

1 **An ancient adaptive episode of convergent molecular**
2 **evolution confounds phylogenetic inference**

3

4 Todd A. Castoe^{1,*}, A. P. Jason de Koning^{1,*}, Hyun-Min Kim¹, Wanjun Gu¹, Brice P.
5 Noonan², Zhi J. Jiang³, Christopher L. Parkinson⁴, and David D. Pollock^{1‡}

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

8 ²Department of Biology, University of Mississippi, Box 1848, University, MS 38677
9 USA

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

12 ⁴Department of Biology, University of Central Florida, 4000 Central Florida Blvd.,
13 Orlando, FL 32816USA

14 *The first two authors contributed equally

15

16 ‡*Corresponding Author*: David D. Pollock, Department of Biochemistry and Molecular
17 Genetics, University of Colorado Health Sciences Center, Aurora, CO, 80045 USA.

18 *Email*: David.Pollock@uchsc.edu *phone*: 303-724-3234 *fax*: 303-724-3215

19

20

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

10 Although selection-driven convergent evolution of morphological characters has been
11 identified as a potential source of error for phylogenetic inference¹⁻³, convergence in
12 molecular datasets is believed to be rare. Definitive evidence of convergence at the
13 molecular level is known from only a small number of proteins⁴⁻¹⁰. There have been,
14 however, few searches for convergent molecular evolution in protein sequences, and the
15 true frequency of molecular convergence in nature is therefore unknown; it may be more
16 common than widely believed but difficult to detect, or simply overlooked^{4,8,11}.
17 Regardless of its true frequency in nature, identifying convergent molecular evolution
18 when it happens is important for understanding mechanisms of functional adaptation, and
19 to prevent it from causing errors in phylogenetic inference. For example, significant
20 differences in phylogenies inferred from different genes are usually taken to indicate
21 differences in the evolutionary histories of those genes arising from differential patterns
22 of lineage sorting, hybridization, recombination, horizontal gene transfer, or gene
23 duplication and loss¹²⁻¹⁷. In the presence of convergent evolution, however, the
24 differences between the trees might be artifactual and the bases for the inferences would
25 then be invalid.

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

1 This is a radical result not suggested by previous studies. If true, this relationship would
2 disrupt the monophyly of the Iguania, but it is contradicted by our own nuclear gene
3 analyses (Fig. 1; see also Fig. S1B), previous larger nuclear gene studies^{2,22,23}, and
4 morphological evidence¹⁸⁻²¹.

5 The mitochondrial signal favoring the radical tree is strong enough that the snake-agamid
6 grouping was also supported in combined analysis of the joint mtDNA and nuclear data
7 (Fig. 1), although all other relationships from the combined estimate are in excellent
8 agreement with our nuclear gene trees and previous nuclear gene-based studies^{2,22,23}.
9 Hereafter, we refer to the tree estimated from the joint mitochondrial plus nuclear data
10 (Fig. 1) as the “MT” topology, and the same tree but with a monophyletic Iguania (see
11 red arrow in Fig. 1) as the “NUC” topology. The Shimodaira-Hasagawa (S-H) test²⁴, a
12 standard likelihood-based tree hypothesis testing approach, significantly rejected the
13 NUC in favor of the MT topology for all mitochondrial sequence data together, and for
14 each of the three codon positions separately ($P < 0.01$). Significant rejection of
15 alternative phylogenetic hypotheses based on an S-H test is commonly accepted as
16 conclusive evidence in evolutionary studies. In this case, however, the result is not
17 credible because so many independent data sources support the NUC phylogeny. It must
18 therefore be considered what is wrong with our interpretation of the mitochondrial data.

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

1 evidence for a strong adaptive burst of mitochondrial protein change early in snake
2 evolution²⁷.

