Evidence for an ancient adaptive episode of convergent molecular evolution

Todd A. Castoe[a,1], A. P. Jason de Koning[a,1], Hyun-Min Kim[a], