Squamate Conserved Loci (SqCL): a unified set of conserved loci for phylogenomics and population genetics of squamate reptiles
Supplementary Online Material

Sonal Singhal, Maggie Grundler, Guarino Colli, Daniel Rabosky

March 3, 2017

Contents

1 Figures

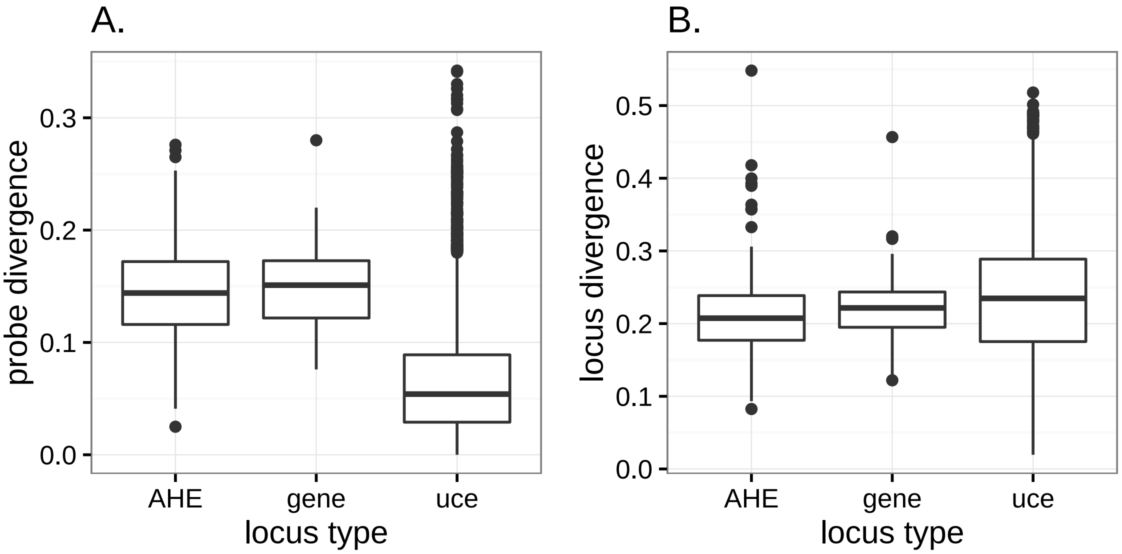


Figure S1: Sequence divergence for (A.) probes and (B.) loci across the three loci types in the SqCL set (anchored hybrid enrichment loci: AHE; traditional phylogenetic genes: gene; ultraconserved elements: UCE). Divergences were measured across the 44 taxa sampled in this experiment. Although UCEs exhibit much less sequence divergence in their probe sequence than AHEs and genes, the divergence across assembled loci are similar across all three loci types.

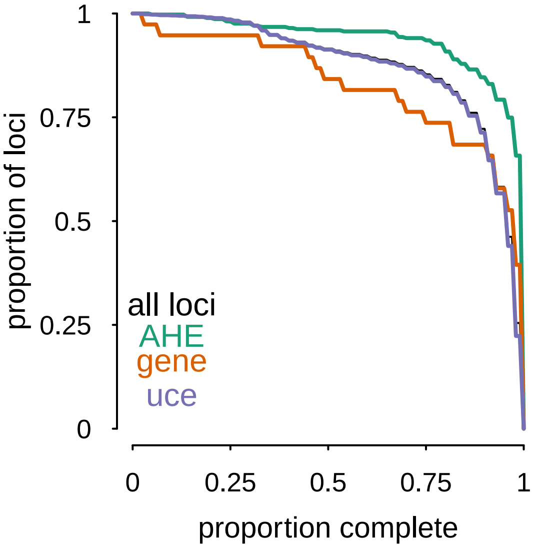


Figure S2: Completeness of the data matrix across the 44 squamate species drawn from 16 major squamate clades and the three types of loci contained within the SqCL set: anchored hybrid enrichment (AHE), standard phylogenetic genes (gene), and ultraconserved elements (UCEs). Black line (all loci) is similar to UCE line. The loci in the SqCL set are recovered well across all taxa, leading to very complete data matrices.

