Received: 14 October 2016 | Revised: 5 March 2017 | Accepted: 3 April 2017

DOI: 10.1111/1755-0998.12681

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Squamate Conserved Loci (SqCL): A unified set of conserved loci for phylogenomics and population genetics of squamate reptiles

Sonal Singhal[1] | Maggie Grundler[1] | Guarino Colli[2] | Daniel L. Rabosky[1]

[1]Museum of Zoology and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA

[2]Departamento de Zoologia, Universidade de Brasília, Brasília, Brazil

Correspondence
Sonal Singhal, Museum of Zoology and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA.
Email: sonal.singhal1@gmail.com

Funding information
Conselho Nacional do Desenvolvimento Científico e Tecnológico – CNPq; Coordenação de Apoio à Formação de Pessoal de Nível Superior – CAPES; Fundação de Apoio à Pesquisa do Distrito Federal – FAPDF; Directorate for Biological Sciences, Grant/Award Number: DBI 1519732; David and Lucile Packard Foundation

Abstract

The identification of conserved loci across genomes, along with advances in target capture methods and high-throughput sequencing, has helped spur a phylogenomics revolution by enabling researchers to gather large numbers of homologous loci across clades of interest with minimal upfront investment in locus design. Target capture for vertebrate animals is currently dominated by two approaches—anchored hybrid enrichment (AHE) and ultraconserved elements (UCE)—and both approaches have proven useful for addressing questions in phylogenomics, phylogeography and population genomics. However, these two sets of loci have minimal overlap with each other; moreover, they do not include many traditional loci that that have been used for phylogenetics. Here, we combine across UCE, AHE and traditional phylogenetic gene locus sets to generate the Squamate Conserved Loci set, a single integrated probe set that can generate high-quality and highly complete data across all three loci types. We use these probes to generate data for 44 phylogenetically disparate taxa that collectively span approximately 33% of terrestrial vertebrate diversity. Our results generated an average of 4.29 Mb across 4709 loci per individual, of which an average of 2.99 Mb was sequenced to high enough coverage (≥10×) to use for population genetic analyses. We validate the utility of these loci for both phylogenomic and population genomic questions, provide a comparison among these locus sets of their relative usefulness and suggest areas for future improvement.

KEYWORDS
comparative population genomics, phylogenomics, squamate reptiles, target capture

1 | INTRODUCTION

For researchers working on biodiversity genomics, a primary challenge in project design is deciding which portion of the genome to sequence for the organisms of interest. Given that whole-genome sequencing remains prohibitively expensive for most organisms and most projects (but see Therkildsen & Palumbi, 2016), sequencing part of the genome allows researchers to affordably sample both more individuals and species. There are many approaches to subsetting the genome for sequencing, including transcriptome sequencing, restriction-aided digest methods (e.g., RAD sequencing) and target sequence capture, each of which poses benefits and challenges (Jones & Good, 2016). In the phylogenetics community, targeting and sequencing conserved elements—that is, anchored hybrid enrichment (AHE; Lemmon, Emme, & Lemmon, 2012) and ultraconserved elements (UCE; Faircloth et al., 2012b)—has been applied to infer phylogenies across broad phylogenetic scales (Crawford et al., 2012; Prum et al., 2015), resolve rapid radiations (Giarla & Esselstyn, 2015; Meiklejohn, Faircloth, Glenn, Kimball, & Braun, 2016) and characterize phylogeographic patterns (Brandley et al., 2015; Smith, Harvey, Faircloth, Glenn, & Brumfield, 2014). Because these loci are fairly conserved across broad phylogenetic scales (i.e., all of arthropods (Faircloth, Branstetter, White, & Brady, 2015) or all of angiosperms (Budenhagen et al., 2016)), researchers can use a common set of publicly available probe sequences for all their species of interest, thus saving energy, time and money.

The approaches targeting AHEs and UCEs are conceptually similar, although they are implemented differently. In both approaches, the basic premise is to identify regions of the genome that are conserved across deep phylogenetic scales and to design probes specific to these regions for use in target capture. AHE loci are long (>1 kb), the probes targeting these loci cover most of the locus sequence and the probe sequences are about 15% divergent among organisms diverged across 200 million years (i.e., snakes and geckos, Appendix S1: Fig. S1A; Zheng & Wiens, 2016). UCE loci tend to be shorter (500–800 bp), the probes only cover the highly conserved central 100–200 bp of these loci, and the probe sequences are very conserved (<5% across snakes and geckos; Appendix S1: Fig. S1A).

Research groups targeting conserved loci have focused on either AHEs or UCEs in generating data for their clades of interest, either of which offers more than enough data to resolve most phylogenetic questions. Unfortunately, AHEs and UCEs only have minor overlap in target loci. This creates a divide in the field. Historically, researchers targeted a common set of mitochondrial and nuclear loci across diverse species, enabling researchers to combine across data sets to create deeper, more fully sampled trees (c.f. Jetz, Thomas, Joy, Hartmann, & Mooers, 2012; Pyron, Burbrink, & Wiens, 2013). However, fully utilizing existing data sets is challenging if different research groups have targeted distinct and largely independent locus sets. In this study, we create a single inclusive locus set with applications to comparative population genomics, phylogeography and phylogenomics of squamate reptiles (lizards and snakes, ~200 million years of evolutionary history, Zheng & Wiens, 2016). This locus set—the squamate conserved locus set (SqCL)—combines across three major sets of loci: AHEs, UCEs and traditional genes used in squamate phylogenetics. We then test this locus set on a phylogenetically diverse set of 56 individuals representing 44 squamate species, confirming its efficacy and its usefulness for both population scale and phylogenetic studies. We further highlight areas of improvement in how these data are collected and analysed.

2 | METHODS

2.1 | Samples

To test the efficacy of the SqCL set, we targeted 16 of the most species-rich families in squamates that span the entire phylogenetic breadth of the clade, resulting in 56 individuals from 44 species. Importantly, this sampling consisted of multiple closely related congeneric species (Figure 1; Appendix S1: Table S1), allowing us to test how these markers resolved both shallower and deeper phylogenetic relationships. These individuals were all collected as part of ecological and macroecological studies in the Brazilian Cerrado over a 10 year span from September 2005 to October 2015 (Colli, Bastos, Araujo, Oliveira, & Marquis, 2002). Full details on the samples used can be found in Appendix S1: Table S1.

2.2 | Probe design

To design the probes for the SqCL set, we started with publicly available sequences for each locus set. For the AHEs, we used the sequence data for the AHE v2, as published in Ruane, Raxworthy, Lemmon, Moriarty Lemmon, and Burbrink (2015). This marker set consists of 394 loci as identified from multiple vertebrate genomes, of which five loci had no match in *Anolis carolinensis*. For the UCEs, we used the probe set Tetrapods-UCE-5Kv1 (accessed from www.ultraconserved.org on 10 February 2016). Because some UCE probes overlap, we assembled them using CAP3 (Huang & Madan, 1999), to result in 5,061 unique targets. For the standard genes used in squamate phylogenetics, we downloaded data matrices from two recent phylogenetic studies; these data sets included 44 genes from approximately 160 tips (Wiens et al., 2012) and 12 genes across 4,161 tips (Pyron et al., 2013). These two gene sets had four overlapping genes, resulting in 52 genes of which five were mitochondrial. Because mitochondrial DNA has much higher copy number than nuclear DNA, capturing both genomic types simultaneously can lead to an excess of sequence reads mapping to the mitochondrial genome (Bi et al., 2012). As such, we dropped these five mitochondrial genes, giving us a total of 47 nuclear loci that have traditionally been obtained using Sanger sequencing.

