An ancient adaptive episode of convergent molecular evolution confounds phylogenetic inference

Todd A. Castoe[1,*], A. P. Jason de Koning[1,*], Hyun-Min Kim[1], Wanjun Gu[1], Brice P. Noonan[2], Zhi J. Jiang[3], Christopher L. Parkinson[4], and David D. Pollock[1‡]

[1]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045 USA

[2]Department of Biology, University of Mississippi, Box 1848, University, MS 38677 USA

[3]Center for Computational Science, University of Miami, 1120 NW 14th Street, Miami, FL 33136, USA

[4]Department of Biology, University of Central Florida, 4000 Central Florida Blvd., Orlando, FL 32816USA

*The first two authors contributed equally

‡*Corresponding Author*: David D. Pollock, Department of Biochemistry and Molecular Genetics, University of Colorado Health Sciences Center, Aurora, CO, 80045 USA.
Email: David.Pollock@uchsc.edu *phone:* 303-724-3234 *fax:* 303-724-3215

Convergence can mislead phylogenetic inference by mimicking shared ancestry, but has been detected only rarely in molecular evolution. Here, we show that significant convergence occurred in snake and agamid lizard mitochondrial genomes. Most evidence, and most of the mitochondrial genome, supports one phylogenetic tree, but a subset of mostly amino acid-altering mitochondrial sites strongly support a radically different phylogeny. These sites are convergent, probably selected, and overwhelm the signal from other sites. This suggests that convergent molecular evolution can seriously mislead phylogenetics, even with large data sets. Radical phylogenies inconsistent with previous evidence should be treated cautiously.

Although selection-driven convergent evolution of morphological characters has been identified as a potential source of error for phylogenetic inference[1-3], convergence in molecular datasets is believed to be rare. Definitive evidence of convergence at the molecular level is known from only a small number of proteins[4-10]. There have been, however, few searches for convergent molecular evolution in protein sequences, and the true frequency of molecular convergence in nature is therefore unknown; it may be more common than widely believed but difficult to detect, or simply overlooked[4,8,11]. Regardless of its true frequency in nature, identifying convergent molecular evolution when it happens is important for understanding mechanisms of functional adaptation, and to prevent it from causing errors in phylogenetic inference. For example, significant differences in phylogenies inferred from different genes are usually taken to indicate differences in the evolutionary histories of those genes arising from differential patterns of lineage sorting, hybridization, recombination, horizontal gene transfer, or gene duplication and loss[12-17]. In the presence of convergent evolution, however, the differences between the trees might be artifactual and the bases for the inferences would then be invalid.

Although the squamate limb on the tree of life is not fully resolved, there is broad consensus that the iguanas, chameleons, and agamid lizards are close relatives and form an exclusive clade, referred to as the Iguania[2,18-23]. Extensive analyses of all 13 mitochondrial protein-coding genes (> 11 kb), however, provided strong support for a close "sister" relationship between agamid lizards and snakes (Fig. 1; see also Fig. S1A).

This is a radical result not suggested by previous studies. If true, this relationship would disrupt the monophyly of the Iguania, but it is contradicted by our own nuclear gene analyses (Fig. 1; see also Fig. S1B), previous larger nuclear gene studies[2,22,23], and morphological evidence[18-21].

The mitochondrial signal favoring the radical tree is strong enough that the snake-agamid grouping was also supported in combined analysis of the joint mtDNA and nuclear data (Fig. 1), although all other relationships from the combined estimate are in excellent agreement with our nuclear gene trees and previous nuclear gene-based studies[2,22,23]. Hereafter, we refer to the tree estimated from the joint mitochondrial plus nuclear data (Fig. 1) as the "MT" topology, and the same tree but with a monophyletic Iguania (see red arrow in Fig. 1) as the "NUC" topology. The Shimodaira-Hasagawa (S-H) test[24], a standard likelihood-based tree hypothesis testing approach, significantly rejected the NUC in favor of the MT topology for all mitochondrial sequence data together, and for each of the three codon positions separately ($P < 0.01$). Significant rejection of alternative phylogenetic hypotheses based on an S-H test is commonly accepted as conclusive evidence in evolutionary studies. In this case, however, the result is not credible because so many independent data sources support the NUC phylogeny. It must therefore be considered what is wrong with our interpretation of the mitochondrial data.

Here, we consider the various possibilities for what may have led to the strongly-supported incorrect phylogeny estimate for the large mitochondrial dataset. We did this by identifying which sites in the mitochondrial genome support the accepted versus the unorthodox topology, and by evaluating whether support for the unexpected topology is consistent with convergent evolution or some other form of bias. We come to the conclusion that a strong episode of convergent molecular evolution occurred between early lineages of snakes and a group of distantly related lizards. This excess of convergent change is highly significant, and much greater than expected due to homoplasy and neutral parallelism under neutral models with constraint[25,26]. A role for adaptation in this burst of convergence seems plausible, and is consistent with previous

evidence for a strong adaptive burst of mitochondrial protein change early in snake evolution[27].

This case demonstrates that convergent evolution can have a much greater impact on phylogenetic inference than is generally appreciated, even in large datasets, and that adding more data will not necessarily solve the problem. We show here that this convergence event involves numerous genes, and that convergence in a small fraction of the data overwhelms an otherwise strong phylogenetic signal. The probable role of adaptation means that the false phylogenetic clustering of lineages due to convergence can be deterministic. Because molecular convergence may be more common than previously thought[4,11], because even small amounts of convergence can exert a strong phylogenetic bias, and because comparative genomics and much of biology in general rely on accurate phylogenies, these results are disturbing. We argue that adaptive convergence should be considered as an explanation whenever there is phylogenetic conflict among data sets.

