are not supported by these data, including Anota, Doliosaurus, and *Sceloporus*. Finally, three remaining clades (Sceloporines, Phrynosomatinae, and Callisaurini) receive mixed support based on either the analysis type or the size of the data matrix. The bootstrap values estimated using the acquisition bias corrections are sensitive to the size of the data set (and amount of missing data), and they each show variation in the bootstrap support for Phrynosomatinae and the Callisaurini.

Phrynosomatidae Phylogeny

Summaries of the phylogenetic trees from analyses of the 14 data matrices using full sequences and SNPs are provided in the Supplementary Material. The ML phylogenetic analysis of the largest data matrix (s5) using full sequences is shown in Figure 9. There is strong support for the major clades in the family, including Sceloporinae, Phrynosomatinae, Callisaurini, and Phrynosomatini. Within Sceloporinae, *Sceloporus* is paraphyletic and includes *Urosaurus*. Within *Phrynosoma*, all of the species are monophyletic, except *P. cerroense*. Two of the four *Phrynosoma* species groups are monophyletic, including Brevicauda and Tapaja. Relationships within *Phrynosoma* that are weakly supported (bootstrap <70%) include the initial divergence within the genus, and the relationships among the species belonging to Anota and Doliosaurus.

The species trees estimated using `SVDquartets` are largely similar to one another and to the expected phylogeny in Figure 1, with a few notable exceptions (Fig. 10). First, the relationships among the sand lizard genera *Callisaurus*, *Cophosaurus*, and *Holbrookia* are not consistent. The earless lizard genera *Cophosaurus* and *Holbrookia* form a clade when the largest data matrix is used, but *Callisaurus* and *Holbrookia* form a clade with the other data matrices (Fig. 10). Second, the relationships within *Phrynosoma* vary with different data assemblies. The first split within *Phrynosoma* either leads to *P. cornutum*, *P. asio*, or a clade containing six species (Fig. 10), depending on the data matrix analyzed. None of the species trees provide strong support for the initial divergences within *Phrynosoma*. Finally, the species trees do not support the monophyly of the genus *Sceloporus*, and this is a result of *S. angustus* forming a clade with either *Urosaurus* or *Uta*.

DISCUSSION

Collecting SNP data for non-model organisms is becoming faster and easier with RADseq methods (Cruaud et al. 2014; Puritz et al. 2014), but these data are introducing new challenges to consider for the application of SNPs to phylogenetic studies. SNPs are useful markers for population genetic studies and phylogenetic analyses of recently diverged species (Brumfield et al. 2003; Liu et al. 2014; Spinks et al. 2014; Streicher et al. 2014), and they can provide accurate measures of population parameters (Kuhner et al. 2000; Arnold et al. 2013). In terms of their phylogenetic utility,

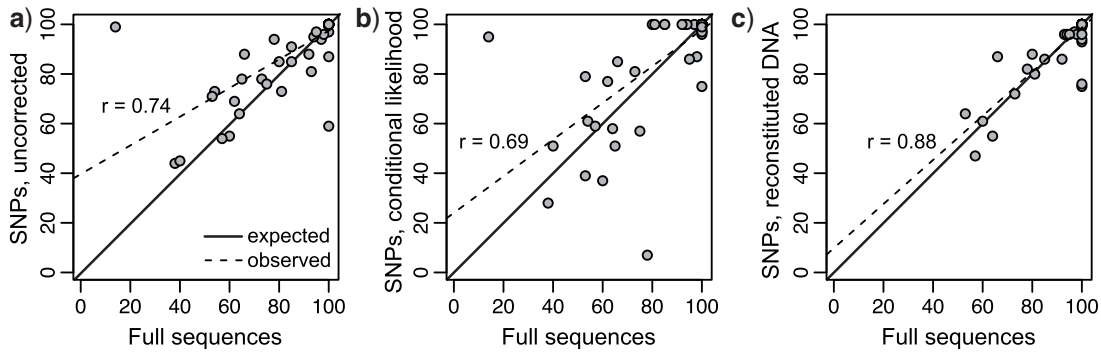

FIGURE 8. Comparison of bootstrap support values from analyses of phrynosomatid lizards using full sequences versus SNPs with no acquisition bias correction (a), the conditional likelihood correction (b), and the reconstituted DNA correction (c). Results are shown for the largest data matrix (s5). On average, analyses of SNPs tend toward slightly higher bootstrap values.

TABLE 3. Bootstrap support for relationships within Phrynosomatidae

Clade	Full sequence	Uncorrected	Conditional likelihood	Reconstituted DNA	Species tree
Sceloporinae	83/100/100	82/99/87	76/98/75	–/97/96	100/98/99
Phrynosomatinae	99/100/100	96/99/100	84/–/97	79/–/–	100/100/100
Callisaurini	98/100/100	96/99/100	84/–/97	79/–/–	82/89/92
Phrynosomatini	100/100/100	100/100/100	100/100/100	100/100/100	100/100/100
Anota	–/–/–	–/–/–	–/–/–	–/–/–	–/–/–
Brevicauda	100/100/100	100/100/100	100/100/100	100/100/100	100/100/100
Doliosaurus	–/–/–	–/–/–	–/–/–	–/–/–	–/–/–
Tapaja	100/100/100	99/100/100	100/100/100	99/100/100	93/99/99
Sceloporus	–/–/–	–/–/–	–/–/–	–/–/–	–/–/–

Notes: Bipartitions that were absent are represented by a "–".
Results are shown for three data matrices: s50/s25/s5.

some studies have suggested that RAD loci are capable of providing accurate interspecific relationships for clades as old as 40–60 myr (Rubin et al. 2012; Cariou et al. 2013). However, these studies obtained RAD loci *in silico* from sequenced genomes, which might provide overly optimistic results compared to empirical data collected in a molecular lab. Our investigation of empirical ddRADseq data for phrynosomatid lizards, together with computer simulations, have helped us explore biases in the phylogenetic analysis of RAD loci. The decision to analyze full sequences versus stripping the data down to SNPs is important, and we implemented new acquisition corrections to aid the analysis of SNP data.

Acquisition Bias Corrections

We developed two new acquisition bias correction models that help, in part, to deal with the challenge of analyzing alignments composed exclusively of SNPs. The conditional likelihood correction can deal with situations where the number of unsampled invariant sites are unknown, and the reconstituted DNA method explicitly incorporates the exact number of unsampled invariant sites. For RAD loci, the latter approach is more appropriate, since the invariant sites are sequenced along with the SNPs. Current acquisition bias corrections can effectively account for the lack of invariant sites in an alignment, but they cannot correct for missing data. One important refinement that should improve the accuracy of the reconstituted DNA method would be to designate ADO samples (which are missing data, not invariant sites) versus samples that are missing invariant sites. The current implementation adds the same number of invariant sites to each sample without accounting for missing data. This data reduction step greatly decreases the number of distinct alignment patterns in the data set, which has the benefit of reducing run-times (Fig. 11), but it affects the topology (Fig. 7). The addition of large amounts of invariant data to the alignment is the likely cause of the branch length underestimation and topological discrepancies that we measured. Extending the acquisition bias corrections to accommodate missing data to determine if this is the cause of the differences in tree topologies that we observed is nontrivial implementation wise, but worthwhile to explore in the future.

Our empirical ddRADseq data for phrynosomatid lizards suggests that stripping RAD loci of invariant sites can lead to branch length overestimation if not modeled correctly, and that acquisition bias correction can alleviate some of this overestimation error. Analyzing variable sites without any correction produces the most extreme tree/branch length overestimation (Lewis 2001), and we see this problem with simulated and empirical data (Fig. 3). The two acquisition bias corrections that we investigated, a conditional likelihood