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

15 **RESULTS**

16 **Site-specific support for the two topologies**

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

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

1 conceivably have been caused by gene conversion or recombination²⁸, although this is
2 unlikely since mtDNA recombination is thought to be rare within vertebrate species²⁹⁻³¹,
3 let alone between such distantly related lineages as snakes and agamid lizards. This
4 hypothesis is further excluded because site-specific support for each tree is widely
5 dispersed throughout the mitochondrial genome (Fig. 2). Gene conversion or
6 recombination should lead to discrete segments of the genome that strongly support one
7 tree over another, and this is not observed. Some genes, including COX1, COX3, CytB,
8 ND1, and ND2, possess more sites that strongly support the MT tree than do other genes,
9 but they still contain a majority of sites that weakly to moderately support the NUC tree
10 (Figs. 2 and 3; also Figs. S2 and S3).

11 Two remaining possibilities for the conflict in phylogenetic signal are that unusual
12 mutation processes led to reconstruction bias, or that positive or negative selection on
13 amino acids led to unusual substitution patterns that misled phylogenetic inference. An
14 important role for the mutation process is strongly contraindicated by a number of
15 independent lines of evidence. First, nucleotide frequency biases at all sites and at four-
16 fold redundant sites are not particularly similar between snakes and agamids (Fig. S4).
17 Second, log-determinant phylogenetic analyses of the mtDNA, which should reduce
18 sensitivity to base frequency biases³², recover the MT tree (Fig. S5). Third, amino acid
19 sequences and 2nd codon positions should be the least affected by mutation biases, but
20 Bayesian phylogenetic analyses of these data both lead to trees essentially identical to the
21 MT topology (data not shown). Furthermore, site-specific support for the MT tree is less
22 common at 3rd codon positions than at 1st or 2nd positions (Fig. 2; also Figs. S2 and S3).
23 Four-fold redundant 3rd codon positions, which do not alter the amino acid sequence
24 when they change, provide almost no differential likelihood support favoring either tree
25 (Fig. 2C).

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

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

6 **Effects of removing taxa**

7 Bayesian phylogenetic analyses of the mtDNA data with either of the two agamid species
8 excluded (either *Xenagama* or *Pogona*) produced trees highly similar to the original MT
9 tree, with agamid lizards paired with snakes (data not shown). Thus, both species of
10 agamid lizards have phylogenetic affinity to snakes in the mtDNA data. When all snakes
11 were excluded, the agamids clustered with amphisbaenian lizards (Fig. S7), whereas
12 previous mtDNA studies that did not include the agamids found strong support for
13 pairing snakes with amphisbaenians^{33,34}. These major changes in phylogenetic
14 relationships with minor changes in taxon sampling are indicators of phylogenetic
15 conflict and the unreliability of the MT tree, and are unexpected given the large size (>
16 11,700 bp) of the mitochondrial dataset.

17 **Effect of removing sites that strongly support the MT topology**

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

1 **Convergent evolution of amino acid sequences**

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

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

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

1 previous analyses of molecular convergence have utilized ML approaches on branch-
2 pairs of *a priori* interest⁸, the difference between the ML and Bayesian results observed
3 here bring the accuracy of ML approaches into question, since they ignore error in the
4 unknown ancestral states. Previous related analyses have shown that failure to integrate
5 over unknown ancestral states can lead to misleading biological conclusions³⁵⁻³⁷. Since
6 bias caused by conditioning on optimal ancestral state reconstructions is expected in ML,
7 we primarily consider the Bayesian results hereafter.

8 Among all branch pairs compared, the number of convergent events between the
9 branches leading to the most recent common ancestors (MRCAs) of snakes and of
10 agamid lizards stands out as being far greater than expected based on the number of
11 divergence events. There were 28 positions in the protein alignments with more than 80%
12 posterior probability of convergent substitution between these two branches. These sites
13 were concentrated in COX1 and ND1, but were present in other proteins as well (Fig.
14 4C). Remarkably, these two branches of *a priori* interest showed the single greatest
15 excess of convergence of all branch-pairs on the tree (0.28 convergent substitutions per
16 divergent substitution, or 1.6 times the empirically determined random (neutral)
17 expectation; Fig. 4A, B). Partial correlation analyses indicated that the extreme excess of
18 convergence between the snake and agamid branch pair cannot be explained by long
19 branch lengths, which is a conceivable source of bias in posterior convergence estimation
20 (see Supplementary Data).