We then used BLAST (Camacho et al., 2009) to identify loci across the three sets that significantly overlapped with each other; we identified and dropped 28 duplicate loci. For the remaining 5,469 targets, we used blast to search for homologous regions of this genome across 11 publicly available squamate reptile genomes (Appendix S1: Table S2), extracted the matching regions and aligned across these regions using MAFFT v7.294 (Katoh & Standley, 2013). We used these alignments to characterize how divergent the targeted sequences were across genomes. We found that although the target sequences exhibited less divergence among snakes, they tended to show equal divergence among “lizards” and between any given “lizard” and any given snake. Given this, for every target, we included sequence representatives from two divergent clades within the phylogeny, to better capture some of this variation in target sequence identity across clades. For AHEs, we used both sequence from *Anolis carolinensis* and from either *Calamaria pavimentata* or *Python molurus*, as originally published in Ruane et al. (2015). For UCEs, we extended the central probes until we accumulated more than 15% sequence divergence across a rolling mean of 10 bp. Previous studies (Hugall, O’Hara, Hunjan, Nilsen, & Moussalli, 2015) have shown that, beyond 15% sequence divergence, capture efficiency begins to decline. We then extracted sequence data with these expanded coordinates from *A. carolinensis* and *Gekko japonicus*. For the few targets for which we

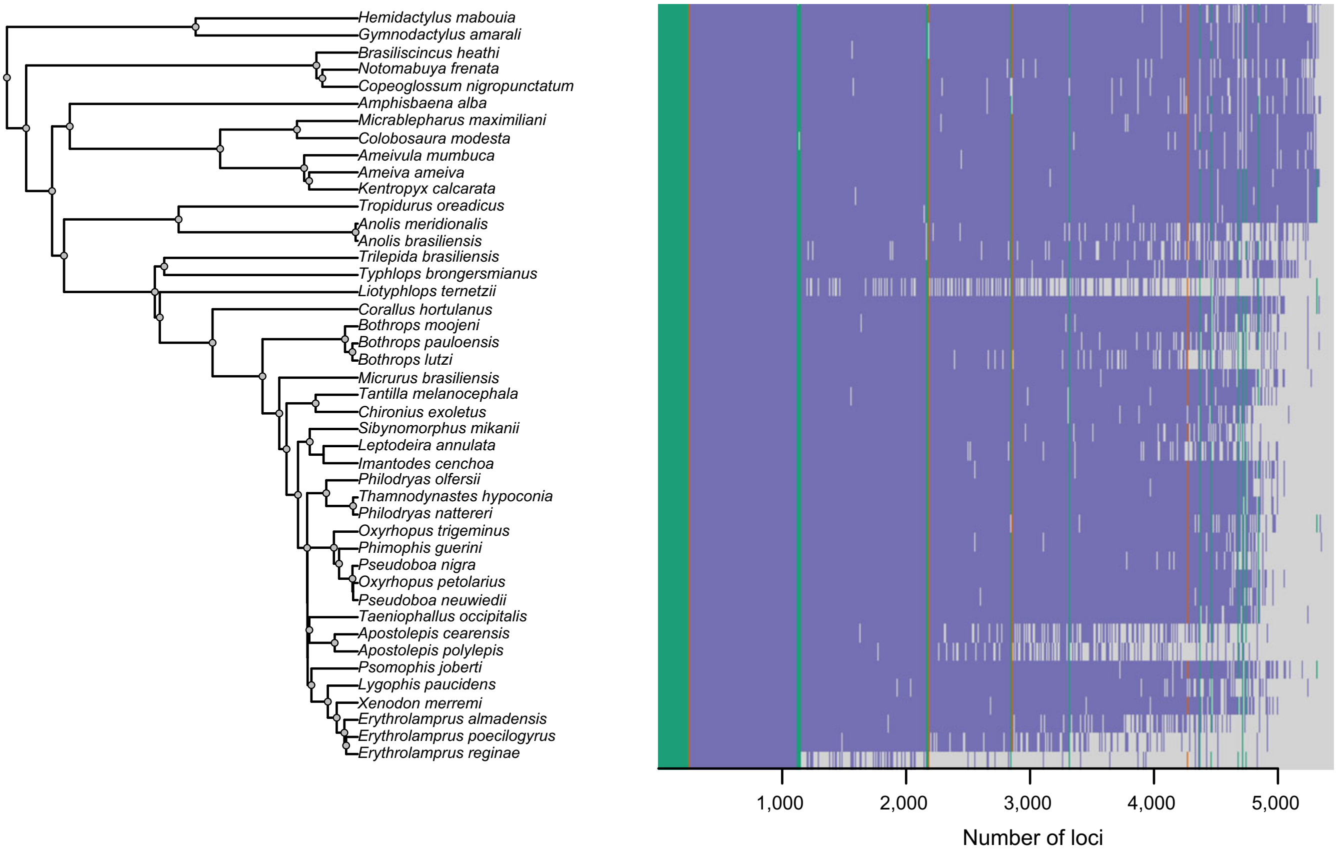


FIGURE 1 A phylogeny for the 44 species used to test the SqCL set along with a matrix, in which each column represents one of the 5,462 loci targeted. Green columns indicate anchored hybrid enrichment (AHE) loci, purple ultraconserved element (UCE) loci, orange traditional phylogenetic genes and grey indicates missing data. Loci are arrayed in order of most to least complete. The phylogeny topology was inferred using ASTRAL-II and BEAST2 for 2,815 loci that were 95% complete across all taxa and rooted with *Gallus gallus* (not shown). Grey dots mark nodes with >0.95 local posterior probability

could not identify a homolog from *G. japonicus*, we instead used *Ophisaurus gracilis*. For traditional phylogenetic genes, we used sequence data from *G. japonicus* and *Boa constrictor*. We then screened all targets against the RepeatMasker database, identifying seven targets that matched significantly to repeats. The final set consisted of 5,462 targets, each represented by two squamate sequences. Probes were designed across these targets at ~2× tiling density by MYcroarray Inc (Ann Arbor, MI, USA) resulting in 38,431 probes. These probes were then further filtered to remove probes that matched to repeats or to multiple places in the *Anolis carolinensis* genome. The final probe set consisted of 37,517 probes targeting 2.25 Mb of unique sequence. The total assembled sequence should be greater as UCEs are designed to capture flanking regions.

2.3 | Data collection

From each individual, we extracted high molecular weight DNA using a high-salt DNA extraction method (Aljanabi & Martinez, 1997) and then measured DNA quantity using a QuBit dsDNA BR Assay Kit (ThermoFisher, cat. no. Q32850) and DNA purity with a NanoDrop (ThermoScientific). MYcroarray then produced dual-barcoded libraries for each sample. Roughly 1.0 to 1.6 ng genomic DNA was sheared using a QSonica Q800RS sonicator and then size selected to approximately 450-bp modal lengths with SPRI beads. Sheared DNA was then end repaired and adapter ligated with the NEBNext Ultra DNA Library Prep Kit for Illumina (NEB, cat. no. E7370), and index amplified with custom primers for six cycles using HotStart HiFi Readymix (Kapa Biosystems, cat. no. KR0370). Following amplification, libraries were purified and quantified with Quant-iT PicoGreen dsDNA assay (ThermoFisher, cat. no. P7589). Roughly 100 ng across each of eight individuals was pooled. Because capture efficiency typically has phylogenetic signal (Bi et al., 2012; Cosart et al., 2011), we reduced bias by pooling individuals by taxonomic family. These pools of 800 ng were then dried to 7 μl via vacuum centrifugation and used as template for standard capture reactions following the MYBAITS PROTOCOL v3. We modified the protocol slightly to include xGEN Universal Blockers (Integrated DNA Technologies, cat. no. 1046636 and 1046639), which have been shown to improve target capture efficiency by up to 4× (Portik, Smith, & Bi, 2016). Following a 12-cycle postcapture PCR, all 56 individuals were combined with an additional eight frog samples from another study (J.G. Larson, unpublished) and sequenced by Hudson Alpha on one 100 paired-end run of a HiSeq 2500 v4.

2.4 | Data analysis

Our pipeline for SqCL facilitates both population genetic and phylogenomic analysis, reflecting the potential use of these loci for questions at both shallow and deep scales of divergence. This pipeline is influenced by the publicly available PHYLUCE pipeline (Faircloth, 2015) but includes two primary modifications. First, we implement species tree methods for phylogenetic inference, because these methods generally outperform concatenated-based approaches (Kubatko & Degnan, 2007; Warnow, 2011) but see (Springer & Gatesy, 2016). Second, we incorporate industry-standard SNP calling, filtering and phasing to enable population genetic analyses. This pipeline, along with documentation explaining its implementation, is available at https://github.com/singhal/SqCL.

Following demultiplexing, we removed adapters and low-quality regions from the reads using TRIMMOMATIC v0.36 (Bolger, Lohse, & Usadel, 2014) and then merged overlapping reads using PEAR v0.9.10 (Zhang, Kobert, Flouri, & Stamatakis, 2014). We then assembled the reads using default settings on the program TRINITY v2.2.0 (Grabherr et al., 2011); for the few samples requiring memory in excess of 64 Gb, we used in silico read normalization to thin the original data set. We matched contigs in each individual assembly to the original targets using BLAT v36 (Kent, 2002). We identified two types of matches: in the first, the contig and the target have a one-to-one unique match; in the second, the target matches to multiple contigs with match scores within 10 orders of magnitude of the best match. We then used these match designations to create a pseudo-reference genome (PRG) for each species. Here, we identified all contigs across all individuals in a given species that match to a given target and then retained either the longest contig or the best matching contig if it was a significantly better match than the next best matching contig (>3 orders of magnitude). We then implemented phylogenomic and population genomic analyses as detailed below.