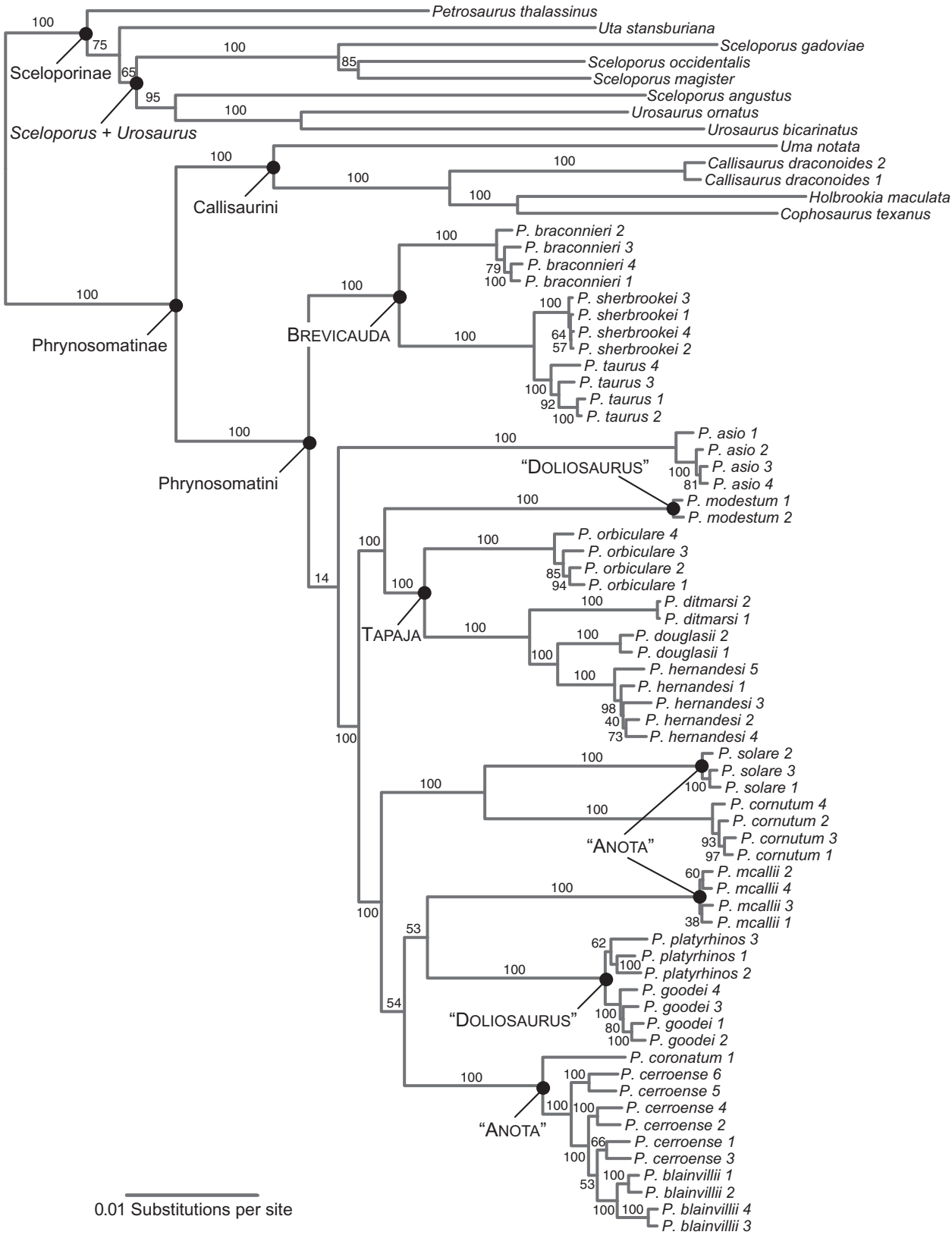

FIGURE 9. Phylogeny of phrynosomatid lizards based on an ML analysis of full sequences (matrix s5: 1,256,221 base pairs, 25,709 loci, and 101,937 variable sites). Bootstrap values are shown on the branches.