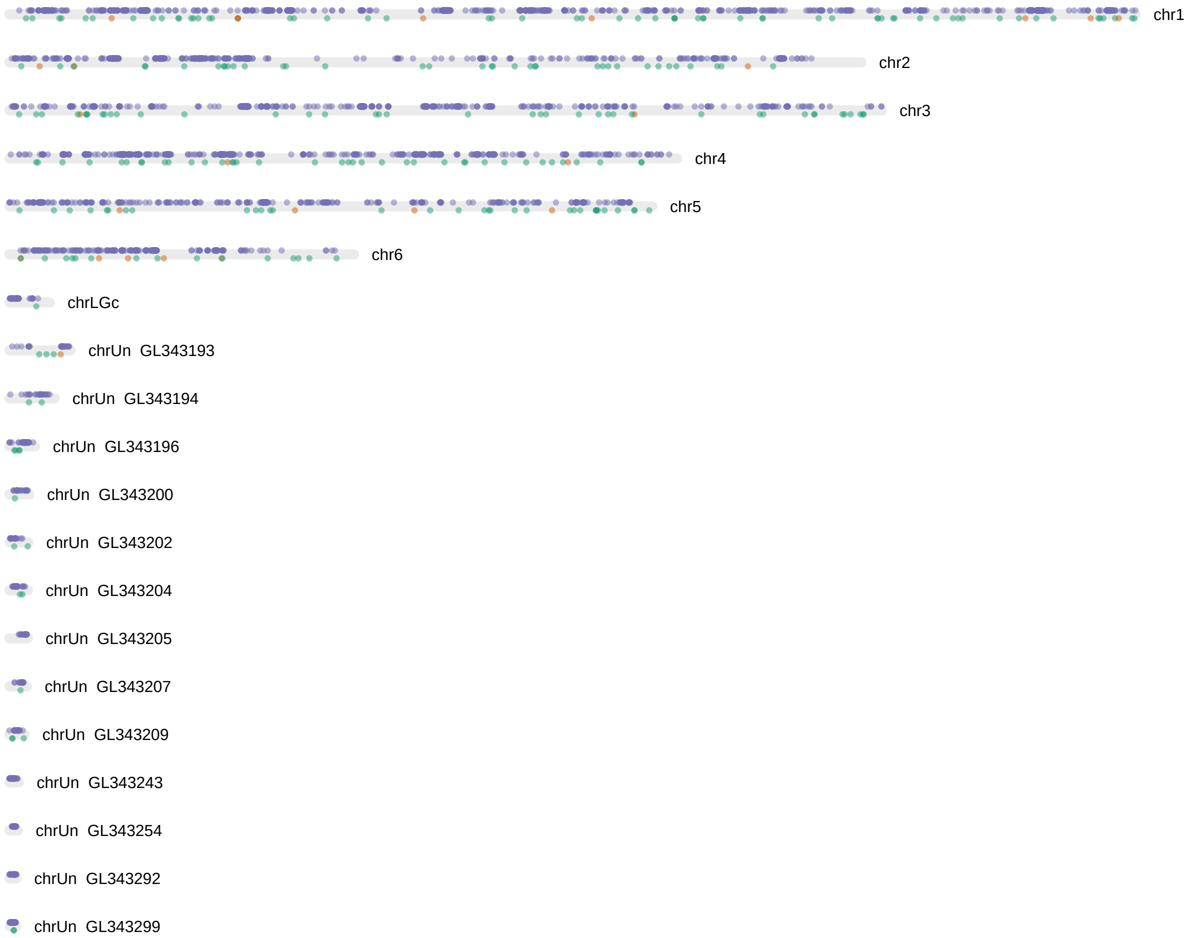


Figure S3: Location of SqCL targets on the *Anolis carolinensis* genome (AnoCar2). Green dots represent anchored hybrid enrichment (AHE) loci, purple dots ultraconserved element (UCE) loci, and orange dots traditional phylogenetic genes. Only those chromosomes and scaffolds containing 25 or more loci are shown. The loci in the SqCL set are well-distributed across the genome, increasing the likelihood that they are evolving independently.

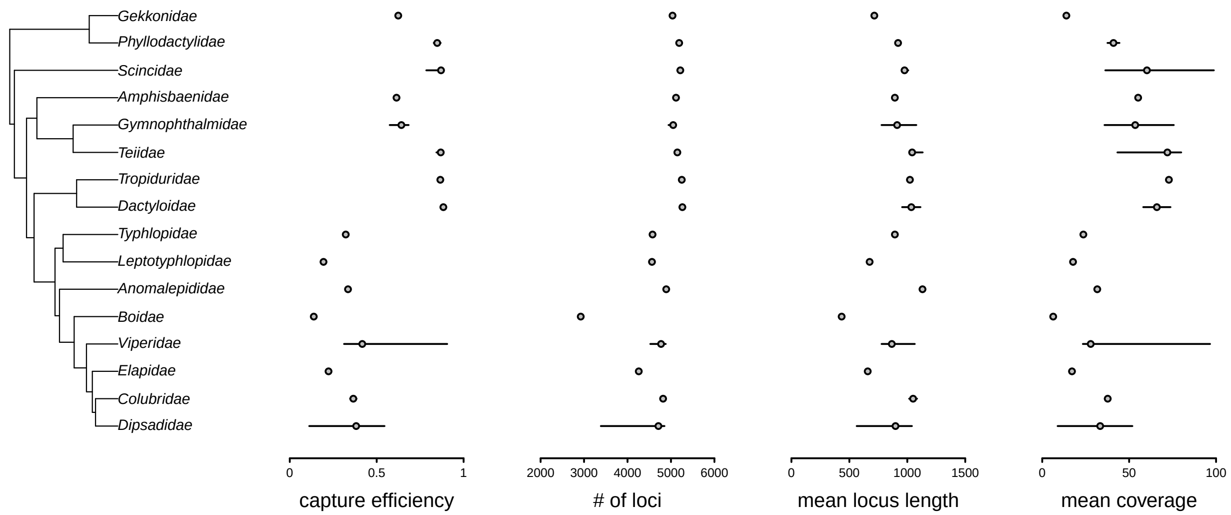


Figure S4: Data quality metrics include: capture efficiency, or the percent of sequenced reads that map to targeted loci, number of loci, or the number of unique loci assembled that match to targeted loci, mean locus length, and mean coverage. Shown are median values and the 95% percentile range across individuals sampled for that family. Not all points are shown with confidence intervals because we only sampled one species in some families. In general, these metrics show the SqCL markers work well, though the snakes generally worked less well than the lizards. A version of this figure showing patterns across the three types of loci in the SqCL set is shown in Fig. 2.

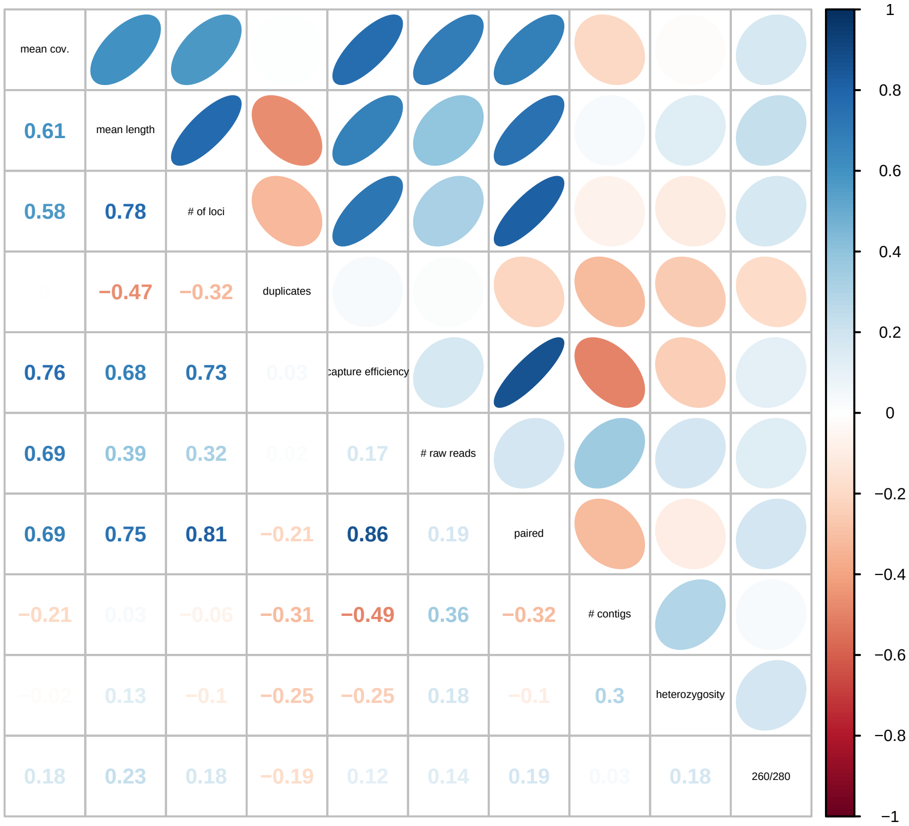


Figure S5: Correlations between 10 metrics describing individual samples: mean coverage – the mean coverage across all loci, mean length – the mean length of targeted loci as assembled, number of loci – the number of targeted loci recovered, duplicates – the percentage of mapped reads that were identified as PCR duplicates, capture efficiency – the percentage of sequenced reads that map on target, raw reads – the number of reads sequenced for an individual, paired – the percentage of mapped reads that mapped as valid pairs, number of contigs – the total number of contigs in the assembly, heterozygosity – the estimated heterozygosity across all targeted loci, and 260 / 280 – a measure of the sample's DNA purity as measured by a NanoDrop UV-Vis Spectrophotometer. Duplicate read rates were high in this experiment (36.7%), and these high rates do not appear to be simply the result of increased sequencing effort.