21 Using empirically predicted levels of convergence from the orthogonal regressions, a
22 series of binomial tests identified this pair of branches as the only pair with a highly
23 significant probability of excess convergence ($P < 0.001$, after accounting for false
24 discovery³⁸). For branch pairs with higher predicted false discovery rates, the expected
25 number of true positives (Fig. S10) is high enough that a further 11 branch-pairs may
26 have experienced an excess of convergence events, although as many as five of these
27 additional branch-pairs are expected to be false positives (Fig. S11).

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

1 the MT topology coincide with peaks in the probability of convergent substitution (Fig.
2 4C). The highly significant correlation ($r = -0.498$, $P < 2.2 \times 10^{-16}$) indicates that sites
3 supporting the incorrect MT tree are likely to represent amino acid convergence. The link
4 between convergent amino acids and phylogenetic error is supported by the observation
5 that removing the 98 codons with the highest probability of convergence (the top 2.5%;
6 see Fig. 4C and site patterns in Fig. S12) was sufficient to cause total likelihood support
7 to switch from favoring the MT topology (total $\Delta\text{SSLS} = -85.77$) to favoring the NUC
8 topology (total $\Delta\text{SSLS} = 2.577$). Removal of these sites brought the cumulative Bayesian
9 posterior number of predicted convergent substitutions down from 113.5 to 58.8, a
10 number statistically indistinguishable from the 69.6 convergent substitutions predicted
11 from the empirical regression on divergent substitutions (Fig. 4A; $P > 0.09$). It therefore
12 appears that removing the excess convergent substitutions allows the correct underlying
13 signal from the majority of the mitochondrial genome to dominate.

14 **DISCUSSION**

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

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

1 proteins have endured the most extreme burst of apparently adaptive protein evolution yet
2 observed in vertebrate mitochondria²⁷; this is consistent with the idea that the excess
3 convergence levels observed here are due to the action of natural selection rather than
4 random homoplasy. It was proposed that the evolutionary burst in snakes may have been
5 driven by selection related to physiological adaptations for metabolic efficiency and to
6 allow radical fluctuations in aerobic metabolic rate²⁷. The molecular convergence
7 between snakes and agamid lizards may thus have resulted from shared adaptive
8 pressures on metabolic function. Since the convergence extends across most regions of
9 the mitochondrial genome, any common adaptive force must have been exceptionally
10 strong and broad in scope.

11 Perhaps the most disturbing aspect of the idea that adaptation may be at the root of this
12 molecular convergence event is the systematic nature of its effects on phylogenetic
13 reconstruction. If such a large convergence event can occur, it is reasonable to suppose
14 that smaller events may be much more common than realized, but are often difficult to
15 detect, or overlooked. A convergence event even a fraction of this magnitude could easily
16 disrupt many topology estimates because of the relative biasing power of each convergent
17 site. Indeed, the first known case of convergent molecular evolution in ruminant
18 lysozymes^{6,7} was shown to lend support to a dramatically wrong phylogeny that placed
19 cows within the primates. This occurred even though only a small number of convergent
20 substitutions apparently took place⁸. In the present case, a small fraction of convergent
21 sites dramatically outweighed the accurate signal at hundreds of other sites (e.g., Figs. 2
22 and 3). Existing evolutionary models assume that convergent molecular evolution is
23 extremely improbable, and thus even small amounts of convergence can be falsely
24 interpreted as extremely strong evidence for an incorrect topology.

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

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

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

10 **METHODS**

11 **Mitochondrial genome sequencing, alignment and phylogeny inference.**

12 Mitochondrial genomes were sequenced and annotated for two snake species, *Anilius*
13 *scytale* and *Tropidophis haetianus*, to increase sampling at the base of snake phylogeny
14 (see Supplementary Methods). All 13 mitochondrial protein-coding genes (~11,700 bp)
15 from complete mitochondrial genomes of squamates available at the time of study, plus
16 the two new species, were aligned using ClustalX³⁹ based on their amino acid translation;
17 multiple species per genus were excluded (Table S1). Representatives of major tetrapod
18 lineages were also included to root the squamate tree. Nucleotide sequences of two
19 nuclear genes, *rag-1* and *c-mos*, were obtained from GenBank and aligned for
20 comparison to the mitochondrial data (Table S2).