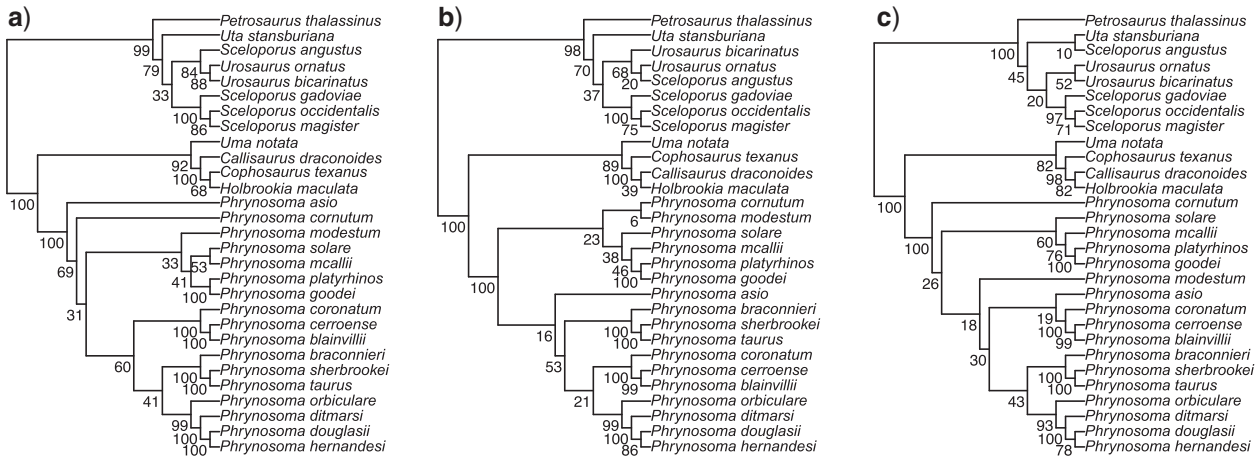

FIGURE 10. Species trees for Phrynosomatidae estimated using SVDquartets for data matrix s5 (a), s25 (b), and s50 (c). Bootstrap values (from 100 replicates) are shown on nodes.

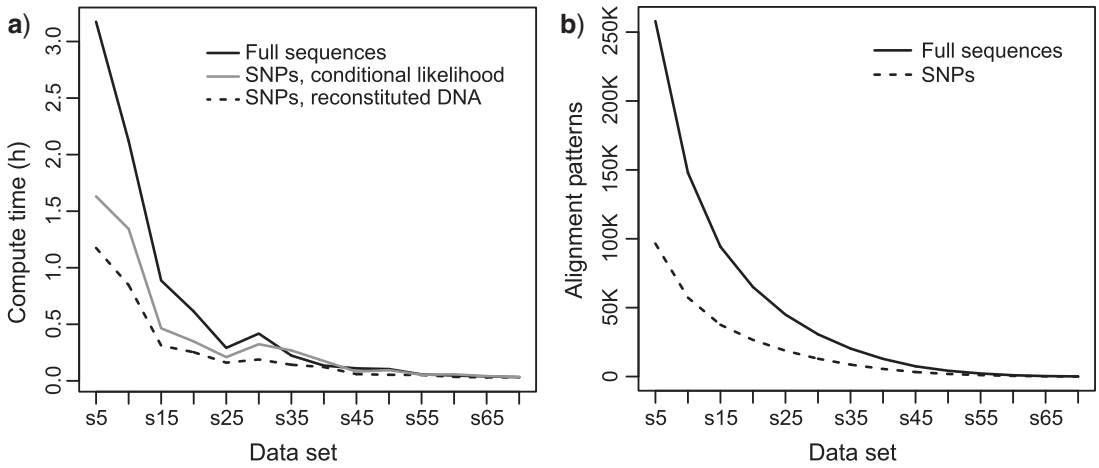

FIGURE 11. RAxML search times are faster for acquisition bias correction models, especially for larger data matrices (a), and the speed increase is a result of removing thousands of distinct alignment patterns from the data matrix that are produced by the missing data (b). Compute times exclude bootstrap calculations. All analyses were run on 16-core Intel E5-2650 CPUs with 32GB of RAM.

and the reconstituted DNA approach, both help reduce overestimation problems (Fig. 3). These acquisition bias corrections provide more accurate branch length estimates for alignments containing less missing data, but biases increase as the alignments become larger with more ADO (Fig. 3). The conditional likelihood method overestimates branch lengths, while the reconstituted DNA approach slightly underestimates them. The degree of branch length overestimation varies across the phrynosomatid phylogeny, and branch lengths are overestimated by 500%, or underestimated by 50% (Fig. 6). Branch length estimation errors of this magnitude will have negative impacts on downstream comparative analyses that utilize branch length information. For example, nonrandom lengthening of branches across the phylogeny can skew node densities across the tree and mislead diversification analyses (Burbrink and Pyron 2011). Researchers intending to use branch length information from SNP phylogenies for divergence dating, diversification studies, or comparative analysis should be cautious when using acquisition bias corrections, and should conduct analyses using different assemblies of the data to determine if estimates are sensitive to the number of SNPs and the amount of missing data. Users of software pipelines that automatically assemble RAD loci and generate phylogenies (Bertels et al. 2014; Lee et al. 2014) should be careful to verify that the proper models are being used for phylogeny estimation, since default settings may not be appropriate for data sets composed entirely of SNPs.