RESULTS

Site-specific support for the two topologies

To identify which nucleotide positions supported the presumably incorrect MT tree, we measured the difference in site-specific log likelihood values for each of the two alternative topologies (ΔSSLS) across the mtDNA dataset. A majority of sites support the accepted NUC tree, but this support is overwhelmed by a relatively small number of sites that strongly support the MT topology. Considering only sites with a notable preference for one tree over another ($|\Delta SSLS| > 0.1$), nearly twice as many sites support the conventional NUC tree as support the MT topology (962 versus 537 sites; Fig. S2). If only sites with strong support ($|\Delta SSLS| > 0.5$) are considered, however, the situation is reversed and around five to nine times more sites, depending on codon position, strongly favor the MT tree over the NUC tree (Fig. S3).

One potential explanation for the conflict in phylogenetic signal is that different sites in the mtDNA genuinely have different phylogenetic histories. Such a situation could

conceivably have been caused by gene conversion or recombination[28], although this is unlikely since mtDNA recombination is thought to be rare within vertebrate species[29-31], let alone between such distantly related lineages as snakes and agamid lizards. This hypothesis is further excluded because site-specific support for each tree is widely dispersed throughout the mitochondrial genome (Fig. 2). Gene conversion or recombination should lead to discrete segments of the genome that strongly support one tree over another, and this is not observed. Some genes, including COX1, COX3, CytB, ND1, and ND2, possess more sites that strongly support the MT tree than do other genes, but they still contain a majority of sites that weakly to moderately support the NUC tree (Figs. 2 and 3; also Figs. S2 and S3).

Two remaining possibilities for the conflict in phylogenetic signal are that unusual mutation processes led to reconstruction bias, or that positive or negative selection on amino acids led to unusual substitution patterns that misled phylogenetic inference. An important role for the mutation process is strongly contraindicated by a number of independent lines of evidence. First, nucleotide frequency biases at all sites and at four-fold redundant sites are not particularly similar between snakes and agamids (Fig. S4). Second, log-determinant phylogenetic analyses of the mtDNA, which should reduce sensitivity to base frequency biases[32], recover the MT tree (Fig. S5). Third, amino acid sequences and 2nd codon positions should be the least affected by mutation biases, but Bayesian phylogenetic analyses of these data both lead to trees essentially identical to the MT topology (data not shown). Furthermore, site-specific support for the MT tree is less common at 3rd codon positions than at 1st or 2nd positions (Fig. 2; also Figs. S2 and S3). Four-fold redundant 3rd codon positions, which do not alter the amino acid sequence when they change, provide almost no differential likelihood support favoring either tree (Fig. 2C).

An amino-acid based explanation of the phylogenetic bias is also favored over a mutational explanation because the probability that a site strongly supports the MT topology is inversely related to the relative rate of evolution at that site (Fig. 4). Slowly evolving (generally conserved) sites most strongly contribute to support for the MT

topology, and fast-evolving sites contribute no notable support (Fig. 4). In particular, the majority of phylogenetic signal favoring the MT topology comes from relatively conserved non-synonymous sites, particularly 2nd codon position transversions that are otherwise conserved (Fig. S6); this is most consistent with selection on protein sequences leading to conflicting signal and phylogenetic error.

Effects of removing taxa

Bayesian phylogenetic analyses of the mtDNA data with either of the two agamid species excluded (either *Xenagama* or *Pogona*) produced trees highly similar to the original MT tree, with agamid lizards paired with snakes (data not shown). Thus, both species of agamid lizards have phylogenetic affinity to snakes in the mtDNA data. When all snakes were excluded, the agamids clustered with amphisbaenian lizards (Fig. S7), whereas previous mtDNA studies that did not include the agamids found strong support for pairing snakes with amphisbaenians[33,34]. These major changes in phylogenetic relationships with minor changes in taxon sampling are indicators of phylogenetic conflict and the unreliability of the MT tree, and are unexpected given the large size (> 11,700 bp) of the mitochondrial dataset.

Effect of removing sites that strongly support the MT topology

We performed a Bayesian phylogenetic analysis excluding the 500 codons with the highest ΔSSLS supporting the MT tree. The result, based on the remaining 10,227 bp, recovered a monophyletic Iguania, placing the Agamidae as the sister group to the Iguanidae with 100% posterior support (Fig. S8); this is the presumed correct relationship found in the NUC tree. Thus, removal of less than 13% of the 11,727 bp dataset not only eliminated support for the MT topology as expected, but also revealed support for the presumed correct placement of the agamids as sister to iguanids. This result would not be expected if the MT topology was in fact true; the recovery of the correct agamid-iguanid relationship upon removal of a small subset of sites is clear evidence that the remaining phylogenetic signal supports the expected squamate tree of life. Other analyses (see below) showed that removing as few as 98 codons was enough to eliminate strong support for the incorrect agamid-snake phylogenetic pairing.