21 For phylogenetic analysis, mitochondrial and nuclear datasets were partitioned by gene
22 and codon position and appropriate partition-specific models were selected
23 (Supplementary Methods). Bayesian phylogenetic trees were estimated in MrBayes
24 3.0b4⁴⁰ with partitioned models for mitochondrial and nuclear, both independently and
25 combined.

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

1 mitochondrial dataset were used to evaluate phylogenetic hypotheses in PAUP* 4.0b10⁴¹
2 (see Supplementary Methods); where relevant, *P*-values less than 0.05 were considered
3 significant. Evidence for non-stationary base frequencies across lineages was evaluated
4 based on chi-squared tests in PAUP*. Support for alternative topologies was evaluated
5 using the Shimodaira-Hasegawa test²⁴. Site-specific likelihood support (SSLS) was
6 estimated using ML and a GTR+ Γ +I model (general time-reversible with gamma-
7 distributed and invariant rates among sites) per codon position.

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

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

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

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

12 **ACKNOWLEDGMENTS**

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

16 **AUTHOR CONTRIBUTIONS**

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

1 **REFERENCES**

- 2 1. Harmon, L.J., Kolbe, J.J., Cheverud, J.M. & Losos, J.B.
3 Convergence and the multidimensional niche. *Evolution* **59**, 409-
4 21 (2005).
- 5 2. Lee, M.S.Y. Convergent evolution and character correlation in
6 burrowing reptiles: towards a resolution of squamate relationships.
7 *Biological Journal of the Linnean Society* **65**, 369-453 (1998).
- 8 3. Wiens, J.J., Chippindale, P.T. & Hillis, D.M. When are
9 phylogenetic analyses misled by convergence? A case study in
10 Texas cave salamanders. *Syst Biol* **52**, 501-14 (2003).
- 11 4. Kitazoe, Y. et al. Multidimensional vector space representation for
12 convergent evolution and molecular phylogeny. *Mol Biol Evol* **22**,
13 704-15 (2005).
- 14 5. Kornegay, J.R., Schilling, J.W. & Wilson, A.C. Molecular
15 adaptation of a leaf-eating bird: stomach lysozyme of the hoatzin.
16 *Mol Biol Evol* **11**, 921-8 (1994).
- 17 6. Stewart, C.B., Schilling, J.W. & Wilson, A.C. Adaptive evolution
18 in the stomach lysozymes of foregut fermenters. *Nature* **330**, 401-4
19 (1987).
- 20 7. Stewart, C.B. & Wilson, A.C. Sequence convergence and
21 functional adaptation of stomach lysozymes from foregut
22 fermenters. *Cold Spring Harb Symp Quant Biol* **52**, 891-9 (1987).
- 23 8. Zhang, J. & Kumar, S. Detection of convergent and parallel
24 evolution at the amino acid sequence level. *Mol Biol Evol* **14**, 527-
25 36 (1997).
- 26 9. Zakon, H.H. Convergent evolution on the molecular level. *Brain*
27 *Behav Evol* **59**, 250-61 (2002).
- 28 10. Zakon, H.H., Lu, Y., Zwickl, D.J. & Hillis, D.M. Sodium channel
29 genes and the evolution of diversity in communication signals of
30 electric fishes: convergent molecular evolution. *Proc Natl Acad Sci*
31 *U S A* **103**, 3675-80 (2006).
- 32 11. Rokas, A. & Carroll, S.B. Frequent and Widespread Parallel
33 Evolution of Protein Sequences. *Mol Biol Evol* (2008).
- 34 12. Fortna, A. et al. Lineage-specific gene duplication and loss in
35 human and great ape evolution. *Plos Biology* **2**, 937-954 (2004).
- 36 13. Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C. & Cristianini, N.
37 Estimating the tempo and mode of gene family evolution from
38 comparative genomic data. *Genome Research* **15**, 1153-1160
39 (2005).