Missing Data and ADO

Empirical RAD loci contain large amounts of missing data, and this problem is more pervasive for distantly related species due to ADO (Arnold et al. 2013). The amount of missing data in the final data matrix is a variable that is controlled by the user, and while setting

a low tolerance for missing data will maximize data overlap, this comes at the cost of retaining far fewer loci (Wagner et al. 2013). Thus, the number of SNPs acquired is positively correlated with missing data. Relying on the data assemblies that minimize missing data contained relatively few SNPs, and those matrices produced the most discordant phylogenies (Fig. 7), probably as a result of having too few characters to resolve the tree (Brandley et al. 2009). The alignments containing the most SNPs produced the most similar topologies for the full sequences, uncorrected, and conditional likelihood models, but these latter approaches also suffer from the most severe branch length and bootstrap value overestimation. Conversely, the reconstituted DNA method provided the most accurate branch lengths and bootstrap values, but the topologies were the most dissimilar compared to topologies from full sequences (Fig. 7).

We expect that there is a strong pattern of average rate differences for matrices of different size. It is possible that RAD loci that are sequenced across the majority of samples are also the most slowly evolving, since these loci lack mutations at restriction sites. Conversely, incompletely sampled loci are more likely to be fast evolving and therefore will be missing sequences from divergent lineages (Huang and Knowles 2014). Rate differences could produce a pattern of missing data resulting from ADO that is strongly nonrandom with respect to taxonomic position and rate of evolution. This specific pattern of missing data can lead to biases in inferred branch lengths and, if sufficiently extreme, topology errors (Lemmon et al. 2009). These problems make the selection of the "best" assembly a difficult problem, since the decision is likely to be a balance between obtaining a large number of SNPs and minimizing missing data. The choice has ramifications on the final topology, branch lengths, and bootstrap support values (Table 3, Fig. 10).

We used our simulated data to characterize the utility of RAD loci containing different levels of missing data and found that as the amount of missing data at a locus increases, the bootstrap support decreases (especially for short branches), topological accuracy decreases, and branch length estimation errors increase (Fig. 4). A recent study by Huang and Knowles (2014) suggested that RAD loci containing high levels of missing data have the advantage of increasing the bootstrap support for shallow divergences. Our simulations suggest that RAD loci with high levels of missing data increase the support for relatively long branches on the phylogeny (Fig. 4d), and not just shallow divergences. The typical procedure for compiling RAD loci entails concatenating loci with different levels of missing data together with more complete loci in a cumulative fashion instead of analyzing loci with large or small amounts of missing data on their own. Although the concatenation approach does lead to higher bootstrap support, we think that it is important to realize that RAD loci containing high levels of missing data are the most error-prone (Fig. 4).

Simulation studies using full sequences have shown that the addition of missing characters to a data matrix can lead to inaccurate bipartitions that are strongly supported, and that it is difficult to distinguish these misleading results from real signal (Lemmon et al. 2009). We addressed the question of whether RAD loci with high levels of missing data produce strongly supported spurious results using our empirical data for phrynosomatid lizards by estimating phylogenetic trees for loci containing different levels of missing data. We estimated a phylogenetic tree in RAxML using only those loci with 90–95% missing data, and the phylogeny provides 100% bootstrap support for at least five incorrect clades that contain outgroup and ingroup species (results not shown). We also tracked the support values for several focal clades of interest in the phrynosomatid phylogeny to investigate the sensitivity of bootstrap values to the size of the data matrix (Table 3). The general pattern that emerges is that concatenating more RAD loci, which at the same time increases the amount of missing data, increases bipartition support. This result is not surprising, since large matrices that contain high levels of missing data tend to produce strong bootstrap support in empirical phylogenies (Emerson et al. 2010; Eaton and Ree 2013; Wagner et al. 2013) and in simulation (Huang and Knowles 2014). We are skeptical about the high levels of support provided by concatenated RAD loci given that concatenation of multilocus data almost always returns high statistical support (Rokas and Carroll 2006; Salichos and Rokas 2013; Simmons and Goloboff 2014) instead of reflecting the true levels of support or incongruence inherent to the data (Seo 2008; Kumar et al. 2012; Salichos et al. 2014).