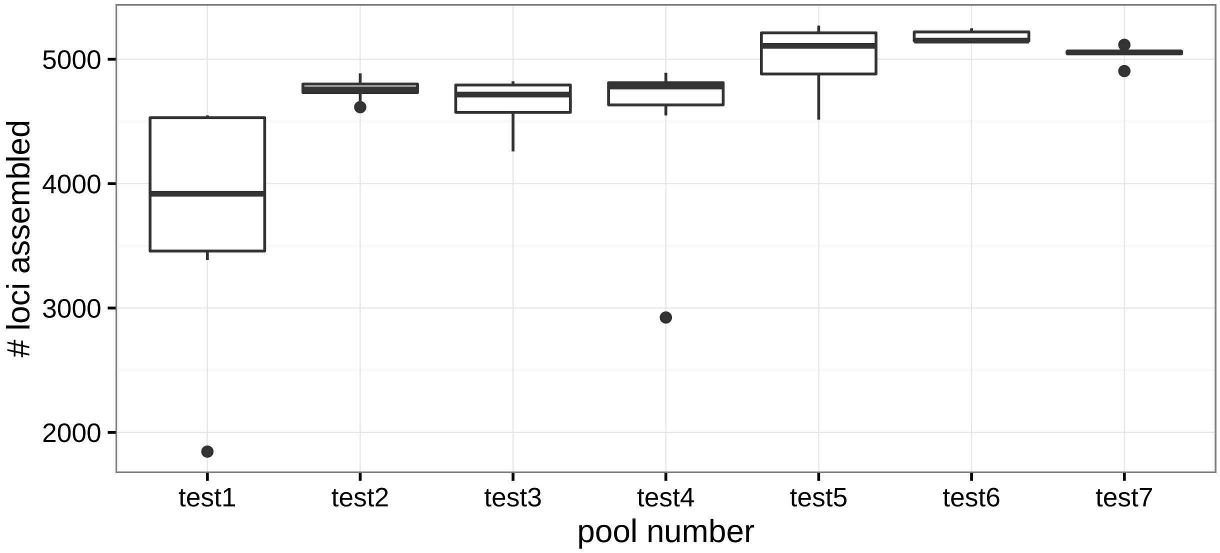


Figure S6: The relationship between the pool in which an individual was captured and the number of loci assembled. We fit a series of single-variable linear models to explain variance in assembly success across individuals, as measured by the number of loci assembled for that individual. Of the multiple factors tested (which includes taxonomic identity of samples, DNA quality, age of tissue sample, number of reads sequenced for that individual), the best predictor of assembly success was the pool in which the individual was captured (adjusted r^2=0.47; p<0.001). Pool is correlated with taxonomic identity, because samples were pooled by taxonomic family. However, both pool ”test1” and ”test2” consist solely of individuals from the family Dipsadidae, yet performed very differently.

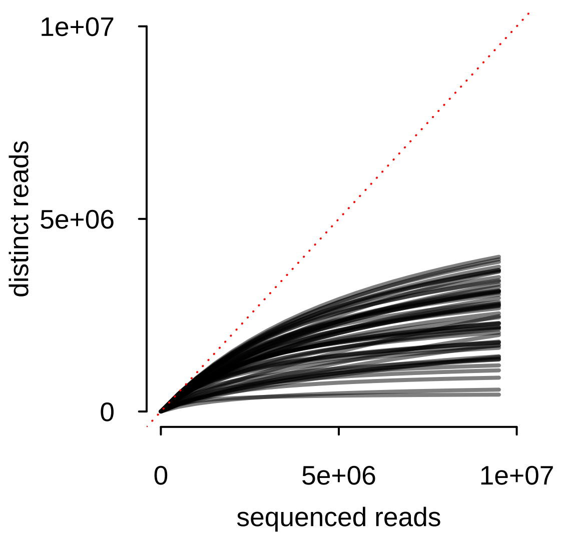


Figure S7: The complexity of sample libraries as evaluated by the module lc_extrap in PreSeq v2.0 (Daley and Smith, 2014). Each line represents one sample; the x-axis indicates the number of reads sequenced and the y-axis indicates the number of distinct reads. The dotted red line is unity, which represents a library with no reduced complexity. These results suggest that the sample libraries vary in complexity and that the reduced complexity of libraries decreases the impact of increased sequencing effort.

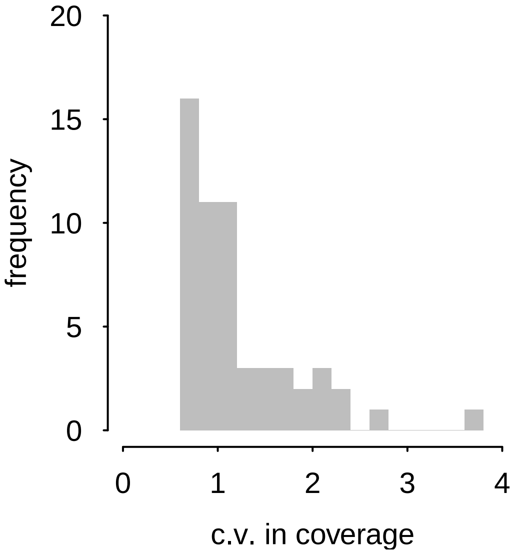


Figure S8: The coefficient of variation of coverage across all loci for a given individual (N=56). Ideally, target capture experiments should show low variance in coverage across loci, as seen here.