Convergent evolution of amino acid sequences

Given that otherwise conserved non-synonymous sites provide the strongest support for the MT topology, it seems likely that this support is due to convergent amino acid evolution between snakes and agamid lizards. To verify this, we used maximum likelihood (ML) and Bayesian posterior approaches to estimate the number of convergent amino acid substitutions between all pairs of branches on the phylogenetic tree. Here, convergent change is defined as changes at the same site along both branches resulting in the same amino acid. The expected number of random convergent changes for each branch-pair will depend on the lengths of the two branches, so to determine the excess above random expectation we compared convergent changes to the estimated number of divergent changes between branch-pairs, which also depends on the branch lengths. Divergent changes are defined here as changes at the same site along both branches, but resulting in different amino acids. In the Bayesian approach, posterior substitution probabilities were calculated by integrating estimates over all possible joint ancestral state assignments at internal nodes (see Methods).

There was a strong linear relationship between the number of divergent and convergent substitutions using both the ML (orthogonal regression $R^2 = 0.812$, $b = 0.103$; Fig. 4A) and Bayesian methods ($R^2 = 0.914$, $b = 0.17$; Fig. 4B). The tightness of this relationship suggests that most convergent substitutions on the tree may have been random (neutral) homoplasies, since they are so well predicted by the number of divergent changes. An important caveat, however, is that these empirical levels of random convergence are far higher than expected if the model used to analyse the data is correct (Figure 4B; see Supplementary Methods). This can be explained by purifying selection, which can alter the random convergence/divergence relationship by constraining the types of amino acids possible at each residue position[11,25].

We also observed substantial differences between the estimates of convergent and divergent changes from the ML and Bayesian analyses. Bayesian estimates predicted fewer divergent substitutions, somewhat less convergence, and overall nearly twice as many convergent changes per divergent substitution (Fig. 4A, B; also Fig. S9). Although

previous analyses of molecular convergence have utilized ML approaches on branch-pairs of *a priori* interest[8], the difference between the ML and Bayesian results observed here bring the accuracy of ML approaches into question, since they ignore error in the unknown ancestral states. Previous related analyses have shown that failure to integrate over unknown ancestral states can lead to misleading biological conclusions[35-37]. Since bias caused by conditioning on optimal ancestral state reconstructions is expected in ML, we primarily consider the Bayesian results hereafter.

Among all branch pairs compared, the number of convergent events between the branches leading to the most recent common ancestors (MRCAs) of snakes and of agamid lizards stands out as being far greater than expected based on the number of divergence events. There were 28 positions in the protein alignments with more than 80% posterior probability of convergent substitution between these two branches. These sites were concentrated in COX1 and ND1, but were present in other proteins as well (Fig. 4C). Remarkably, these two branches of *a priori* interest showed the single greatest excess of convergence of all branch-pairs on the tree (0.28 convergent substitutions per divergent substitution, or 1.6 times the empirically determined random (neutral) expectation; Fig. 4A, B). Partial correlation analyses indicated that the extreme excess of convergence between the snake and agamid branch pair cannot be explained by long branch lengths, which is a conceivable source of bias in posterior convergence estimation (see Supplementary Data).

Using empirically predicted levels of convergence from the orthogonal regressions, a series of binomial tests identified this pair of branches as the only pair with a highly significant probability of excess convergence ($P < 0.001$, after accounting for false discovery[38]). For branch pairs with higher predicted false discovery rates, the expected number of true positives (Fig. S10) is high enough that a further 11 branch-pairs may have experienced an excess of convergence events, although as many as five of these additional branch-pairs are expected to be false positives (Fig. S11).

A sliding window plot of site-specific support for the MT versus NUC topologies and the predicted number of convergent substitutions shows that peaks in site-specific support for

the MT topology coincide with peaks in the probability of convergent substitution (Fig. 4C). The highly significant correlation (r = -0.498, $P < 2.2x10^{-16}$) indicates that sites supporting the incorrect MT tree are likely to represent amino acid convergence. The link between convergent amino acids and phylogenetic error is supported by the observation that removing the 98 codons with the highest probability of convergence (the top 2.5%; see Fig. 4C and site patterns in Fig. S12) was sufficient to cause total likelihood support to switch from favoring the MT topology (total ΔSSLS = -85.77) to favoring the NUC topology (total ΔSSLS = 2.577). Removal of these sites brought the cumulative Bayesian posterior number of predicted convergent substitutions down from 113.5 to 58.8, a number statistically indistinguishable from the 69.6 convergent substitutions predicted from the empirical regression on divergent substitutions (Fig. 4A; $P > 0.09$). It therefore appears that removing the excess convergent substitutions allows the correct underlying signal from the majority of the mitochondrial genome to dominate.

DISCUSSION

This study presents evidence for large-scale molecular convergence between snake and agamid lizard mitochondrial genomes at the amino acid level. These convergent replacements misled phylogenetic reconstruction and falsely joined these two groups as sister taxa, even though they are separated by over a hundred million years of divergence. The degree of convergence observed is well outside expectations based on both empirical distributions and model-based calculations, and was sufficiently large to overcome the correct signal in over 11 kb of sequence data. This result has disturbing implications for the reliability of phylogenetic reconstruction in the presence of convergent evolution, even using statistical methods that are otherwise typically robust and believed free of systematic biases. We discovered this molecular convergence phenomenon because it was so extreme and because it severely disrupted the phylogeny in such a nonsensical way. Smaller and less obvious phylogenetic errors caused by convergence might often be mistakenly accepted as being reliable.

An obvious potential explanation for this phylogenetically-misleading example of convergent evolution is adaptation. It was previously shown that snake mitochondrial

proteins have endured the most extreme burst of apparently adaptive protein evolution yet observed in vertebrate mitochondria[27]; this is consistent with the idea that the excess convergence levels observed here are due to the action of natural selection rather than random homoplasy. It was proposed that the evolutionary burst in snakes may have been driven by selection related to physiological adaptations for metabolic efficiency and to allow radical fluctuations in aerobic metabolic rate[27]. The molecular convergence between snakes and agamid lizards may thus have resulted from shared adaptive pressures on metabolic function. Since the convergence extends across most regions of the mitochondrial genome, any common adaptive force must have been exceptionally strong and broad in scope.