Phrynosomatidae Phylogeny

The ddRADseq data presented here do not provide a definitive phylogeny for phrynosomatid lizards. The topology changes with different numbers of RAD loci, and the bootstrap support varies across data sets as well (Table 3, Fig. 10). The analysis of full sequences with the largest concatenated data set (Fig. 9) supports many of the accepted relationships for the family (Fig. 1), but important ambiguities remain. The relationships within Sceloporinae are congruent with estimates from previous studies (albeit with low support), although these data do not support the monophyly of *Sceloporus*. In general, the longest branches in the phylogeny were seemingly "easy" to reconstruct even with limited numbers of RAD loci, including Sceloporinae, Phrynosomatinae, Callisaurini, Phrynosomatini, and Brevicauda (Table 3). However, the relationships among the genera within Sceloporinae (e.g., *Uta*, *Petrosaurus*, *Urosaurus*, and *Sceloporus*) are among the most difficult to resolve in the entire family, and the support for these short branches was inconsistent and dependent on the size of the data matrix (Table 3). A recent study of phrynosomatid lizard

relationships using targeted sequence capture data and ddRADseq data also found that these short branches in the phylogeny were the most difficult to resolve, and that different pyRAD assemblies have the potential to support conflicting topologies, often with strong support (Leaché et al. 2015).

The phylogenetic relationships among *Phrynosoma* that we estimated from ddRADseq data in this study reaffirm the close relationships among some species at shallow levels of the tree (Fig. 10). These include the placement of *P. sherbrookei* within Brevicauda, and the specific relationships within Tapaja; these relationships are congruent with previous analyses of nuclear loci and mtDNA (Leaché and McGuire 2006; Nieto-Montes de Oca et al. 2014; Leaché and Linkem 2015). These ddRADseq data, whether analyzed using full sequences or with SNPs, fail to support the monophyly of Anota and Doliosaurus (Table 3, Fig. 10), two clades that are supported by phylogenetic analyses of concatenated nuclear gene sequences (Leaché and McGuire 2006; Nieto-Montes de Oca et al. 2014; Leaché and Linkem 2015). The initial divergences within *Phrynosoma* appear to have occurred in rapid succession (Figs. 9 and 10). The coalescent-based species tree analyses provide weak support for these short branches (Fig. 10), which could be an indication that there is incongruence among the RAD loci, presumably from incomplete lineage sorting. However, the concatenation analysis squelches this signal and provides strong support for several of these short branches (Fig. 9). Determining whether incomplete lineage sorting is responsible for lowered support values in the coalescent analysis, or if ADO is responsible for increased support values in the concatenation analysis, or if a mix of both is occurring, will have important implications for future phylogenetic studies of RAD loci. There is a need for the continued development of species tree estimation approaches that can handle large SNP data sets that contain high levels of missing data.

Conclusions

The use of SNP data in phylogenetics is increasing as reduced representation library sequencing approaches (like RADseq) become more common. An assessment of the best practises for using these data in a phylogenetic context is important for identifying the costs and benefits associated with different approaches. We developed and assessed two new acquisition bias corrections for SNP-based phylogenetic analysis, and compared these approaches to phylogenetic analyses of full sequences. We found that using full sequences from RAD loci is preferable to omitting invariant sites and analyzing the SNPs on their own. Analyzing SNPs comes with the benefit of decreasing computation times when removing thousands of sites with missing data, but branch length and topological accuracy are compromised when using the acquisition bias correction models. Despite these drawbacks, the conditional likelihood and reconstituted DNA corrections provide new alternatives for the phylogenetic analysis of exceptionally large data sets that are often prohibitively slow with full sequences. The acquisition bias corrections are both sensitive to missing data, which is usually extensive with respect to RAD loci, but branch length accuracy is improved in the reconstituted DNA approach compared to the conditional likelihood approach. More detailed studies are needed to address how structured missing data, model misspecification, and rate variation among loci impact phylogenetic analyses of RAD loci and SNP data.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.t9r3g.

Funding

This work was funded by grants from the National Science Foundation (DBI-1144630) and the Royalty Research Fund (UW, A61649) awarded to A.D.L. J.F. was supported by the National Science Foundation (DEB-1019583).