2.5 | Assessing informativeness for phylogenomics

To facilitate phylogenomic analyses, we first extracted homologues for our target loci from *Gallus gallus* (Hillier et al., 2004) for use as an outgroup. We then used MAFFT to generate alignments for each locus sampled for ≥4 species (Katoh & Standley, 2013) and trimmed alignments to remove regions of low quality using GBLOCKS (Castresana, 2000). We inferred gene trees for each alignment using RAXML v8.2.4 (Stamatakis, 2006).

For each locus, we measured (i) phylogenetic informativeness, (ii) certainty of gene trees inferred with that locus, and (iii) how clock-like a locus is. Empirical results show that maximizing these metrics can improve the accuracy of topology and branch length inference. First, empirical results suggest that the ideal loci are phylogenetically informative across evolutionary time—that is, they should contain variable sites at recent timescales while not exhibiting homoplasy at deeper time scales (Dornburg, Townsend, Friedman, & Near, 2014; Gilbert et al., 2015). To characterize phylogenetic informativeness (PI) for each locus, we used the method introduced by Townsend (2007) and implemented in TAPIR v1.1 (Faircloth, Chang, & Alfaro, 2012a; Guindon et al., 2010). Measuring phylogenetic informativeness requires an ultrametric tree. We used the tree inferred from our combined ASTRAL + BEAST analysis (see below); we rescaled the tree using the R package "GEIGER" to have a root age that reflects estimates from the literature (Harmon, Weir, Brock, Glor, & Challenger, 2008; Zheng & Wiens, 2016). Second, empirical results suggest gene trees with high tree certainty lead to more accurate phylogenies (Blom, Bragg, Potter, & Moritz, 2016; Salichos & Rokas, 2013). We used a tree certainty measure that calculates the relative frequency of each bipartition in a set of trees with respect to the frequency for the most common conflicting bipartition (Salichos & Rokas, 2013). Higher scores reflect a topology that shows greater stability across replicates. Here, we inferred 100 bootstraps with RAXML to use as replicates (Salichos, Stamatakis, & Rokas, 2014). Finally, empirical results suggest trees inferred with more clock-like genes are more accurate (Doyle, Young, Naylor, & Brown, 2015). We measured clocklikeness following the approach outlined in Doyle et al., 2015; in which we compared tree likelihoods estimated by PAUP v4 for a gene tree forced to be ultrametric to one that was not (Swofford, 2003).

We then employed three phylogenetic approaches. First, we generated concatenated alignments by marker type, defined partitions using PARTITIONFINDER2 with the "RCLUSTER" algorithm (Lanfear, Frandsen, Wright, Senfeld, & Calcott, 2017), and then inferred phylogenies with RAXML v8.2.4 (Stamatakis, 2006). Second, we implemented a species tree approach. RAXML generates fully bifurcating gene trees even if some nodes have no support. Using the di2multi function in the R package "ape" (Paradis, Claude, & Strimmer, 2004), we first collapsed all such nodes in the gene trees; these nodes have branch lengths <1e−5. We then used these gene trees to infer species tree using ASTRAL v4.10.7 (Mirarab & Warnow, 2015). ASTRAL only infers tree topology, so to infer branch lengths, we used BEAST v2.4.5 (Bouckaert et al., 2014). To ensure reasonable run times, we randomly subsampled the data sets to 100 loci each and ran five independent samples. We did not set fossil or mutation rate priors as we were interested primarily in comparing relative branch lengths. Because we were only interested in inferring branch lengths, we fixed the topology to the ASTRAL tree by turning off the subtree slide, Wilson-Balding and narrow and wide exchange operators. We used an uncorrelated relaxed clock across branches and ran each locus set for 100e6 steps with a 20% burn-in. Trees were visualized and compared using the R packages "GGTREE" and "TREESCAPE" (Jombart, Kendall, Almagro-Garcia, & Colijn, 2015; Yu, Smith, Zhu, Guan, & Lam, 2017). Third, because the SqCL set does not target any mitochondrial loci, we used the program MITOBIM v1.8 (Hahn, Bachmann, & Chevreux, 2013) to reconstruct partial to whole mitochondrial genomes from by-catch reads. For each individual, we identified their closest phylogenetic relative from the 271 squamates that have publically available mitochondrial genomes and used this genome as the seed genome. We then generated a concatenated alignment of the mitochondrial gene sequences and used RAXML to infer the mitochondrial gene tree.

2.6 | Assessing informativeness for population genomics

To facilitate population genetic analyses, we aligned trimmed reads from each individual to its PRG using BWA v0.7.12 (Li, 2013), fixed mate-pair information using SAMTOOLS v1.3.1 (Li et al., 2009), marked duplicate read pairs using PICARD v2.4.1 (accessed from https://broadinstitute.github.io/picard/) and identified and realigned indels using GATK v3.6 (McKenna et al., 2010). We then called a raw set of variants across all individuals in a species using GATK in UNIFIEDGENOTYPER mode, filtered the variants to retain only high-quality variants occurring at sites ≥10× and used this filtered variant set to perform base quality score recalibration of the read alignment files. We used GATK's UNIFIEDGENOTYPER to call both nonvariant and variants from these recalibrated alignment files and filtered the variants to remove low-quality sites and to set genotypes to missing where coverage was <10×. Finally, we used GATK's ReadBackedPhasing to phase variants. The resulting variants were used to infer nucleotide diversity (π; (Tajima, 1983)) and F_{ST} (Reich, Thangaraj, Patterson, Price, & Singh, 2009).

3 | RESULTS AND DISCUSSION

3.1 | Data quality

The data collected were of high quality and confirmed the efficacy of the SqCL probe set. Of the 5,462 targets, only 150 targets failed (seven AHE loci, 140 UCE loci and two genes); we define failed loci as those that were recovered at <10× coverage for all individuals (Figure 1). In total, we were able to generate an average of 4.29 Mb across 4,709 loci of sequence data per individual, of which an average of 69.8% was sequenced to high coverage (>10×). We were able to assemble most targets in most individuals, leading to a fairly complete data set particularly for AHE loci (Appendix S1: Fig. S2). Because missing data can often complicate phylogenomic inference (Hosner, Faircloth, Glenn, Braun, & Kimball, 2016; Wiens, 2003), researchers using this locus set should be able to restrict analyses to just well-sampled loci and still have sufficient data to power most phylogenetic analyses.

The 5,312 captured targets are distributed across the nuclear genome and across all chromosomes in *Anolis carolinensis* (Appendix S1: Fig. S3). This dispersed genomic distribution makes it likely that these loci are independently evolving, as assumed in many population genomic and phylogenomic analyses (Brito & Edwards, 2009). For one individual (here, we chose *A. brasiliensis* because it is closely related to the squamate species for which we have the best annotated genome, *A. carolinensis*), we determined the percentage of coding loci in the capture data set. Of the 5.90 Mb of assembled sequence for *A. brasiliensis*, 5.31 Mb could be aligned to the *A. carolinensis* genome, of which 825 Kb (15.5%) spanned exons and 2.32 Mb (43.8%) fell within gene coordinates. Because only a fraction of the assembled sequence is coding, these loci are not appropriate for researchers interested in some molecular evolutionary questions (i.e., looking at substitution rates for nonsynonymous vs. synonymous sites) although other questions (i.e., levels of heterozygosity in natural populations) can still be addressed. In all subsequent analyses, we analyse both population genetic and phylogenetic inference across all sequence.

We calculated several other quality metrics, including capture efficiency (or, the proportion of sequenced reads that map onto targeted loci), the number of total loci recovered, mean locus length, mean coverage across loci and percentage of duplicate reads (Figure 2, Appendix S1: Figs. S4, S5, Table S1). In general, we see good results for all metrics across all of the diversity sampled. Most notably, on average 93% of AHE loci were captured at an average length of 1,556 bp, 82% of genes at 1,040 bp and 86% of UCEs at 841 bp. Our experiment had a relatively high average capture efficiency rate (60.0%)—capture efficiencies reported in the literature for AHEs and UCEs can range from 10% to 80% (Faircloth et al., 2012b; McCormack, Tsai, & Faircloth, 2016; Ruane et al., 2015). On average, our locus assemblies were 30% and 70% longer than the total target length for AHEs and genes, respectively, and these assemblies were of high quality—80% of our paired reads mapped properly. Snakes generally performed less well than other squamates, particularly for UCE loci (Figure 2). This reduced data quality is partially because one of our eight pools performed poorly during the target capture step of the laboratory experiment. A linear model found that the pool identity best explained variation in the number of loci assembled across individuals (adjusted $r^2 = 0.47$; Appendix S1: Fig. S6). Pool number and taxonomy are conflated because we pooled individuals by families. However, both pools 1 and 2 consisted solely of species from the family Dipsadidae, and yet, they had markedly different success rates.