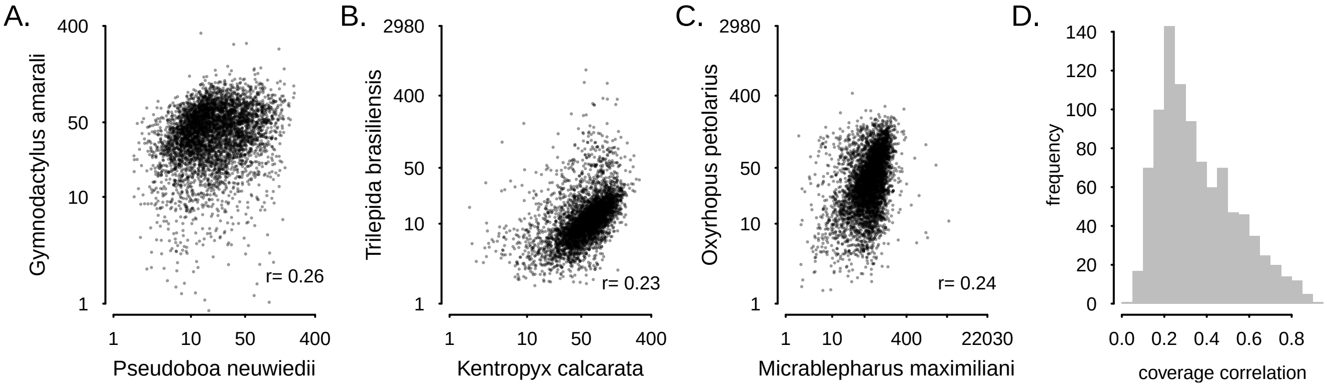


Figure S9: The correlation in coverage among loci between (A - C) three randomly sampled pairs of individuals and (D) the correlations in coverage across all pairwise comparisons. Ideally, target capture experiments should show high correlation in coverage across loci across individuals, because this leads to more complete data matrices.

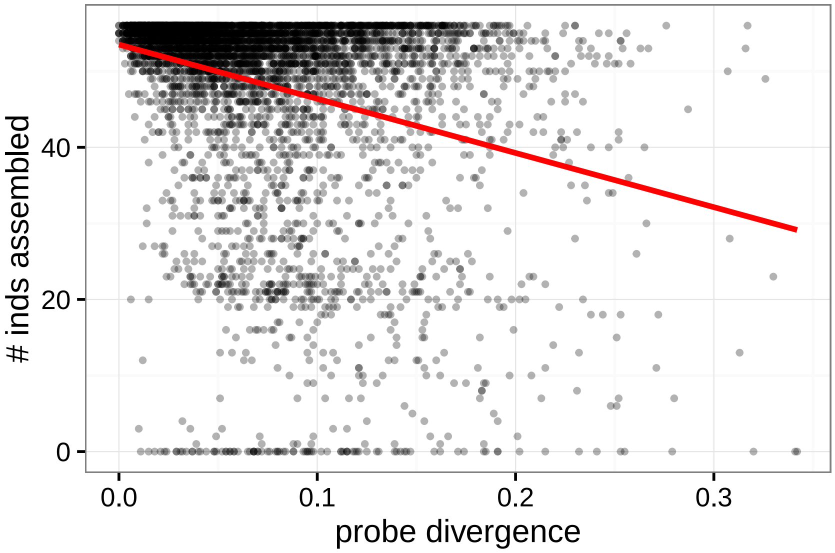


Figure S10: The relationship between divergence in probe sequence across taxa and the number of individuals assembled for the locus targeted by that probe. We fit a series of single-variable linear models to explain variance in assembly success across loci, as measured by the number of individuals assembled for that locus. Of the multiple factors tested (which includes GC content of probes and loci, repeat content of both probes and loci, nucleotide divergence across sampled taxa for probes and loci, number of probes used for the loci), the best predictor of assembly success was the divergence in probe sequence (adjusted r^2=0.09; $p < 0.001$).

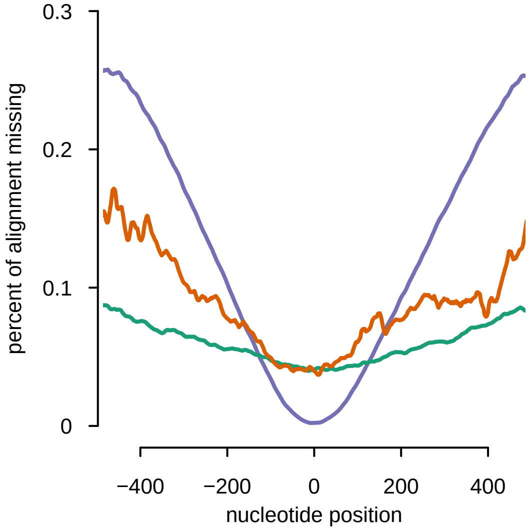


Figure S11: The average amount of missing data across phylogenetic alignments for the three types of loci in the SqCL set (green: anchored hybrid enrichment (AHE) loci, purple: ultraconserved element (UCE) loci, and orange: traditional phylogenetic genes). Coordinates are given with respect to the midpoint of the probe sequences used to capture the loci. These results show that the loci types exhibit different patterns of missingness across the length of the loci. In particular, UCE markers show a u-shaped curve, with increased levels of missingness towards the ends of the alignment. These patterns of missingness also suggest alignments should be trimmed before used for phylogeny aligment.

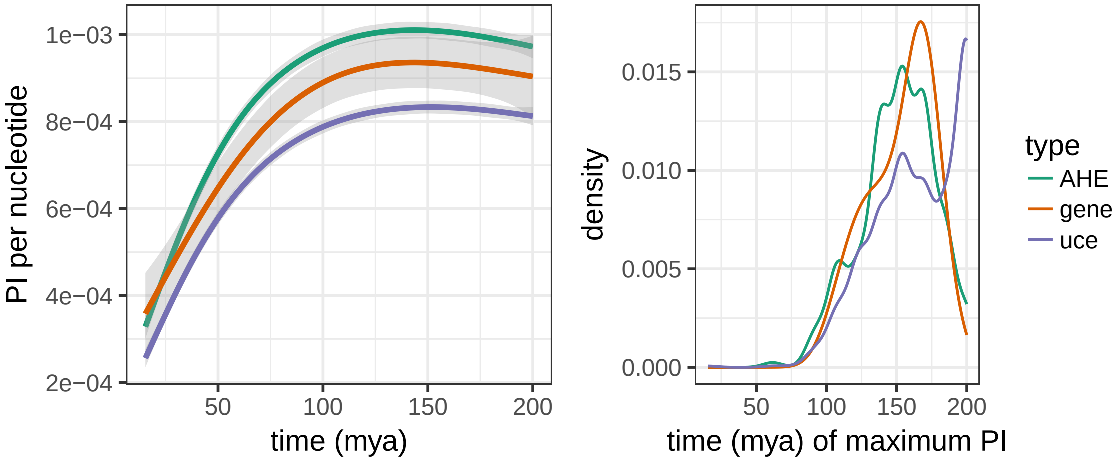


Figure S12: Phylogenetic informativeness (PI) across the three types of loci in the SqCL set (anchored hybrid enrichment: AHE; traditional phylogenetic genes: gene, ultraconserved element: UCE). PI was measured per nucleotide across the depth of the phylogeny. We used the phylogeny shown in Fig. 1, rescaling it so its crown age matched that inferred by other phylogenetic studies in squamates. Also shown is the tree depth at which locu show maximum PI. These results suggest that all three types of markers remain useful at resolving relationships at deeper time scales.