Perhaps the most disturbing aspect of the idea that adaptation may be at the root of this molecular convergence event is the systematic nature of its effects on phylogenetic reconstruction. If such a large convergence event can occur, it is reasonable to suppose that smaller events may be much more common than realized, but are often difficult to detect, or overlooked. A convergence event even a fraction of this magnitude could easily disrupt many topology estimates because of the relative biasing power of each convergent site. Indeed, the first known case of convergent molecular evolution in ruminant lysozymes[6,7] was shown to lend support to a dramatically wrong phylogeny that placed cows within the primates. This occurred even though only a small number of convergent substitutions apparently took place[8]. In the present case, a small fraction of convergent sites dramatically outweighed the accurate signal at hundreds of other sites (e.g., Figs. 2 and 3). Existing evolutionary models assume that convergent molecular evolution is extremely improbable, and thus even small amounts of convergence can be falsely interpreted as extremely strong evidence for an incorrect topology.

We focused here largely on the evidence for convergent molecular evolution in snakes and lizards and the alarming impact it can have on phylogenetic inference. Nevertheless, the implication that this dramatic convergence may have been caused by adaptive pressure on protein function suggests that further study may reveal valuable insight into the function of these proteins. The tendency for convergent amino acid substitutions to

occur at otherwise conserved positions also suggests that many of these convergent changes are likely to have had notable structural and functional effects.

The confounding effects of convergence on phylogeny and its potential informativeness on the sequence, structure, and function relationships mean that the presence and influence of convergent molecular evolution should be scrutinized more aggressively than is currently standard. Convergence should also be incorporated into probabilistic phylogenetic models, if possible. This will provide important insights into molecular evolutionary processes and greater confidence in the phylogenetic inferences that underlie comparative biology and, increasingly, genomics.

METHODS

Mitochondrial genome sequencing, alignment and phylogeny inference. Mitochondrial genomes were sequenced and annotated for two snake species, *Anilius scytale* and *Tropidophis haetianus*, to increase sampling at the base of snake phylogeny (see Supplementary Methods). All 13 mitochondrial protein-coding genes (~11,700 bp) from complete mitochondrial genomes of squamates available at the time of study, plus the two new species, were aligned using ClustalX[39] based on their amino acid translation; multiple species per genus were excluded (Table S1). Representatives of major tetrapod lineages were also included to root the squamate tree. Nucleotide sequences of two nuclear genes, *rag-1* and *c-mos*, were obtained from GenBank and aligned for comparison to the mitochondrial data (Table S2).

For phylogenetic analysis, mitochondrial and nuclear datasets were partitioned by gene and codon position and appropriate partition-specific models were selected (Supplementary Methods). Bayesian phylogenetic trees were estimated in MrBayes 3.0b4[40] with partitioned models for mitochondrial and nuclear, both independently and combined.

Molecular evolutionary analyses and hypothesis testing. Maximum parsimony (MP), log-determinant distance methods, and maximum likelihood (ML) analyses of the

mitochondrial dataset were used to evaluate phylogenetic hypotheses in PAUP* 4.0b10[41] (see Supplementary Methods); where relevant, *P*-values less than 0.05 were considered significant. Evidence for non-stationary base frequencies across lineages was evaluated based on chi-squared tests in PAUP*. Support for alternative topologies was evaluated using the Shimodaira-Hasegawa test[24]. Site-specific likelihood support (SSLS) was estimated using ML and a GTR+Γ+I model (general time-reversible with gamma-distributed and invariant rates among sites) per codon position.

Maximum likelihood analysis of convergent evolution. We used *PAML*[42] to estimate the most likely ancestral states (by marginal ancestral reconstruction using mtREV24+F and a 5-category discrete gamma distribution) across all internal nodes of the NUC topology. We used a Perl script to count the divergent and convergent double amino acid replacements (changes at the same site in two branches) for all pair-wise comparisons of branches. Only counts along separate lineages (*i.e.*, those not sharing a common ancestor) within the squamates were used. Change per branch was estimated based on the maximum likelihood ancestral sequence reconstructions by comparing states at ancestral and descendant nodes per branch. For amino acid sites at which changes occurred along two compared branches, sites with different amino acids in the descendants were defined as divergent, and those with the same amino acid in the descendant were defined as convergent. Analyses of the inferred number of changes were performed in *R*, where a linear model was fit to the numbers of convergent and divergent changes for each branch-pair, using orthogonal regression forced through the origin.

Bayesian analyses of convergent evolution. For our Bayesian approach, we modified the *codeml* program of *PAML*[42] to calculate the posterior probability of all possible amino-acid substitutions along every branch in the phylogeny, while accounting for rate variation across sites (using mtREV24+F+Γ). The posterior probabilities of substitution were used to calculate the probability of all possible convergent and divergent substitutions, and were therefore implicitly integrated over all possible ancestral states. The probability of convergent and divergent substitutions were calculated as the sum of the joint probabilities of all possible pairs of substitutions that end in the same state

(convergent) or in a different state (divergent), between the two branches in question. The details of these calculations are given in the Supplementary Methods.