- 1 14. Khaitovich, P. et al. Parallel patterns of evolution in the genomes
2 and transcriptomes of humans and chimpanzees. *Science* **309**,
3 1850-1854 (2005).
- 4 15. Maddison, W.P. & Knowles, L.L. Inferring phylogeny despite
5 incomplete lineage sorting. *Systematic Biology* **55**, 21-30 (2006).
- 6 16. Pollard, D.A., Iyer, V.N., Moses, A.M. & Eisen, M.B. Widespread
7 discordance of gene trees with species tree in *Drosophila*:
8 Evidence for incomplete lineage sorting. *Plos Genetics* **2**, 1634-
9 1647 (2006).
- 10 17. Wang, X.X., Grus, W.E. & Zhang, J.Z. Gene losses during human
11 origins. *Plos Biology* **4**, 366-377 (2006).
- 12 18. Camp, C.L. Classification of the lizards. *Bulletin of the American*
13 *Museum of Natural History* **48**, 289-435 (1923).
- 14 19. Estes, R., De Queiroz, K. & Gauthier, J.A. Phylogenetic
15 relationships within Squamata. in *Phylogenetic Relationships of*
16 *the Lizard Families, Essays Commemorating Charles L. Camp* (ed.
17 Estes, R.) 119-281 (Stanford University Press, Stanford, 1988).
- 18 20. Frost & Etheridge. A phylogenetic analysis and taxonomy of
19 Iguanian lizards (Reptilia: Squamata). *Miscellaneous Publications*
20 *of the Museum of Natural History, University of Kansas* **81**, 1-65
21 (1989).
- 22 21. Fry, B.G. et al. Early evolution of the venom system in lizards and
23 snakes. *Nature* **439**, 584-588 (2006).
- 24 22. Townsend, T.M., Larson, A., Louis, E. & Macey, J.R. Molecular
25 phylogenetics of Squamata: The position of snakes,
26 Amphisbaenians, and Dibamids, and the root of the Squamate tree.
27 *Systematic Biology* **53**, 735-757 (2004).
- 28 23. Vidal, N. & Hedges, S.B. The phylogeny of squamate reptiles
29 (lizards, snakes, and amphisbaenians) inferred from nine nuclear
30 protein-coding genes. *Comptes Rendus Biologies* **328**, 1000-1008
31 (2005).
- 32 24. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-
33 likelihoods with applications to phylogenetic inference. *Molecular*
34 *Biology and Evolution* **16**, 1114-1116 (1999).
- 35 25. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-
36 branch attraction artefacts in the animal phylogeny using a site-
37 heterogeneous model. *BMC Evol Biol* **7 Suppl 1**, S4 (2007).
- 38 26. Rokas, A., Williams, B.L., King, N. & Carroll, S.B. Genome-scale
39 approaches to resolving incongruence in molecular phylogenies.
40 *Nature* **425**, 798-804 (2003).