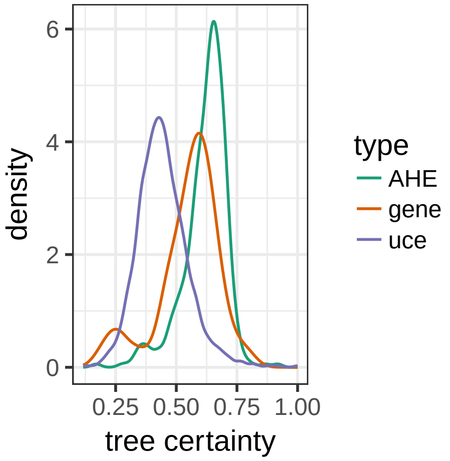


Figure S13: Estimates of tree certainty (TC) across gene trees inferred for each locus as defined by Salichos and Rokas (2013). Tree certainty measures the relative frequency of a bipartition in context of the most common conflicting bipartition in a set of trees. Higher tree certainty scores represent trees that show greater concordance across a set of trees. Here, the set of trees was 100 bootstrap trees per locus, as inferred by RAxML. These results show that trees inferred from anchored hybrid enrichment (AHE) and traditional phylogenetic gene (genes) have greater certainty than those inferred from ultraconserved elements. This TC difference is partially the result of the longer length of AHEs and gene alignments.

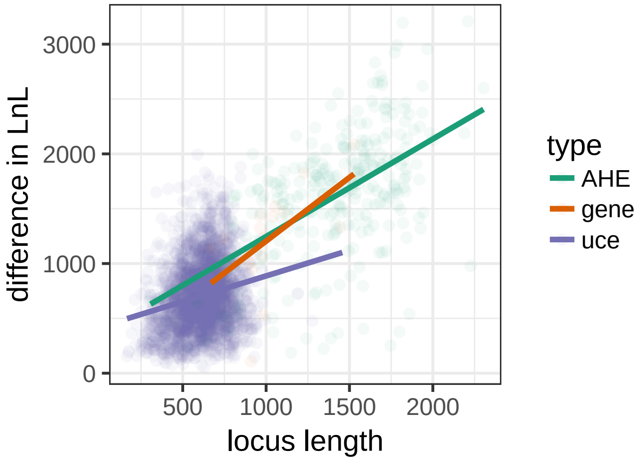
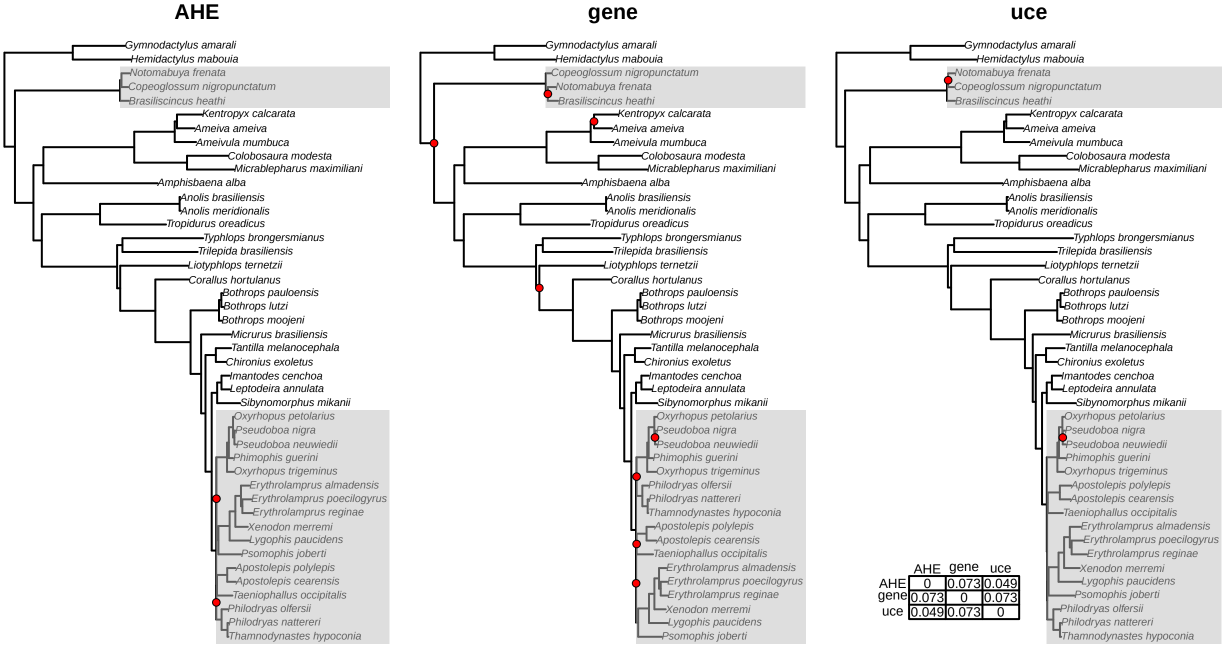


Figure S14: The clock-likeness of anchored hybrid enrichment (AHE) loci, traditional phylogenetic genes, and ultraconserved element (UCE) loci. For each locus, we inferred both unconstrained and ultrametric trees in PAUP and calculated the difference in their log likelihoods. More clock-like behavior results in a lower difference in log-likelihoods. Because this difference is highly correlated with locus length, we show these differences as a function of locus length. Comparing slopes across marker types with an ANOVA, UCEs are significantly more clock-like behavior than AHEs and traditional phylogenetic genes.



	AHE	gene	uce
AHE	0	0.073	0.049
gene	0.073	0	0.073
uce	0.049	0.073	0

Figure S15: Species trees inferred with RAxML using anchored hybrid enrichment (AHE), traditional phylogenetic (gene) loci, and ultraconserved element (UCE) loci that were 95% complete. Trees were rooted with *Gallus gallus* (not shown). Gray boxes mark clades that exhibit unstable topologies across marker sets, and the matrix shows normalized Robinson-Foulds distances between trees. Nodes with <0.95 bootstrap are shown in red. Particularly in the AHE-UCE comparison, we see topological discordance between trees at nodes that have high topological support.