Acknowledgments

We thank the Secretariat of Environment and Natural Resources (SEMARNAT) for permission to conduct scientific collecting in Mexico (Permit No. 05034/11 to A.D.L., and Permit No. FAUT 0093 to A.N.M.O.) We thank the following institutions for tissue loans: Museum of Vertebrate Zoology (University of California, Berkeley), Burke Museum of Natural History and Culture (University of Washington), California Academy of Sciences, Ambrose Monell Cryo Collection (American Museum of Natural History), Los Angeles County Museum, Royal Ontario Museum, and the Museo de Zoologia "Alfonso L. Herrera" (Universidad Nacional Autónoma de México). We thank A. Gottscho for the *Uma* ddRADseq data. The University of Washington eSciences Institute provided computing infrastructure. Vladimir Minin and the Department of Statistics provided server space to host our phrynomics shiny application. Derrick Zwickl helped solve some problems with early versions of the RAxML implementation of the acquisition correction model. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303. We thank J. Brown, A. Chavez, M. Fujita, N. Goldman, J. Grummer, R. Harris, C. Linkem, A. Wright, and two anonymous reviewers for their comments. Axios Review provided valuable feedback that greatly improved the manuscript.

References

Arnold B., Corbett-Detig R., Hartl D., Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol. Ecol. 22:3179–3190.

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.

Bertels F., Silander O.K., Pachkov M., Rainey P.B., van Nimwegen E. 2014. Automated reconstruction of whole-genome phylogenies from short-sequence reads. Mol. Biol. Evol. 31:1077–1088.

Brandley M.C., Warren D.L., Leaché A.D., McGuire J.A. 2009. Homoplasy and clade support. Syst. Biol. 58:184–198.

Brumfield R.T., Beerli P., Nickerson D.A., Edwards S.V. 2003. The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol. Evol. 18:249–256.

Burbrink F.T., Pyron R.A. 2011. The impact of gene-tree/species-tree discordance on diversification-rate estimation. Evolution 65: 1851–1861.

Cariou M., Duret L., Charlat S. 2013. Is RAD-seq suitable for phylogenetic inference? an *in silico* assessment and optimization. Ecol. Evol. 3:846–852.

Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30:3317–3324.

Cruaud A., Gautier M., Galan M., Foucaud J., Sauné L., Genson G., Dubois E., Nidelet S., Deuve T., Rasplus J.-Y. 2014. Empirical assessment of RAD sequencing for interspecific phylogeny. Mol. Biol. Evol. 31:1272–1274.

Darriba D., Taboada G.L., Doallo R., Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9:772–772.

Davey J.W., Cezard T., Fuentes-Utrilla T., Eland C., Gharbi K., Blaxter M.L. 2013. Special features of RAD sequencing data: implications for genotyping. Mol. Ecol. 22:3151–3164.

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.

Eaton D.A., Ree R.H. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). Syst. Biol. 62:689–706.

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

Edgar R.C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

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. Nat. Acad. Sci. USA 107:16196–16200.

Felsenstein J. 1992. Phylogenies from restriction sites: a maximum-likelihood approach. Evolution 46:159–173.

Huang H., Knowles L.L. 2014. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. Syst. Biol. doi:10.1093/sysbio/syu046; first published online July 4, 2014.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. Mamm. Protein Metabolism 3:21–132.

Kuhner M.K., Beerli P., Yamato J., Felsenstein J. 2000. Usefulness of single nucleotide polymorphism data for estimating population parameters. Genetics 156:439–447.

Kumar S., Filipski A.J., Battistuzzi F.U., Pond S.L.K., Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457–472.

Leaché A.D., Chavez A.S., Jones L.N., Grummer J.A., Gottscho A.D., Linkem C.W. 2015. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biol. Evol. 7:706–719.

Leaché, A.D., Linkem C.W. 2015. Phylogenomics of horned lizards (genus: *Phrynosoma*) using targeted sequence capture cata. Copeia. doi: 10.1643/CH-15-248.

Leaché A.D., McGuire J.A. 2006. Phylogenetic relationships of horned lizards (*Phrynosoma*) based on nuclear and mitochondrial data: evidence for a misleading mitochondrial gene tree. Mol. Phylogenet. Evol. 39:628–644.

Lee T.-H., Guo H., Wang X., Kim C., Paterson A.H. 2014. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 15:162.

Lemmon A.R., Brown J.M., Stanger-Hall K., Lemmon E.M. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58:130–145.

Lewis P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.