Probe design also explains some of this variation in capture efficiency across individuals. The probe design included a gecko and an anole for UCEs and a snake and an anole for AHEs, which we believe led to geckos' hybridization with UCEs outcompeting their hybridization with AHEs and vice versa for the snakes. The data confirm this hypothesis. We see geckos have lower AHE recovery compared to squamates as a whole but see no performance reduction for UCEs, and snakes have lower UCE recovery compared to squamates as a whole but have no performance reduction for AHEs (Figure 2). As such, we suggest future users use a modified version of this initial probe set (SqCL v2; available at github.com/singhal/SqCL), in which we include, for 96% of loci, a representative sequence from the lizard *Anolis carolinensis* and the snake *Python molurus*. The remaining 4% have poor matches to either the *A. carolinensis* or the *P. molurus* genomes, so we instead use sequence from *Gallus gallus*, one of seven snake species, or the lizard *Ophisaurus gracilis*.

Perhaps the biggest area for improvement is to increase library complexity. Library complexity measures how many of the reads in a library share identical start sites; lower complexity libraries lead to more sequenced reads being exact duplicates of existing reads. Anywhere from 22% to 54% of our reads were marked as duplicates via computational methods, and duplication rates were correlated to library complexity ($r = -0.349$, $p = .009$). Our libraries should be low

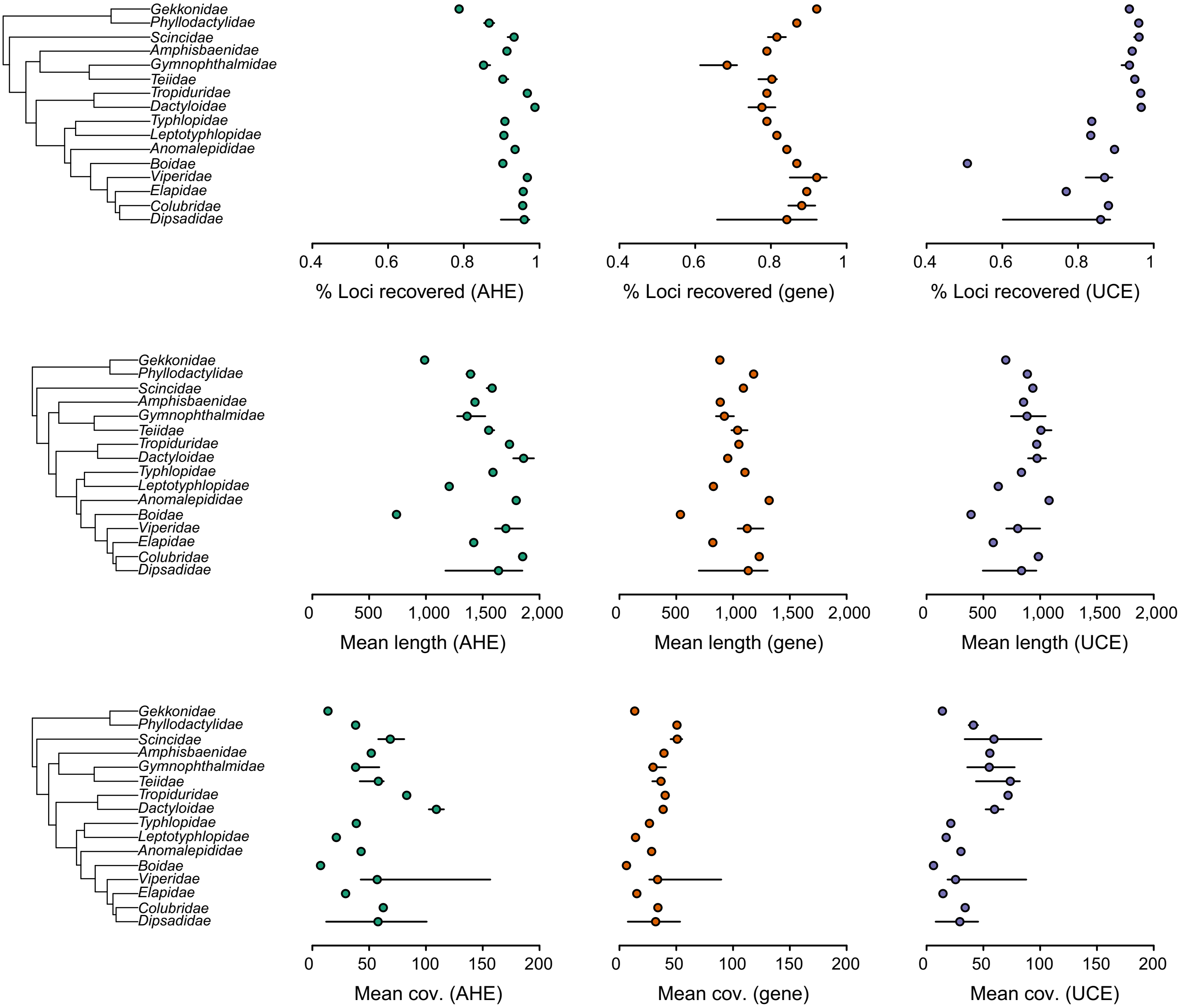


FIGURE 2 Several metrics of data quality summarized across the three types of loci in the SqCL set (AHE, anchored hybrid enrichment loci; traditional phylogenetic genes: gene; UCE, ultraconserved elements) across the 16 squamate families sampled. Results show the SqCL set works well across taxa that last shared a common ancestor more than 200 million years ago. Data quality metrics are the percent of loci targeted that were recovered, the mean locus length and mean coverage across the locus. Shown are median values and the 95% percentile range across individuals sampled for that family. Not all points are shown with confidence intervals because we only sampled one species in some families. A version of this figure showing patterns across additional metrics is shown in Appendix S1: Fig. S3

complexity because we targeted a subset of the genome, but we see variance around that expectation—libraries in this experiment achieve saturation at different sequencing depths (Appendix S1: Fig. S7). Improving library complexity, both using higher quality DNA, increasing conversion rates during library generation and increasing capture efficiency—allowing us to reduce the number of PCR cycles used to amplify libraries—should make these experiments more efficacious, ensuring that more reads sequenced are unique and can be retained for downstream analyses.

Standard quality metrics for target capture experiments have not yet been reported for either AHE or UCE loci, such as the correlation in coverage across loci across individuals and the variance in coverage across loci within an individual. Another standard measure, sensitivity or the percentage of bases of the original target that are at least covered by one read, is less relevant to report here given that UCEs are designed to capture loci much longer than the original probe sequence. These metrics are particularly useful when target capture loci are used for population genomics, because variant calling quality is sensitive to sequencing depth (Nielsen, Paul, Albrechtsen, & Song, 2011). If coverage is uneven across loci and across probes, it can lead to sparse data matrices. Thus, in an ideal capture set, variance in coverage across loci within an individual would be minimal. Although we expect some probes will work better because of their GC-content, melting temperature and divergence across loci, minimizing variance helps ensure a more complete data matrix across loci and individuals. Concordantly, in an ideal target set, average coverage across loci should be correlated across individuals, reflecting the differential efficacy of targets. Low correlations suggest that experimental error is increasing variance across samples. We report an average coefficient of variation of 1.20 across loci within individuals, with lower values for AHEs (0.86) and genes (0.98) than UCEs (1.23) (Appendix S1: Fig. S8). Coverage across loci and across individuals was correlated at an average $r = 0.373$ (Appendix S1: Fig. S9), and we recovered higher correlations for AHEs ($r = 0.48$) and genes ($r = 0.65$) than UCEs ($r = 0.36$).

We see both higher coefficients of variation across loci coverage and lower correlation among individuals than has been reported in exome capture experiments (Bi et al., 2012; Bragg, Potter, Bi, & Moritz, 2015; Hugall et al., 2015; Portik et al., 2016). Most exome capture experiments are conducted at a much narrower taxonomic scale (i.e., across species diverged tens of millions of years) than the taxonomic scale used here (i.e., hundreds of millions of years). This increased variance could simply reflect the increased divergence between the probes and the target genomic sequences. To test this hypothesis, we fit a linear model for which factors best predict how well a given locus worked across all individuals, including factors such as the average divergence of the probe sequence across the species considered, the number of probes used for that species, the GC and repeat content of the probes and the loci themselves, and the type of locus (i.e., AHE, UCE or gene). Our best-fitting single-variable model showed that more divergent probes lead to lower rates of locus recovery (Appendix S1: Fig. S10). To ameliorate these effects, future work could reconstruct the ancestral sequence for a given probe across the species of interest and include this sequence in probe sets. A similar approach allowed researchers to target exome data successfully across 250 million years of evolution in the invertebrate class Ophiuroidea (Hugall et al., 2015). Making these improvements would indubitably help, but our linear model explains a relatively small portion of the variance ($r^2 = 0.09$, $p < .001$; Appendix S1: Fig. S10). Future work should consider how we can reduce variance in assembly success and coverage across loci to improve the completeness of our data sets.