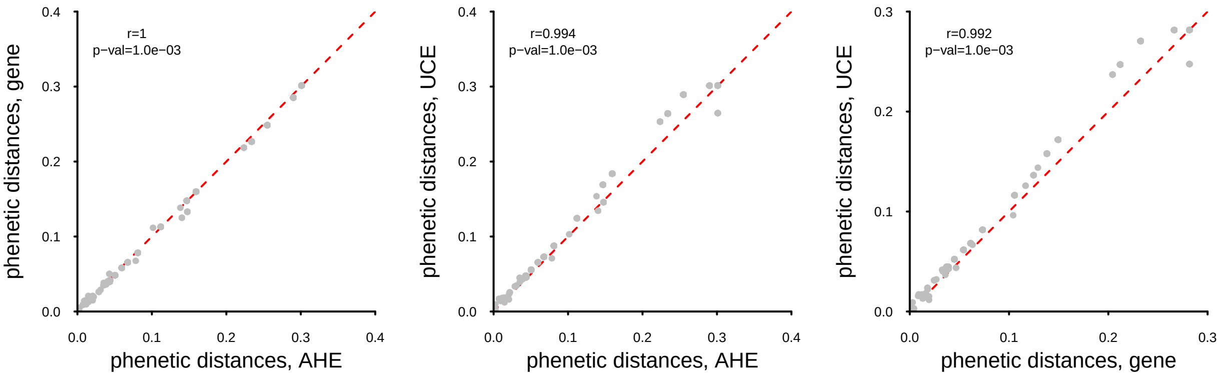


Figure S16: Relationships of co-phenetic distances between taxa from trees inferred with anchored hybrid enrichment (AHE) loci, ultraconserved element (UCE) loci, and traditional phylogenetic genes (gene). These tree topologies were inferred using ASTRAL-II and ultrametric trees for these topologies were inferred with BEAST2. Trees were rescaled to the same height prior to testing for correlations, because these trees were dated without using fossil calibrations or substitution rates. Correlations were measured using a Mantel test; red lines represent a one-to-one relationship. Distances between taxa are highly correlated across marker types, suggesting branch length estimation is robust to differences amongst markers.

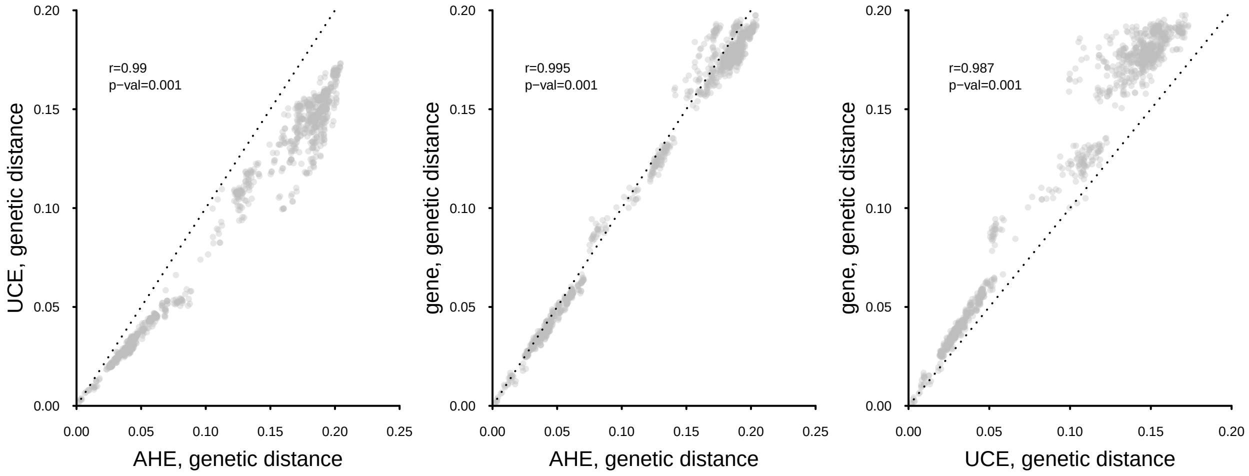


Figure S17: Relationships of raw genetic distances between taxa using concatenated alignments of anchored hybrid enrichment (AHE) loci, ultraconserved element (UCE) loci, and traditional phylogenetic genes (gene). Correlations were measured using a Mantel test; dotted black lines represent a one-to-one relationship. Genetic distances among taxa are highly correlated across marker types, even though they deviate from unity.

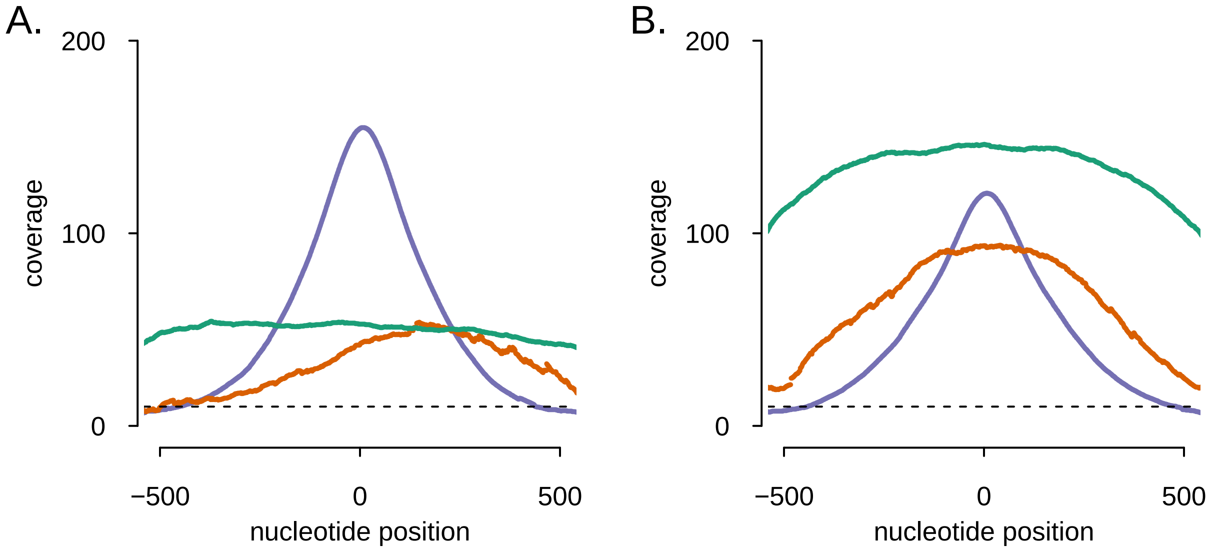


Figure S18: Average coverage across the three types of loci in the SqCL set (green: anchored hybrid enrichment (AHE) loci, purple: ultraconserved element (UCE) loci, and orange: traditional phylogenetic genes) for (A) the six individuals sampled for the lizard *Colobosaura modesta* and (B) the five individuals sampled for the snake *Bothrops moojeni*. The dotted line represents 10× coverage, below which we do not call variants. Coordinates are given with respect to the midpoint of the probe sequences used to capture the loci. These results show that the different loci types exhibit different coverage patterns across the length of the loci, which will affect the completeness of data matrices.