Using the posterior expected number of convergent substitutions with predicted levels of convergence (from orthogonal linear regressions), we performed one-sided binomial tests for each branch-pair to assess the expected probability of the observed amount of convergence under the null hypothesis provided by the linear regression-based model. The test therefore assumed each site was a drawn from a binomial distribution with a probability of being convergent (p) defined by the expected amount of convergence divided by the number of sites, and a number of trials (n) equal to the number of sites. False discovery controls were applied to all tests, unless otherwise specified. All binomial tests and false discovery controls were performed using scripts written in *R*.

ACKNOWLEDGMENTS

We acknowledge the support of the National Institutes of Health (NIH; GM065612-01, GM065580-01) to DDP, National Science Foundation (DEB–0416000) and a UCF startup package to CLP, and an NIH training grant (LM009451) to TAC.

AUTHOR CONTRIBUTIONS

TAC, APJdK, and DDP co-wrote the manuscript and designed the study. TAC sequenced the new snake mitochondrial genomes and conducted much of the analyses. All authors participated in editing the manuscript. ZJJ conducted some analyses and annotated the new snake genomes. APJdK developed and implemented the Bayesian methods for detecting convergence, and APJdK, WG and HMK performed some of the computational analyses including writing required analytical programs. BPN conducted some analyses and helped with supercomputing. CLP supervised the mitochondrial genome sequencing, and DDP supervised the study.

REFERENCES

1. Harmon, L.J., Kolbe, J.J., Cheverud, J.M. & Losos, J.B. Convergence and the multidimensional niche. *Evolution* **59**, 409-21 (2005).

2. Lee, M.S.Y. Convergent evolution and character correlation in burrowing reptiles: towards a resolution of squamate relationships. *Biological Journal of the Linnean Society* **65**, 369-453 (1998).

3. Wiens, J.J., Chippindale, P.T. & Hillis, D.M. When are phylogenetic analyses misled by convergence? A case study in Texas cave salamanders. *Syst Biol* **52**, 501-14 (2003).

4. Kitazoe, Y. et al. Multidimensional vector space representation for convergent evolution and molecular phylogeny. *Mol Biol Evol* **22**, 704-15 (2005).

5. Kornegay, J.R., Schilling, J.W. & Wilson, A.C. Molecular adaptation of a leaf-eating bird: stomach lysozyme of the hoatzin. *Mol Biol Evol* **11**, 921-8 (1994).

6. Stewart, C.B., Schilling, J.W. & Wilson, A.C. Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330**, 401-4 (1987).

7. Stewart, C.B. & Wilson, A.C. Sequence convergence and functional adaptation of stomach lysozymes from foregut fermenters. *Cold Spring Harb Symp Quant Biol* **52**, 891-9 (1987).

8. Zhang, J. & Kumar, S. Detection of convergent and parallel evolution at the amino acid sequence level. *Mol Biol Evol* **14**, 527-36 (1997).

9. Zakon, H.H. Convergent evolution on the molecular level. *Brain Behav Evol* **59**, 250-61 (2002).

10. Zakon, H.H., Lu, Y., Zwickl, D.J. & Hillis, D.M. Sodium channel genes and the evolution of diversity in communication signals of electric fishes: convergent molecular evolution. *Proc Natl Acad Sci U S A* **103**, 3675-80 (2006).

11. Rokas, A. & Carroll, S.B. Frequent and Widespread Parallel Evolution of Protein Sequences. *Mol Biol Evol* (2008).

12. Fortna, A. et al. Lineage-specific gene duplication and loss in human and great ape evolution. *Plos Biology* **2**, 937-954 (2004).

13. Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research* **15**, 1153-1160 (2005).

14. Khaitovich, P. et al. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309**, 1850-1854 (2005).

15. Maddison, W.P. & Knowles, L.L. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* **55**, 21-30 (2006).

16. Pollard, D.A., Iyer, V.N., Moses, A.M. & Eisen, M.B. Widespread discordance of gene trees with species tree in Drosophila: Evidence for incomplete lineage sorting. *Plos Genetics* **2**, 1634-1647 (2006).

17. Wang, X.X., Grus, W.E. & Zhang, J.Z. Gene losses during human origins. *Plos Biology* **4**, 366-377 (2006).

18. Camp, C.L. Classification of the lizards. *Bulletin of the American Museum of Natural History* **48**, 289-435 (1923).

19. Estes, R., De Queiroz, K. & Gauthier, J.A. Phylogenetic relationships within Squamata. in *Phylogenetic Relationships of the Lizard Families, Essays Commemorating Charles L. Camp* (ed. Estes, R.) 119-281 (Stanford University Press, Stanford, 1988).

20. Frost & Etheridge. A phylogenetic analysis and taxonomy of Iguanian lizards (Reptilia: Squamata). *Miscellaneous Publications of the Museum of Natural History, University of Kansas* **81**, 1-65 (1989).

21. Fry, B.G. et al. Early evolution of the venom system in lizards and snakes. *Nature* **439**, 584-588 (2006).

22. Townsend, T.M., Larson, A., Louis, E. & Macey, J.R. Molecular phylogenetics of Squamata: The position of snakes, Amphisbaenians, and Dibamids, and the root of the Squamate tree. *Systematic Biology* **53**, 735-757 (2004).

23. Vidal, N. & Hedges, S.B. The phylogeny of squamate reptiles (lizards, snakes, and amphisbaenians) inferred from nine nuclear protein-coding genes. *Comptes Rendus Biologies* **328**, 1000-1008 (2005).

24. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* **16**, 1114-1116 (1999).

25. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol* **7 Suppl 1**, S4 (2007).