3.2 | Data informativeness for phylogenetics

Because previous work has clearly shown the utility of AHE and UCE markers for phylogenetic inference (Faircloth et al., 2012b; Lemmon et al., 2012), we focus our discussion on how marker type influences phylogenetic inference. First, although the probes targeting UCE loci are much more conserved than the probes targeting AHE loci, the sequence divergence of the loci themselves is comparable across locus types (Appendix S1: Fig. S1). Further, these loci show broad distributions in how variable they are across sampled individuals (Appendix S1: Fig. S1). Because the evolutionary rates of these loci vary, these loci should be able to resolve both broad and shallow radiations. In fact, as others have found, both locus types contain many variable sites across both broad and more shallow radiations (Figure 3)—the average AHE, gene and UCE locus contains 0.44, 0.39 and 0.43 variable sites/bp across the broad array of squamates sampled. However, where these variable sites occur across loci varies by locus type. AHEs exhibit a fairly uniform density of variable sites across the length of the locus and, as reported previously (Faircloth et al., 2012b), UCEs show a U-shaped pattern, with the density of variable sites increasing away from the locus centre. Further, UCEs show a decline in variable site density at loci ends. Because assembled locus length varies across individuals, the per

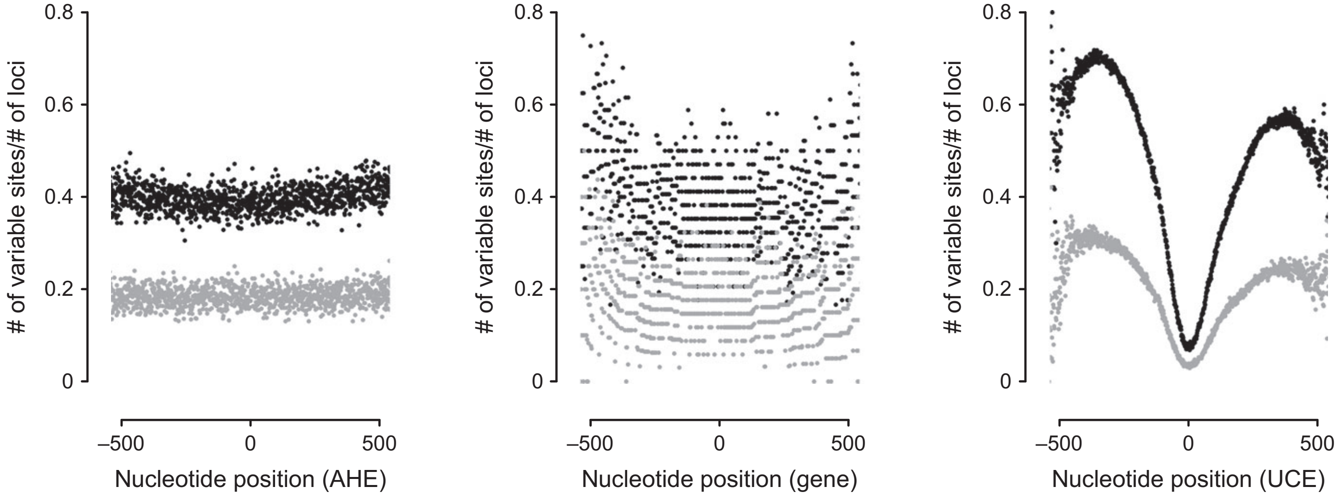
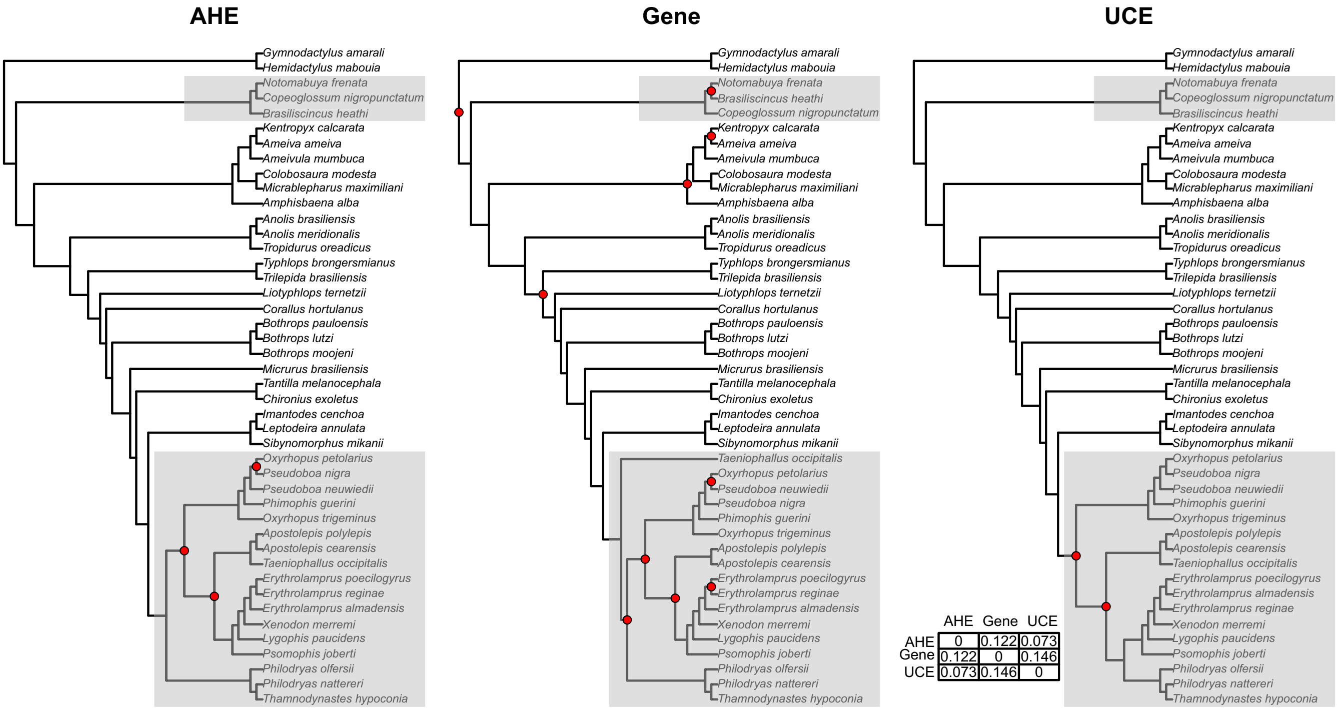


FIGURE 3 Density of variable sites in multispecies alignments for the three types of loci in the SqCL set (anchored hybrid enrichment: AHE; traditional phylogenetic genes: gene, ultraconserved element: UCE). Black dots indicate variable site density for the 44 squamate taxa sequenced in this study; grey dots variable site density for the 30 snake taxa sequenced. The squamates span 200 million years of divergence, and the snakes span 120 million years of divergence (Zheng & Wiens, 2016). The frequency of variable sites differs between the two comparisons, reflecting the difference in phylogenetic depth. Different loci types exhibit different variable density patterns across the length of the loci, which is both a function of locus design and variation in levels of missingness across the locus alignment

cent of missing characters at any given column of an alignment increases towards the alignment ends (Appendix S1: Fig. S11). This pattern underscores the importance of trimming alignments to remove regions with high density of missing data (Lemmon & Lemmon, 2013).