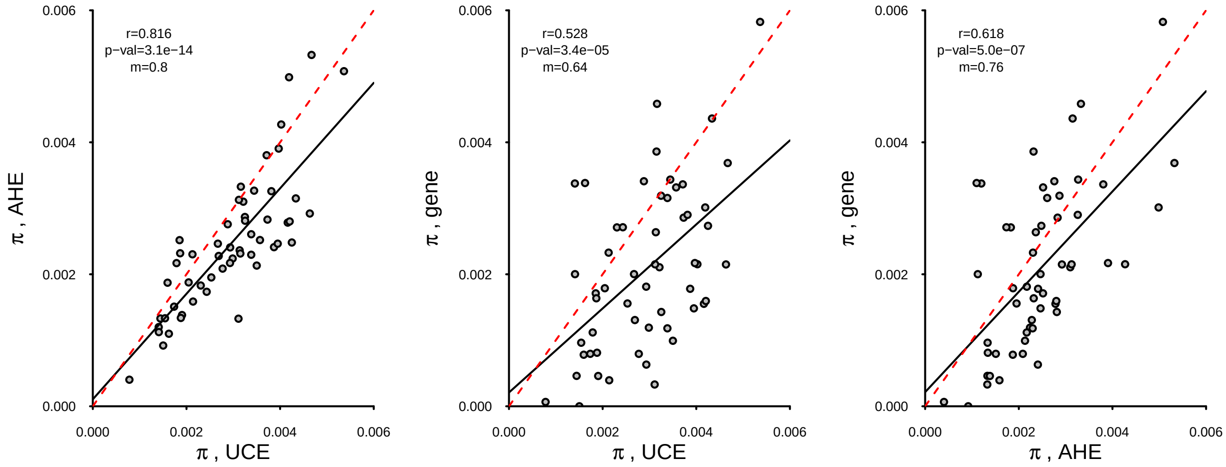


Figure S19: Estimates of genetic diversity (π) as measured for the 56 individuals included in this study at the three marker types: anchored hybrid enrichment (AHE) loci, ultraconserved element (UCE) loci, and traditional phylogenetic genes (gene). The red dotted line is at unity; the black line shows the best-fit linear model between the estimates. Also shown are the Pearson correlations. Esimates of genetic diversity are highly correlated between AHE - UCE markers and less-so between genes and other markers. This is likely because the genes only contained (on average) 5% and 0.8% of the sequence sampled for AHEs and UCEs. All relationships deviate from unity, suggesting that the average effective population sizes of these markers differs.

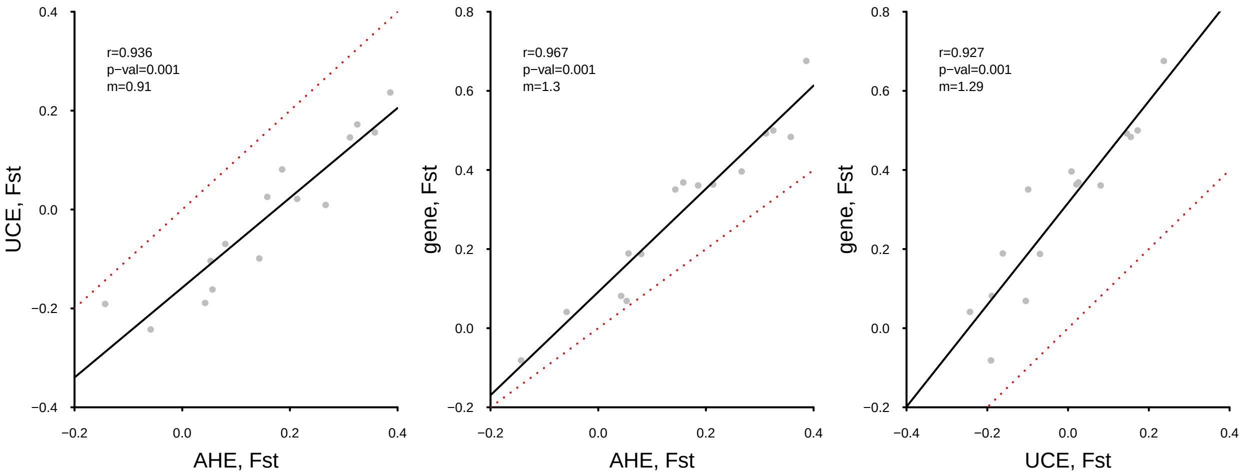


Figure S20: Estimates of genetic differentiation (F_{ST}) as measured for the 6 individuals of *Colobosaura modesta* at three maker types: anchored hybrid enrichment (AHE) loci, ultraconserved element (UCE) loci, and traditional phylogenetic genes (gene). The red dotted line is at unity; the black line shows the best-fit linear model between the estimates. Also shown are the correlations as measured by a Mantel test. Estimates of genetic differentation are significantly and highly correlated between marker types. All relationships deviate from unity, suggesting that the markers have different effective population sizes on average.