26. Rokas, A., Williams, B.L., King, N. & Carroll, S.B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798-804 (2003).

27. Castoe, T.A., Jiang, Z.J., Gu, W., Wang, Z.O. & Pollock, D.D. Adaptive evolution and functional redesign of core metabolic proteins in snakes. *PLoS ONE* **3**, e2201 (2008).

28. Schierup, M.H. & Hein, J. Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879-91 (2000).

29. Clayton, D.A. Replication of animal mitochondrial DNA. *Cell* **28**, 693-705 (1982).

30. Piganeau, G., Gardner, M. & Eyre-Walker, A. A broad survey of recombination in animal mitochondria. *Molecular Biology and Evolution* **21**, 2319-2325 (2004).

31. Tsaousis, A.D., Martin, D.P., Ladoukakis, E.D., Posada, D. & Zouros, E. Widespread recombination in published animal mtDNA sequences. *Molecular Biology and Evolution* **22**, 925-933 (2005).

32. Lockhart, P.J., Steel, M.A., Hendy, M.D. & Penny, D. Recovering evolutionary trees under a more realistic model of sequence. *Molecular Biology and Evolution* **11**, 605-612 (1994).

33. Douglas, D.A., Janke, A. & Arnason, U. A mitogenomic study on the phylogenetic position of snakes. *Zoologica Scripta* **35**, 545-558 (2006).

34. Kumazawa, Y. Mitochondrial genomes from major lizard families suggest their phylogenetic relationships and ancient radiations. *Gene* **388**, 19-26 (2007).

35. Krishnan, N.M., Seligmann, H., Stewart, C.B., De Koning, A.P. & Pollock, D.D. Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference. *Mol Biol Evol* **21**, 1871-83 (2004).

36. Williams, P.D., Pollock, D.D., Blackburne, B.P. & Goldstein, R.A. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol* **2**, e69 (2006).

37. Yang, Z. Adaptive molecular evolution. in *Handbook of Statistical Genetics* (eds. Balding, D., Bishop, M. & Cannings, C.) 229-254 (Wiley, New York, 2003).

38. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**, 289-300 (1995).

39. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. & Higgins, D.G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**, 4876-4882 (1997).

40. Ronquist, F. & Huelsenbeck, J.P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574 (2003).

41. Swofford, D.L. PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). (Sinauer Associate, Sunderland, Massachusetts, 1997).

42. Yang, Z.H. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**, 555-556 (1997).

FIGURE LEGENDS

Figure 1. Squamate phylogenetic tree. This Bayesian tree was estimated using all 13 mitochondrial protein-coding genes and two nuclear genes. All nodes had 100% posterior probability support, except the three nodes indicated. In contrast to this topology, the agamid lizards are thought to form a group with the iguanid lizards (both in blue), as indicated by the red arrow. Trees based on mitochondrial genes tend to be similar to that shown (the MT topology). In contrast, trees based on nuclear genes place them with the Iguanidae (the NUC topology), in agreement with expectations from morphological studies.

Figure 2. Differences in site-specific likelihood support (ΔSSLS) for the MT and NUC topologies. Positive values of ΔSSLS indicate greater support for the NUC tree, and negative values indicate greater support for the MT tree. ΔSSLS across sites in all mitochondrial protein-coding genes are shown for (A) 2^{nd} codon positions; (B) 3^{rd} codon positions; and (C) four-fold degenerate sites. Values are shown in blue if the ΔSSLS magnitude is less than 0.5, and are shown in red if support levels are greater than 0.5. This highlights strong support levels for one tree or the other.

Figure 3. Relationship between evolutionary rates and site specific support for competing trees. The difference in site likelihood support (ΔSSLS) between the MT and NUC tree is broken down by relative rates of evolution for each of the three codon positions for all protein-coding mitochondrial genes. Slower evolving sites contribute the highest support to the MT tree, whereas a majority of all sites provide moderate support for the NUC tree, regardless of evolutionary rate.

Figure 4. Convergent evolution of protein sequences. The number of convergent and divergent substitutions in all pairs of branches along independent lines of descent were estimated A) using the ML marginal ancestral reconstructions, and B) using a Bayesian approach that calculated the posterior probability of all possible substitutions (see text). The numbers of convergent substitutions were related to the numbers of divergent substitutions using orthogonal regressions (red line; R^2 shown on graph). The snake-

agamid branch-pair is well above the other branch pairs, regardless of the methodology used (red point; panels A and B). The asymptotic calculation of the random expected fraction of convergent substitutions, conditional on the ML parameter estimates from the observed data is shown for reference (blue line, panel B). C) Site-specific posterior probabilities of convergent substitutions between the snake-agamid branch pair for all codon positions using the Bayesian method. Sites with a high probability of having experienced convergent changes (red) are present in all protein-coding genes, but are clustered particularly in COX1 and ND1. D) Sliding window plots of the site-specific likelihood support in favor of the presumed false MT topology (blue) and the regional posterior probability of convergent substitutions (red).