Liu S., Lorenzen E.D., Fumagalli M., Li B., Harris K., Xiong Z., Zhou L., Korneliussen T.S., Somel M., Babbitt C., Wray G., Li J., He W., Wang Z., Fu W., Xiang X., Morgan C.C., Doherty A., O'Connell M.J., McInerney J.O., Born E.W., Dalen L., Dietz R., Orlando L., Sonne C., Zhang G., Nielsen R., Willerslev E., Wang J. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell 157:785–794.

Mastretta-Yanes A., Arrigo N., Alvarez N., Jorgensen T.H., Piñero D., Emerson B.C. 2014. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for populationgenetic inference. Mol. Ecol. Resour. 15:28–41.

McGill J.R., Walkup E.A., Kuhner M.K.. 2013. Correcting coalescent analyses for panel-based SNP ascertainment. Genetics 193:1185–1196.

Nieto-Montes de Oca A., Arenas-Moreno D., Beltrán-Sanchez E., Leaché A.D. 2014. A new species of horned lizard (genus *Phrynosoma*) from Guerrero, Mexico, with an updated multilocus phylogeny. Herpetologica 70:241–257.

Pante E., Abdelkrim J., Viricel A., Gey D., France S., Boisselier M.-C., Samadi S. 2014. Use of RAD sequencing for delimiting species. Heredity 114:450–459.

Pattengale N.D., Alipour M., Bininda-Emonds O.R., Moret B.M., Stamatakis A. 2010. How many bootstrap replicates are necessary? J. Comput. Biol. 17:337–354.

Peterson B.K., Weber J.N., Kay E.H., Fisher H.S., Hoekstra H.E. 2012. Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PLoS ONE 7:e37135.

Puritz J.B., Matx M.V., Toonen R.J., Weber J.N., Bolnick D.I., Bird C.E. 2014. Demystifying the RAD fad. Mol. Ecol. 23:5937–5942.

Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.

Robinson D., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

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

Rubin B.E., Ree R.H., Moreau C.S. 2012. Inferring phylogenies from RAD sequence data. PLoS ONE 7:e33394.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Salichos L., Stamatakis A., Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31:1261–1271.

Schliep K.P. 2011. Phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593.

Seeb J., Carvalho G., Hauser L., Naish K., Roberts S., Seeb L. 2011. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Mol. Ecol. Resour. 11:1–8.

Seo T.-K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol. 25:960–971.

Simmons M.P., Goloboff P.A. 2014. Dubious resolution and support from published sparse supermatrices: the importance of thorough tree searches. Mol. Phylogenet. Evol. 78:334–348.

Snir S., Rao S. 2012. Quartet MaxCut: a fast algorithm for amalgamating quartet trees. Mol. Phylogenet. Evol. 62:1–8.

Spinks P.Q., Thomson R.C., Shaffer H.B. 2014. The advantages of going large: genome-wide SNPs clarify the complex population history and systematics of the threatened western pond turtle. Mol. Ecol. 23:2228–2241.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Streicher J.W., Devitt T.J., Goldberg C.S., Malone J.H., Blackmon H., Fujita M.K. 2014. Diversification and asymmetrical gene flow across

time and space: lineage sorting and hybridization in polytypic barking frogs. Mol. Ecol. 23:3273–3291.

Sukumaran J., Holder M.T. 2010. Dendropy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Thomson M.J. 2014. Review article: High-throughput SNP genotyping to accelerate crop improvement. Plant Breed. Biotechnol. 2:195–212.

Wagner C.E., Keller I., Wittwer S., Selz O.M., Mwaiko S., Greuter L., Sivasundar A., Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22:787–798.

Wiens J.J., Kozak K.H., Silva N. 2013. Diversity and niche evolution along aridity gradients in North American lizards (Phrynosomatidae). Evolution 67:1715–1728.

Wiens J.J., Kuczynski C.A., Arif S., Reeder T.W. 2010. Phylogenetic relationships of phrynosomatid lizards based on nuclear and mitochondrial data, and a revised phylogeny for *Sceloporus*. Mol. Phylogenet. Evol. 54:150–161.

Yoder J.B., Briskine R., Mudge J., Farmer A., Paape T., Steele K., Weiblen G.D., Bharti A.K., Zhou P., May G.D., Young N.D., Tiffin P. 2013. Phylogenetic signal variation in the genomes of *Medicago* (Fabaceae). Syst. Biol. 62:424–438.