We then explored how these alignments—and their resulting gene trees—differ across several metrics that empirical data suggest can influence phylogenetic inference. First, we inferred phylogenetic informativeness (PI) (Townsend, 2007). PI profiles for the three marker types are comparable (Appendix S1: Fig. S12), and all markers are able to resolve deep relationships. None of the marker types shows appreciable declines after they reach their maximum informativeness, unlike what is typically seen in more quickly evolving loci, like mitochondrial genes (Dornburg et al., 2014). As such, all three marker types should be useful for phylogenetic inference. Second, we measured tree certainty as measured by Salichos and Rokas (2013). Our results showed that AHE and gene markers have greater tree certainty than UCE markers (Appendix S1: Fig. S13). This difference in part reflects a trade-off between locus length and tree certainty. Longer loci (like AHE loci) tend to be result in better resolved gene trees (Arcila et al., 2017; Blom et al., 2016), although they are also more likely to contain recombination events that violate most gene tree inference methods. Finally, we characterized how well these loci fit to a clock-like model for molecular evolution, finding UCE markers appear to be more clock-like than AHEs (Appendix S1: Fig. S14). No one marker type emerges as superior to the others across these metrics. Rather, these loci exhibit significant variation across these metrics, suggesting that sampling more loci will allow users to carefully filter loci as required by their desired analysis.

We then inferred phylogenies for these loci using concatenated and species tree approaches. We do not discuss our concatenated results (Appendix S1: Fig. S15) because concatenation (particularly with phylogenomic data) can often converge on the wrong tree with high support (Kubatko & Degnan, 2007). Our species tree analyses recover largely similar topologies across the three marker types (Figure 4), particularly between the topologies inferred with AHE and UCE markers. Across all comparisons, the nodes that disagree also tend to have low support. Additionally, our divergence dating analyses across marker types showed that branch length estimates were highly correlated across inferred trees (Appendix S1: Fig. S16), a result that is unsurprising given that raw pairwise genetic divergences between tips are also highly correlated (Appendix S1: Fig. S17). In sum, phylogenetic inference—in both topology and branch length estimation—is robust to marker type. Further, while the increased data content of both AHE and UCE marker sets allow us to resolve some tricky nodes in the phylogeny, some nodes remain poorly resolved. Future work will explore (i) filtering loci to see whether filtered data sets lead to more resolved tress (Blom et al., 2016; Doyle et al., 2015; Salichos & Rokas, 2013), (ii) using methods that resolve tricky nodes by constraining the topology space explored (Arcila et al., 2017) or (iii) accounting for phylogenetic uncertainty in any tree-based analyses. Further explorations



	AHE	Gene	UCE
AHE	0	0.122	0.073
Gene	0.122	0	0.146
UCE	0.073	0.146	0

FIGURE 4 Species trees inferred using anchored hybrid enrichment (AHE), traditional phylogenetic (gene) loci and ultraconserved element (UCE) loci that were 95% complete across the 44 species with ASTRAL-II. Trees were rooted with *Gallus gallus* (not shown). Grey boxes mark clades that exhibit unstable topologies across marker sets, and the matrix shows normalized Robinson–Foulds distances between trees. Nodes with <0.95 local posterior probability are shown in red. Topologies are largely concordant across marker sets, and conflicting nodes generally have low support

into the causes for this topological discordance are beyond the scope of this study.

3.3 | Recovering the mtDNA genome

Including mitochondrial targets in the probe set is not recommended. Because of the difference in copy number between mitochondrial and nuclear genomes in vertebrates, mitochondrial DNA generally outcompetes nuclear DNA for binding, leading to far greater coverage of the mtDNA genome than the nDNA genome (Bi et al., 2012). However, mtDNA is the traditional workhorse for phylogenetics, and genealogical discordance between mtDNA and nuclear data is often used to test for introgression between taxa (Toews & Brelsford, 2012). As such, we evaluated our ability to recover mtDNA from these taxa. We were able to assemble portions of the mtDNA genome for 55 of our 56 individuals, although two of these individuals had no sequence data for any of the 13 mtDNA polypeptide genes. Of the remaining 53 individuals, we recovered 89.2% of the total length of the mitochondrial genome. We used these data to infer a mtDNA gene tree (Appendix S1: Fig. S21), which differs from the SqCL-based tree at deeper nodes although it recovers many of the same species relationships within families. The quality of our mtDNA assembly (here, measured by the portion of sequence that was recovered) was negatively correlated with capture efficiency ($r = 0.453$, $p < .001$). In individuals with more reads mapping on target, there are fewer reads randomly sequenced from the mtDNA genome, and thus, recovering a complete mtDNA genome is less likely.

3.4 | Data informativeness for population genomics

The utility of AHE and UCE loci for population genomics and phylogeography has already been reported in a number of papers (Brandley et al., 2015; Harvey, Smith, Glenn, Faircloth, & Brumfield, 2016; Zarza et al., 2016). Here, we further compare and contrast across patterns of variation across the locus types. Although we expect AHEs, UCEs and conserved genes to be less variable than other loci type used in population genomics—that is, exons or RAD data (Bragg et al., 2015; Harvey et al., 2016)—we recover sufficient variation across all three locus types to power population genomic analyses (Figure 5). In particular, summarizing across all data types, we were able to generate robust estimates of isolation-by-distance slopes for both *C. modesta* and *B. moojeni* (Figure 6), illustrating the utility of these markers to study

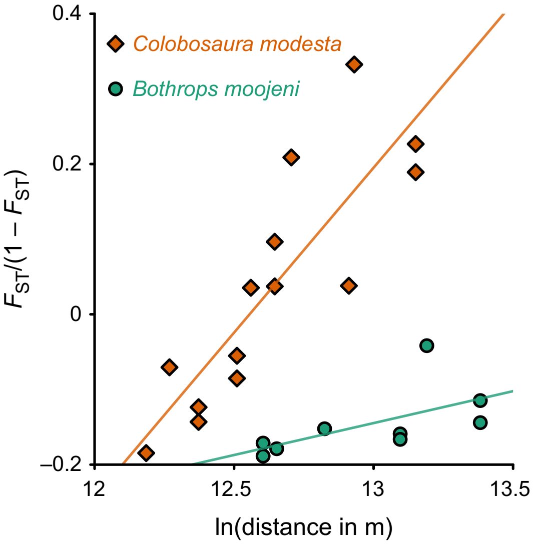


FIGURE 6 Isolation-by-distance estimates for *Colobosaura modesta* and *Bothrops moojeni*. Each point represents a pairwise comparison between two individuals. F_{ST} estimates are based on an average of 37K variant sites. The two species have very different isolation-by-distance relationships, illustrating the power of SqCL markers to address questions about population-level variation

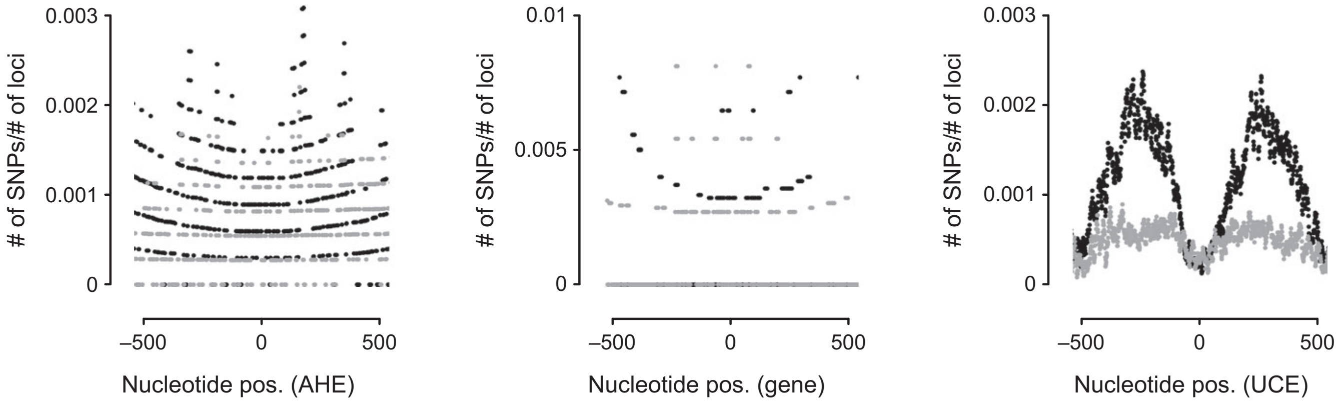


FIGURE 5 Density of single nucleotide polymorphisms (SNPs) for the three types of loci in the SqCL set (anchored hybrid enrichment: AHE; traditional phylogenetic genes: gene, ultraconserved element: UCE). Black dots indicate single nucleotide polymorphism (SNP) density across six individuals of the lizard *Colobosaura modesta*; grey dots SNP density across five individuals of the snake *Bothrops moojeni*. Different loci types exhibit different SNP densities across the length of the loci, which is both a function of locus design and average sequencing coverage across the loci length

population-level processes. The average AHE, gene and UCE locus contains 0.0035 (10%–90% distribution: 0.001–0.006), 0.0025 (0.0–0.004) and 0.0042 (0.001–0.008) segregating sites per bp across *Colobosaura modesta*, the lizard for which we sampled six individuals, and 0.0019 (0.0–0.004), 0.0022 (0.0–0.007) and 0.0021 (0.0–0.004) segregating sites per bp across *Bothrops moojeni*, the snake for which we sampled five individuals. The pattern of SNP density mimics the pattern of variable site density (Figures 3, 5). As seen with our phylogenetic results, locus design influences both coverage and patterns of variation across the length of loci (Appendix S1: Fig. S18). Despite this, patterns of both genetic diversity and differentiation were highly correlated across marker types (Appendix S1: Figs. S19, S20). The slope of these relationships generally deviated from unity, which reflects these loci's different evolutionary histories. Selection, recombination and their interaction likely influence effective population sizes across these markers differentially (Charlesworth, 2009).