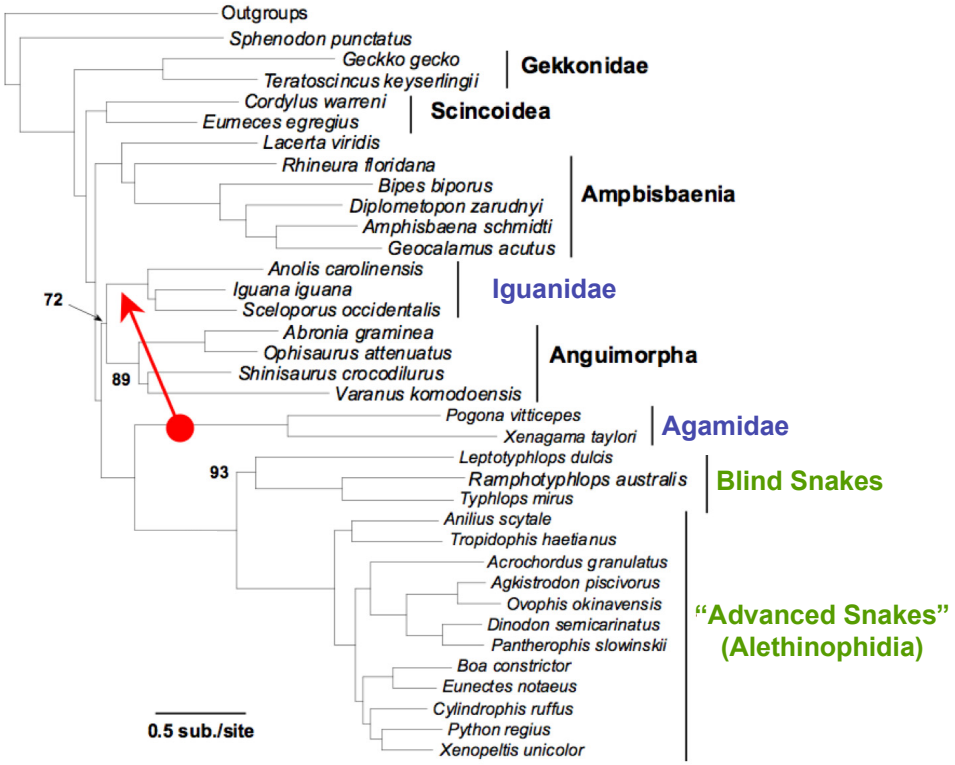
Outgroups
Sphenodon punctatus
Gekko gecko
Teratoscincus keyserlingii
Gekkonidae
Cordylus warreni
Eumeces egregius
Scincoidea
Lacerta viridis
Rhineura floridana
Bipes biporus
Diplometopon zarudnyi
Amphisbaena schmidti
Geocalamus acutus
Ampbisbaenia
Anolis carolinensis
Iguana iguana
Sceloporus occidentalis
Iguanidae
72
Abronia graminea
Ophisaurus attenuatus
Shinisaurus crocodilurus
Varanus komodoensis
Anguimorpha
89
Pogona vitticepes
Xenagama taylori
Agamidae
Leptotyphlops dulcis
93
Ramphotyphlops australis
Typhlops mirus
Blind Snakes
Anilius scytale
Tropidophis haetianus
Acrochordus granulatus
Agkistrodon piscivorus
Ovophis okinavensis
Dinodon semicarinatus
Pantherophis slowinskii
Boa constrictor
Eunectes notaeus
Cylindrophis ruffus
Python regius
Xenopeltis unicolor
"Advanced Snakes"
(Alethinophidia)
0.5 sub./site

Fig. 1

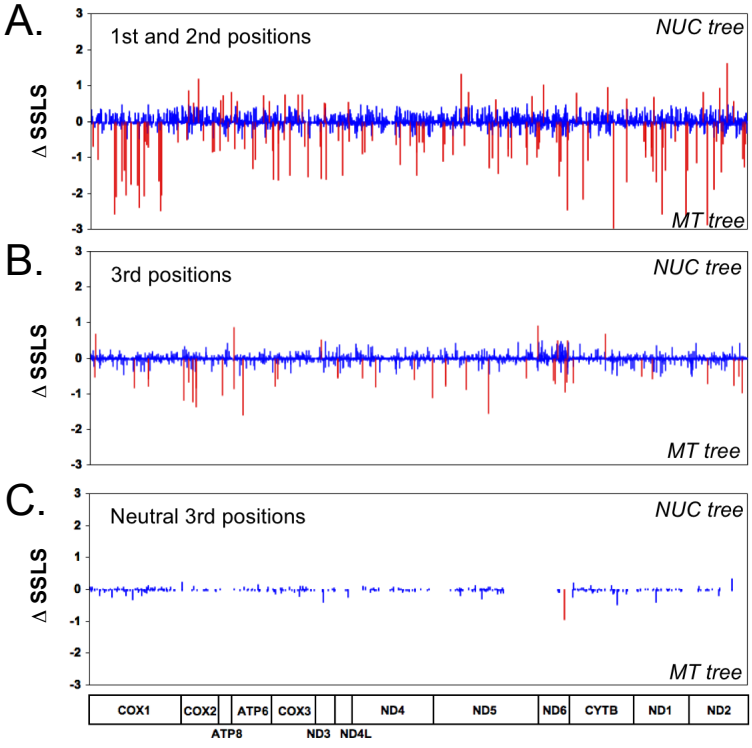
A.
1st and 2nd positions
NUC tree
MT tree
Δ SSLS
B.
3rd positions
NUC tree
MT tree
Δ SSLS
C.
Neutral 3rd positions
NUC tree
MT tree
Δ SSLS
COX1
COX2
ATP8
ATP6
COX3
ND3
ND4L
ND4
ND5
ND6
CYTB
ND1
ND2