- 1 27. Castoe, T.A., Jiang, Z.J., Gu, W., Wang, Z.O. & Pollock, D.D.
2 Adaptive evolution and functional redesign of core metabolic
3 proteins in snakes. *PLoS ONE* **3**, e2201 (2008).
- 4 28. Schierup, M.H. & Hein, J. Consequences of recombination on
5 traditional phylogenetic analysis. *Genetics* **156**, 879-91 (2000).
- 6 29. Clayton, D.A. Replication of animal mitochondrial DNA. *Cell* **28**,
7 693-705 (1982).
- 8 30. Piganeau, G., Gardner, M. & Eyre-Walker, A. A broad survey of
9 recombination in animal mitochondria. *Molecular Biology and*
10 *Evolution* **21**, 2319-2325 (2004).
- 11 31. Tsaousis, A.D., Martin, D.P., Ladoukakis, E.D., Posada, D. &
12 Zouros, E. Widespread recombination in published animal mtDNA
13 sequences. *Molecular Biology and Evolution* **22**, 925-933 (2005).
- 14 32. Lockhart, P.J., Steel, M.A., Hendy, M.D. & Penny, D. Recovering
15 evolutionary trees under a more realistic model of sequence.
16 *Molecular Biology and Evolution* **11**, 605-612 (1994).
- 17 33. Douglas, D.A., Janke, A. & Arnason, U. A mitogenomic study on
18 the phylogenetic position of snakes. *Zoologica Scripta* **35**, 545-558
19 (2006).
- 20 34. Kumazawa, Y. Mitochondrial genomes from major lizard families
21 suggest their phylogenetic relationships and ancient radiations.
22 *Gene* **388**, 19-26 (2007).
- 23 35. Krishnan, N.M., Seligmann, H., Stewart, C.B., De Koning, A.P. &
24 Pollock, D.D. Ancestral sequence reconstruction in primate
25 mitochondrial DNA: compositional bias and effect on functional
26 inference. *Mol Biol Evol* **21**, 1871-83 (2004).
- 27 36. Williams, P.D., Pollock, D.D., Blackburne, B.P. & Goldstein, R.A.
28 Assessing the accuracy of ancestral protein reconstruction
29 methods. *PLoS Comput Biol* **2**, e69 (2006).
- 30 37. Yang, Z. Adaptive molecular evolution. in *Handbook of Statistical*
31 *Genetics* (eds. Balding, D., Bishop, M. & Cannings, C.) 229-254
32 (Wiley, New York, 2003).
- 33 38. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate
34 - a practical and powerful approach to multiple testing. *Journal of*
35 *the Royal Statistical Society Series B-Methodological* **57**, 289-300
36 (1995).
- 37 39. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. &
38 Higgins, D.G. The CLUSTAL_X windows interface: flexible
39 strategies for multiple sequence alignment aided by quality
40 analysis tools. *Nucleic Acids Research* **25**, 4876-4882 (1997).

- 1 40. Ronquist, F. & Huelsenbeck, J.P. MrBayes 3: Bayesian
2 phylogenetic inference under mixed models. *Bioinformatics* **19**,
3 1572-1574 (2003).
- 4 41. Swofford, D.L. PAUP*. Phylogenetic Analysis Using Parsimony
5 (* and Other Methods). (Sinauer Associate, Sunderland,
6 Massachusetts, 1997).
- 7 42. Yang, Z.H. PAML: a program package for phylogenetic analysis
8 by maximum likelihood. *Computer Applications in the Biosciences*
9 **13**, 555-556 (1997).
- 10
- 11
- 12