3.5 | Practicality of approach

We pooled fewer individuals to a lane than most other target capture experiments, which regularly multiplex 100 individuals to a single lane of sequencing. Thus, we sequenced our libraries to a much greater depth than is typical. To test how reduced sequencing would affect the quality of the data recovered, we conducted a series of subsampling experiments in which we took the 11 individuals in *Colobosaura modesta* and *Bothrops moojeni* and randomly sampled 5e5, 1e6, 1.5e6 and 2e6 paired reads (for a total of 1e6, 2e6, 3e6 and 4e6 reads). With current sequencing yields on the Illumina HiSeq 2500 v4 sequencing platform of approximately 250 million paired reads, this represents pooling of approximately 500, 250, 166 and 125 individuals to a single sequencing lane. Even with significantly reduced sequencing, we still assembled a large number of loci for a given individual, with only modest improvements for additional sequencing beyond 2e6 reads (Figure 7). However, sequencing more reads led to a linear increase in the number of sites with sufficiently high coverage to call variants (Figure 7). Researchers interested in population genomic analyses might want to use lower levels of multiplexing than those interested solely in phylogenomics. This analysis is contingent on both capture efficiency and library complexity, and improving the number of reads mapping on target and/or reducing the library duplication rate will allow researchers to multiplex even further.

Using the SqCL probe set presents additional costs. More probes must be synthesized than if either locus set was used in isolation. In our study, the cost for probes per sample increased from $25 for solely capturing UCEs to $31.25 for the entire SqCL set. Further, sequencing both loci requires further investment in sequencing than sequencing either set alone. However, our subsetting experiment (Figure 7) suggests that researchers should still be able to multiplex at similar levels as used in other projects using AHE and UCE loci (Meiklejohn et al., 2016; Prum et al., 2015), despite the increase in overall target length. Thus, we anticipate that using SqCL loci will result in only modest increases in cost for a given project, while generating a much more inclusive dataset.

4 | CONCLUSIONS

The AHE and UCE locus sets made an important contribution to the field of biodiversity genomics by allowing researchers to efficiently query homologous loci across a diversity of organisms. However, the presence of two largely nonoverlapping locus sets has created an unfortunate divide, in that many research groups have invested in either AHEs or UCEs for their clade of interest. This lack of overlap will hinder future attempts at synthesis in both population genomics and phylogenetics, limiting the utility of existing datasets. We have provided a simple resolution to this problem by presenting a probe set that includes AHEs, UCEs and ~50 additional loci that have served as "workhorse genes" for squamate phylogenetics. Because target capture also often allows us to recover the mitochondrial genome (Appendix S1: Fig. S21), the SqCL probe set thus provides maximal integration with most existing phylogenetic data. These data

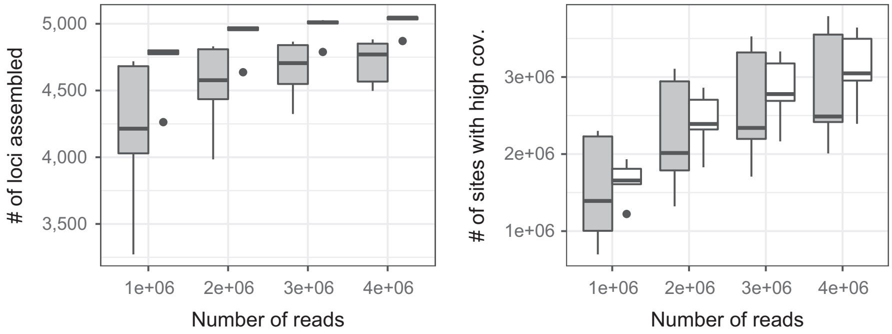


FIGURE 7 Results of an in silico experiment testing the effect of reducing sequencing depth on the number of loci assembled and the number of sites with ≥10× coverage, or those sites at we call single nucleotide polymorphisms (SNPs). For the five individuals in the species *Bothrops moojeni* (shown in gray) and the six individuals in *Colobosaura modesta* (in white), we used SEQTK (https://github.com/lh3/seqtk) to randomly subsample 5e5, 1e6, 1.5e6 and 2e6 paired reads (for a total of 1e6, 2e6, 3e6 and 4e6 reads) and analysed the data using our bioinformatics pipeline. In this study, we sequenced an average of 3.5e6 paired reads for these 11 individuals. We could reduce sequencing depth by 70% and still recover 86.7% of the loci. Decreasing sequencing depth, however, does decrease the number of sites recovered at high coverage

also allow population and conservation genetics researchers to generate data sets that can ultimately be integrated into broad-scale comparative analyses. We advocate the use of this integrated probe set for questions currently being addressed with UCEs or AHEs, thus ensuring that future data sets for squamate reptiles are compatible with much of the existing phylogenetic data generated over the last thirty years. Importantly, both population genomic and phylogenetic inferences are robust across marker types.

Although we refined the AHE, UCE and traditional gene sets for their application to squamate phylogenetics only, our approach can easily be applied to other tetrapod systems and could be used to create a probe set of general use across the phylogeny, thus further supporting the development of community-wide, inclusive locus set for use in phylogenomics and comparative population genomics. This study took effort to customize these probe sets for squamates; however, published probe sequences could simply be synthesized and applied to tetrapod systems of interest (Faircloth et al., 2012b; Lemmon et al., 2012). AHE probes tend to diverge more quickly across phylogenetic distance than UCE probes (Appendix S1: Fig. S1A). Thus, to ensure efficient capture, researchers should ideally synthesize AHE probes specific to their broad clade of interest (i.e., amphibians, reptiles or mammals).

ACKNOWLEDGEMENTS

For logistical support, we thank Alison Devault & Jake Enk at MYcroarray, Marcella Baiz, Robbin Murrell, Gabriel Costa, Izabella Paim da Silva and Fabricius Maia Domingos. Brant Faircloth and Mozes Blom provided useful advice. This work is funded by a grant from the David and Lucile Packard Foundation to DLR and a NSF Postdoctoral Fellowship in Research Biology to SS. GRC thanks Coordenação de Apoio à Formação de Pessoal de Nível Superior—CAPES, Conselho Nacional do Desenvolvimento Científico e Tecnológico—CNPq, and Fundação de Apoio à Pesquisa do Distrito Federal—FAPDF for financial support.

AUTHOR CONTRIBUTIONS

S.S. was involved with project design, laboratory work, data analysis and paper writing, M.G. helped with lab work, G.C. contributed samples and D.L.R. helped design the project. All authors read and approved the manuscript.

DATA ACCESSIBILITY

Raw reads: associated with BioProject PRJNA382381. Probe sequences for SqCL v1 and v2 available at https://github.com/singhal/SqCL. Assemblies for all 44 species available at https://doi.org/10.5061/dryad.r0q02. VCF files for the two species for which we called variants available at https://doi.org/10.5061/dryad.r0q02. Scripts used in probe design and data analysis, along with README, available at https://github.com/singhal/SqCL_analysis.

REFERENCES

Aljanabi, S. M., & Martinez, I. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research*, *25*, 4692–4693.

Arcila, D., Ortí, G., Vari, R., Armbruster, J. W., Stiassny, M. L. J., Ko, K. D., . . . Betancur-R, R. (2017). Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology & Evolution*, *1*, 0020.

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, *13*, 1.

Blom, M. P., Bragg, J. G., Potter, S., & Moritz, C. (2016). Accounting for uncertainty in gene tree estimation: Summary-coalescent species tree inference in a challenging radiation of Australian lizards. *Systematic Biology*, *66*(3), 352–366. https://doi.org/10.1093/sysbio/syw089.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., . . . Drummond, A. J. (2014). beast 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *10*, e1003537.

Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2015). Exon capture phylogenomics: Efficacy across scales of divergence. *Molecular Ecology Resources*, *16*, 1059–1068.