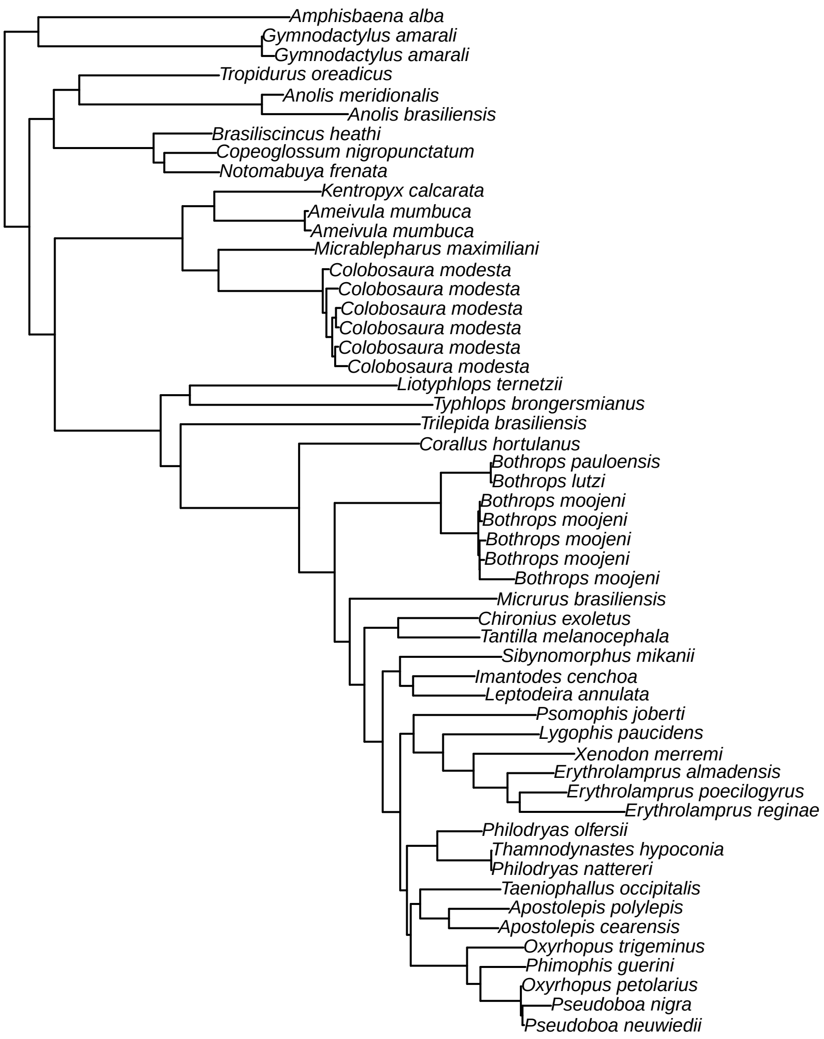


Figure S21: A mitochondrial gene tree for 53 of the 56 individuals sequenced in this study; this gene tree consists of concatenated alignments for the 13 polypeptide genes in the mitochondrial genome and was inferred using RAxML. On average, we recovered 89.2% of the total length of the mitochondrial genome for the 53 individuals shown here. There are some topological differences between this tree and the tree based in nuclear data (shown in Fig. 1) largely at deeper nodes. However, many of the relationships within a family are the same.

2 Tables

Table S1: Data on the individual samples used in this study, including the sample names, their family and species, their locality, their latitude and longitude, the date they were collected, DNA quality measured by 260/280, the pool in which they are captured, the raw number of reads sequenced, the percent of reads mapping to target, the percent of mapped reads that were identified as duplicates, the number of targeted loci recovered, the mean length of those loci, the mean coverage of these loci, and the number of sites with $>10\times$ coverage. These samples were collected under guidelines by the appropriate Brazilian environmental agencies under ICMBio permits 23164-1 and 13324-1.

Species	Family	Notes	Citation
Gekko japonicus	Gekkonidae		(Liu *et al.*, 2015)
Anolis carolinensis	Dactyloidae		(Alföldi *et al.*, 2011)
Ophisaurus gracilis	Anguidae		(Song *et al.*, 2015)
Pogona vitticeps	Agamidae		(Georges *et al.*, 2015)
Vipera berus	Viperidae		http://www.ncbi.nlm.nih.gov/genome/14467
Crotalus mitchellii	Viperidae		(Vonk *et al.*, 2013)
Ophiophagus hannah	Elapidae		(Gilbert *et al.*, 2014)
Pantheropis guttatus	Colubridae		(Ullate-Agote *et al.*, 2015)
Python molurus	Pythonidae		(Castoe *et al.*, 2013)
Thamnophis sirtalis	Colubridae		(McGlothlin *et al.*, 2014)
Boa constrictor	Boidae	used genome 7C assembled with SGA	(Bradnam *et al.*, 2013)

Table S2: Reptile genomes used to design SqCL probe set.

References

Alföldi, J., F. Di Palma, M. Grabherr, C. Williams, L. Kong, E. Mauceli, P. Russell, C. B. Lowe, R. E. Glor, J. D. Jaffe, *et al.*, 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature 477:587–591.

Bradnam, K. R., J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, *et al.*, 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2:1.

Castoe, T. A., A. J. De Koning, K. T. Hall, D. C. Card, D. R. Schield, M. K. Fujita, R. P. Ruggiero, J. F. Degner, J. M. Daza, W. Gu, *et al.*, 2013. The burmese python genome reveals the molecular basis for extreme adaptation in snakes. Proceedings of the National Academy of Sciences 110:20645–20650.

Daley, T. and A. D. Smith, 2014. Modeling genome coverage in single-cell sequencing. Bioinformatics P. btu540.

Georges, A., Q. Li, J. Lian, D. OMeally, J. Deakin, Z. Wang, P. Zhang, M. Fujita, H. R. Patel, C. E. Holleley, *et al.*, 2015. High-coverage sequencing and annotated assembly of the genome of the australian dragon lizard pogona vitticeps. Gigascience 4:1.

Gilbert, C., J. Meik, D. Dashevsky, D. Card, T. Castoe, and S. Schaack, 2014. Endogenous hepadnaviruses, bornaviruses and circoviruses in snakes. Proceedings of the Royal Society of London B: Biological Sciences 281:20141122.

Liu, Y., Q. Zhou, Y. Wang, L. Luo, J. Yang, L. Yang, M. Liu, Y. Li, T. Qian, Y. Zheng, *et al.*, 2015. Gekko japonicus genome reveals evolution of adhesive toe pads and tail regeneration. Nature communications 6.

McGlothlin, J. W., J. P. Chuckalovcak, D. E. Janes, S. V. Edwards, C. R. Feldman, E. D. Brodie, and M. E. Pfrender, 2014. Parallel evolution of tetrodotoxin resistance in three voltage-gated sodium channel genes in the garter snake thamnophis sirtalis. Molecular biology and evolution 31:2836–2846.

Song, B., S. Cheng, Y. Sun, X. Zhong, J. Jin, R. Guan, R. W. Murphy, J. Che, Y. Zhang, and X. Liu, 2015. A genome draft of the legless anguid lizard, ophisaurus gracilis. GigaScience 4:1.

Ullate-Agote, A., M. C. Milinkovitch, and A. C. Tzika, 2015. The genome sequence of the corn snake (pantherophis guttatus), a valuable resource for evodevo studies in squamates. International Journal of Developmental Biology 58:881–888.

Vonk, F. J., N. R. Casewell, C. V. Henkel, A. M. Heimberg, H. J. Jansen, R. J. McCleary, H. M. Kerkkamp, R. A. Vos, I. Guerreiro, J. J. Calvete, *et al.*, 2013. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. Proceedings of the National Academy of Sciences 110:20651–20656.