1 **FIGURE LEGENDS**

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

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

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

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

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

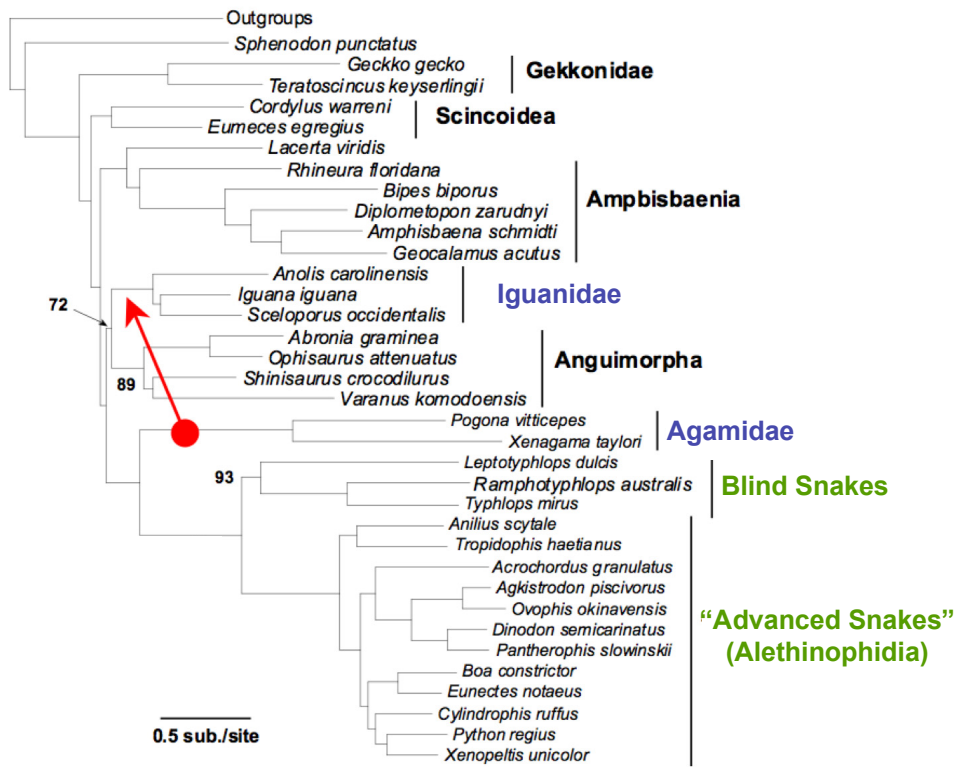


Fig. 1

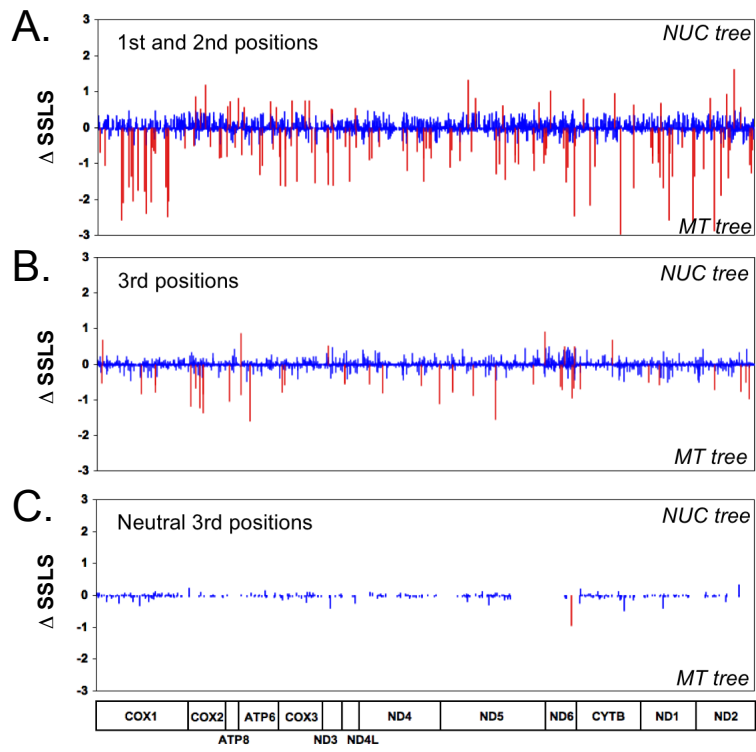


Fig. 2

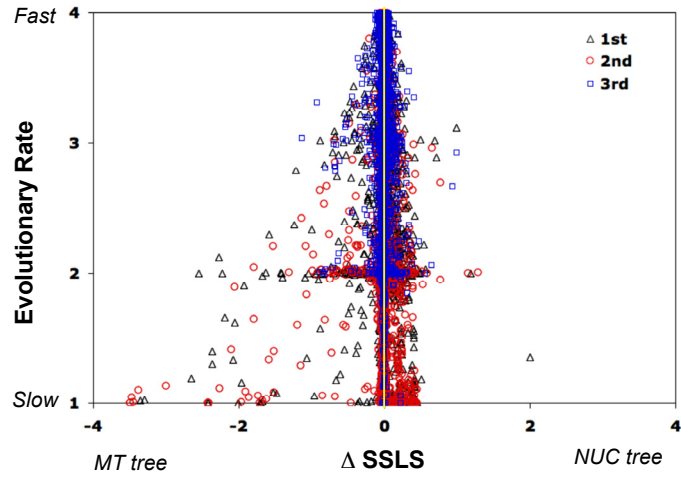


Fig. 3

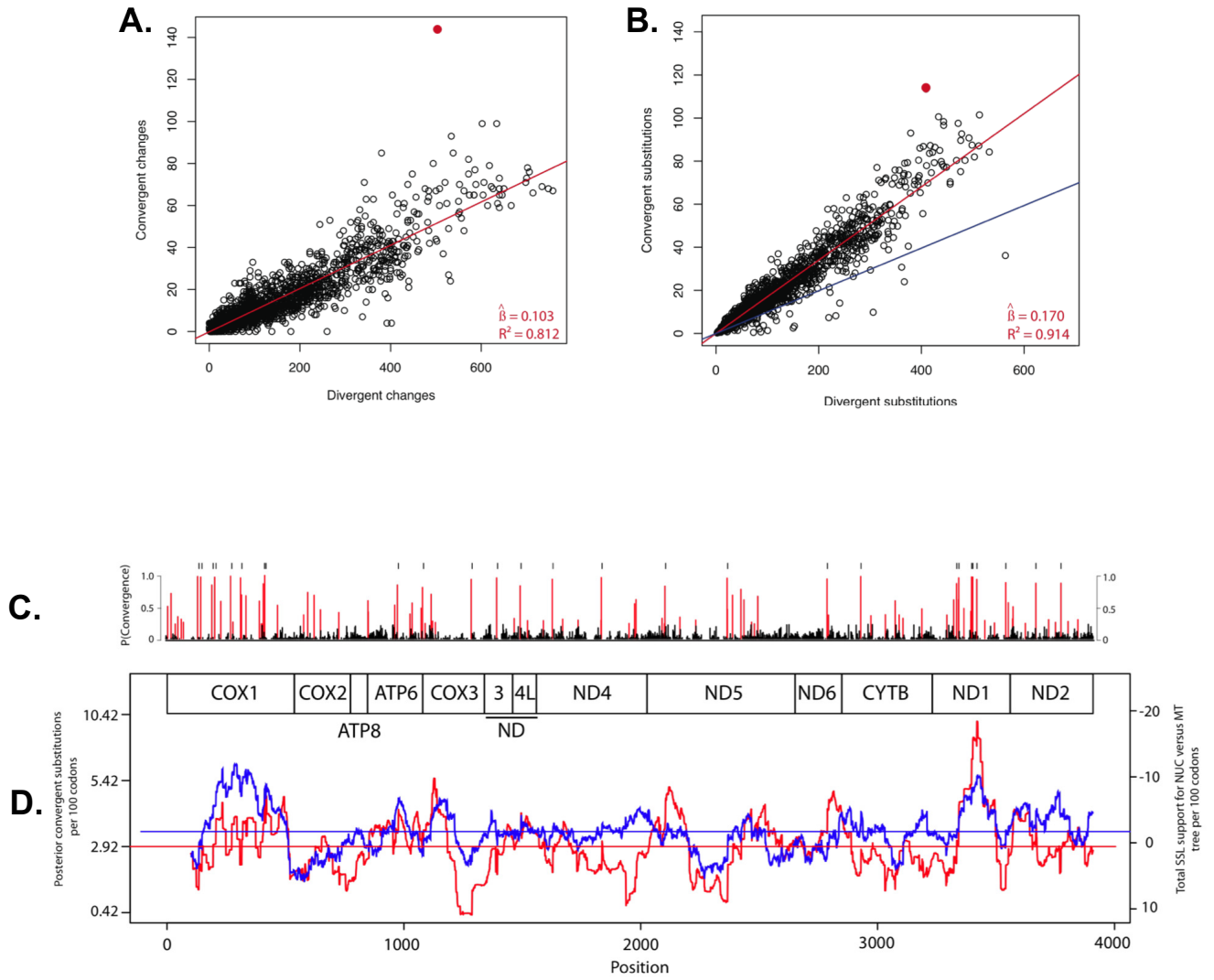


Fig. 4