Brandley, M. C., Bragg, J. G., Singhal, S., Chapple, D. G., Jennings, C. K., Lemmon, A. R., . . . Moritz, C. (2015). Evaluating the performance of anchored hybrid enrichment at the tips of the tree of life: A phylogenetic analysis of Australian Eugongylus group scincid lizards. *BMC Evolutionary Biology*, *15*, 1.

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

Budenhagen, C., Lemmon, A. R., Lemmon, E. M., Bruhl, J., Cappa, J., Clement, W. L., . . . Mast, A. (2016). Anchored phylogenomics of angiosperms I: Assessing the robustness of phylogenetic estimates. *bioRxiv*, 086298, doi: https://doi.org/10.1101/086298.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 1.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*, 540–552.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*, 195–205.

Colli, G. R., Bastos, R. P., Araujo, A. F., Oliveira, P., & Marquis, R. (2002). The character and dynamics of the Cerrado herpetofauna. In P. S. Oliveira & R. J. Marquis (Eds.), *The Cerrados of Brazil: Ecology and natural history of a neotropical savanna*, *1*, 223–241. New York, NY: Columbia University Press.

Cosart, T., Beja-Pereira, A., Chen, S., Ng, S. B., Shendure, J., & Luikart, G. (2011). Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics*, *12*, 347.

Crawford, N. G., Faircloth, B. C., McCormack, J. E., Brumfield, R. T., Winker, K., & Glenn, T. C. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, *8*, 783–786.

Dornburg, A., Townsend, J. P., Friedman, M., & Near, T. J. (2014). Phylogenetic informativeness reconciles ray-finned fish molecular divergence times. *BMC Evolutionary Biology*, *14*, 169.

Doyle, V. P., Young, R. E., Naylor, G. J., & Brown, J. M. (2015). Can we identify genes with increased phylogenetic reliability? *Systematic Biology*, *64*, 824–837.

Faircloth, B. C. (2015). phyluce is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *2*(5), 786–788. https://doi.org/10.1093/bioinformatics/btv646.

Faircloth, B. C., Branstetter, M. G., White, N. D., & Brady, S. G. (2015). Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Molecular Ecology Resources*, *15*, 489–501.

Faircloth, B. C., Chang, J., & Alfaro, M. E. (2012). TAPIR enables high-throughput estimation and comparison of phylogenetic informativeness using locus-specific substitution models. arXiv preprint arXiv:1202.1215.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*(5), 717–726. https://doi.org/10.1093/sysbio/sys004

Giarla, T. C., & Esselstyn, J. A. (2015). The challenges of resolving a rapid, recent radiation: Empirical and simulated phylogenomics of Philippine shrews. *Systematic Biology*, *64*, 727–740.

Gilbert, P. S., Chang, J., Pan, C., Sobel, E. M., Sinsheimer, J. S., Faircloth, B. C., & Alfaro, M. E. (2015). Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. *Molecular Phylogenetics and Evolution*, *92*, 140–146.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*, 644–652.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PHYML 3.0. *Systematic Biology*, *59*, 307–321.

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—A baiting and iterative mapping approach. *Nucleic Acids Research*, *41*(13), e129. https://doi.org/10.1093/nar/gkt371.

Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E., & Challenger, W. (2008). GEIGER: Investigating evolutionary radiations. *Bioinformatics*, *24*, 129–131.

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., & Brumfield, R. T. (2016). Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Systematic Biology*, *65*(5), 910–924. https://doi.org/10.1093/sysbio/syw036.

Hillier, L. W., Miller, W., Birney, E., Warren, W., Hardison, R. C., Ponting, C. P., ... Smith, A. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, *432*, 695–716.

Hosner, P. A., Faircloth, B. C., Glenn, T. C., Braun, E. L., & Kimball, R. T. (2016). Avoiding missing data bases in phylogenomic inference: An empirical study in the landfowl (Aves: Galliformes). *Molecular Biology and Evolution*, *33*, 1110–1125.

Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, *9*, 868–877.

Hugall, A. F., O'Hara, T. D., Hunjan, S., Nilsen, R., & Moussalli, A. (2015). An exon-capture system for the entire class Ophiuroidea. *Molecular Biology and Evolution*, *33*(1), 281–294. https://doi.org/10.1093/molbev/msv216

Jetz, W., Thomas, G., Joy, J., Hartmann, K., & Mooers, A. (2012). The global diversity of birds in space and time. *Nature*, *491*, 444–448.

Jombart, T., Kendall, M., Almagro-Garcia, J., & Colijn, C. (2015). TREESCAPE: Statistical exploration of landscapes of phylogenetic trees. *R package version*, **1, 15.**

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*, 185–202.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*, 772–780.

Kent, W. J. (2002). BLAT—The BLAST-like alignment tool. *Genome Research*, *12*, 656–664.

Kubatko, L. S., & Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, *56*, 17–24.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., & Calcott, B. (2017). PARTITIONFINDER2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, *34*(3), 772–773. https://doi.org/10.1093/molbev/msw260

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, *61*(5), 727–744. https://doi.org/10.1093/sysbio/sys049

Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *44*, 99–121.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

McCormack, J. E., Tsai, W. L., & Faircloth, B. C. (2016). Sequence capture of ultraconserved elements from bird museum specimens. *Molecular Ecology Resources*, *16*, 1189–1203.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. A. (2010). The genome analysis toolkit: A MAPREDUCE framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303.

Meiklejohn, K. A., Faircloth, B. C., Glenn, T. C., Kimball, R. T., & Braun, E. L. (2016). Analysis of a rapid evolutionary radiation using ultraconserved elements: evidence for a bias in some multispecies coalescent methods. *Systematic Biology*, *65*(4), 612–627. https://doi.org/10.1093/sysbio/syw014

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, *31*, i44–i52.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*, 443–451.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*, 289–290.

Portik, D. M., Smith, L. L., & Bi, K. (2016). An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Molecular Ecology Resources*, *16*, 1069–1083.

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Moriarty Lemmon, E., & Lemmon, A. R. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, *526*, 569–573.

Pyron, R. A., Burbrink, F. T., & Wiens, J. J. (2013). A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evolutionary Biology*, *13*, 1.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, *461*, 489–494.

Ruane, S., Raxworthy, C., Lemmon, A., Moriarty Lemmon, E., & Burbrink, F. (2015). Comparing species tree estimation with large anchored phylogenomic and small Sanger-sequenced molecular datasets: An empirical study on Malagasy pseudoxyrhophiine snakes. *BMC Evolutionary Biology*, *15*, https://doi.org/10.1186/s12862-12015-10503-12861

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. *Molecular Biology and Evolution*, *31*(5), 1261–1271. https://doi.org/10.1093/molbev/msu061

Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C., & Brumfield, R. T. (2014). Target capture and massively parallel sequencing of ultraconserved elements (UCEs) for comparative studies at shallow evolutionary time scales. *Systematic Biology*, *63*, 83–95.

Springer, M. S., & Gatesy, J. (2016). The gene tree delusion. *Molecular Phylogenetics and Evolution*, *94*, 1–33.

Stamatakis, A. (2006). RAXML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*, 2688–2690.

Swofford, D. L. (2003). PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, *105*, 437–460.

Therkildsen, N. O., & Palumbi, S. R. (2016). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, *17*, 194–208.

Toews, D. P., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, *21*, 3907–3930.

Townsend, J. P. (2007). Profiling phylogenetic informativeness. *Systematic Biology*, *56*, 222–231.

Warnow, T. (2011). Concatenation analyses in the presence of incomplete lineage sorting. *PLoS Currents*, *7*, doi: https://doi.org/10.1371/currents.tol.8d41ac0f13d1abedf4c4a59f5d17b1f7.

Wiens, J. J. (2003). Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology*, *52*, 528–538.

Wiens, J. J., Hutter, C. R., Mulcahy, D. G., Noonan, B. P., Townsend, T. M., Sites, J. W. Jr., & Reeder, T. W. (2012). Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. *Biology Letters*, *8*, 1043–1046.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). GGTREE: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*, 28–36.

Zarza, E., Faircloth, B. C., Tsai, W. L. E., Bryson, R. W. Jr., Klicka, J., & McCormack, J. E. (2016). Hidden histories of gene flow in highland birds revealed with genomic markers. *Molecular Ecology*, *25*, 5144–5157.

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*, 614–620.

Zheng, Y., & Wiens, J. J. (2016). Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. *Molecular Phylogenetics and Evolution*, *94*, 537–547.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Singhal S, Grundler M, Colli G, Rabosky DL. Squamate Conserved Loci (SqCL): A unified set of conserved loci for phylogenomics and population genetics of squamate reptiles. *Mol Ecol Resour*. 2017;17:e12–e24. https://doi.org/10.1111/1755-0998.12681