Fig. 2

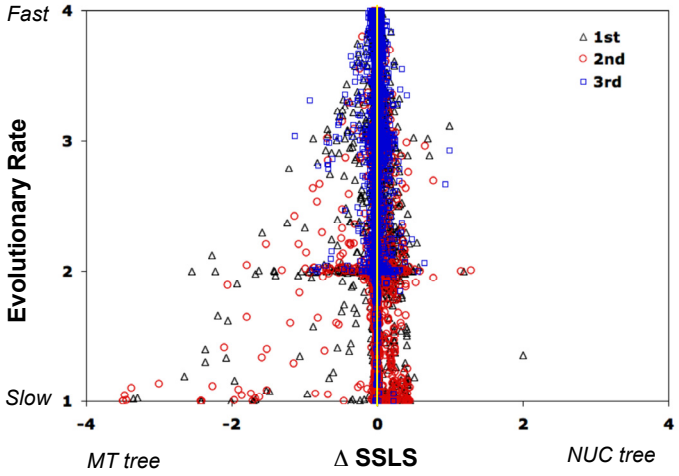
Fast
Slow
Evolutionary Rate
4
3
2
1
-4
-2
0
2
4
MT tree
Δ SSLS
NUC tree
1st
2nd
3rd

Fig. 3

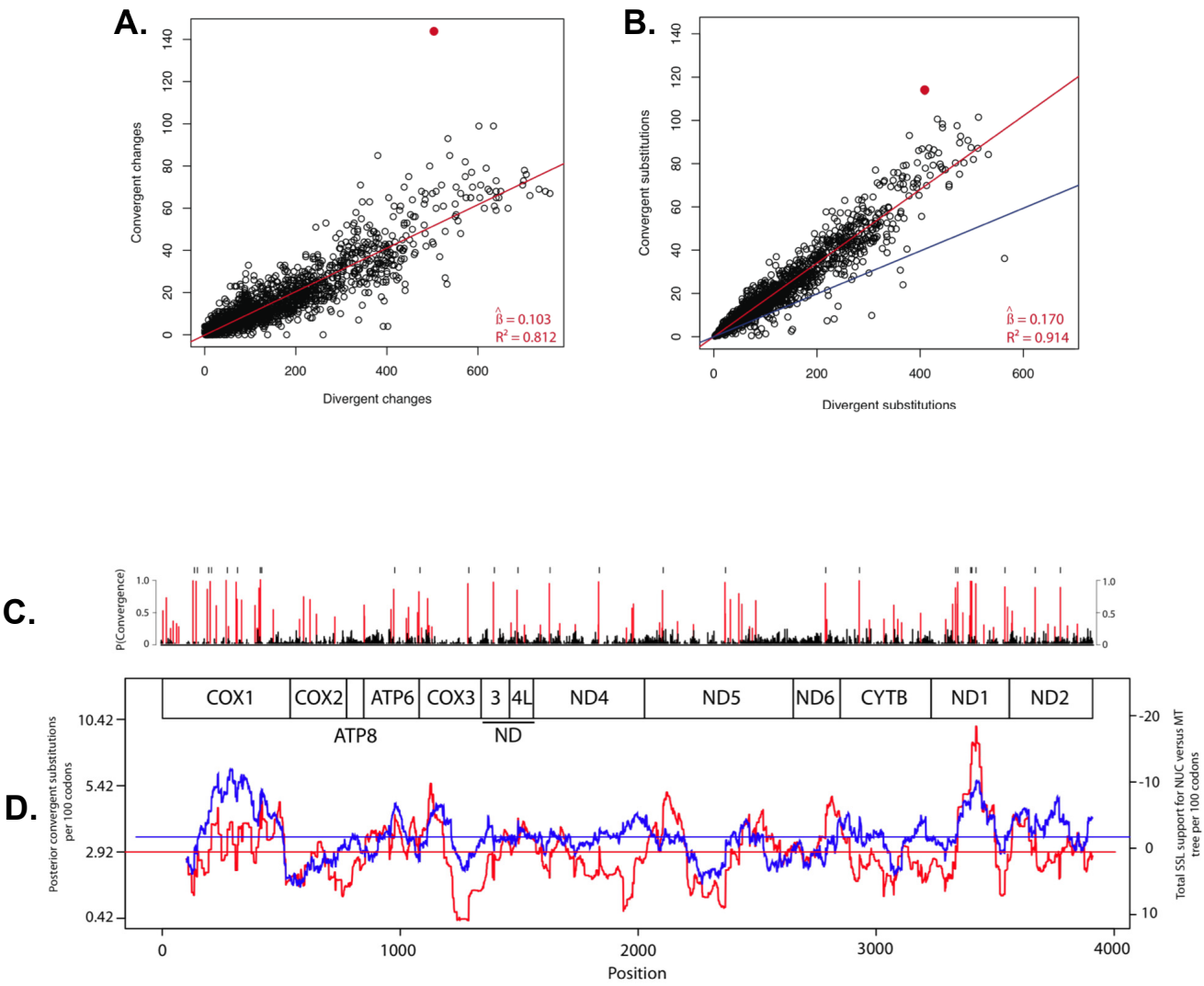
A.
Convergent changes
Divergent changes
$\hat{\beta} = 0.103$
$R^2 = 0.812$
B.
Convergent substitutions
Divergent substitutions
$\hat{\beta} = 0.170$
$R^2 = 0.914$
C.
P(Convergence)
D.
COX1
COX2
ATP6
COX3
3
4L
ND4
ND5
ND6
CYTB
ND1
ND2
ATP8
ND
Posterior convergent substitutions per 100 codons
Total SSL support for NUC versus MT tree per 100 codons
Position

Fig. 4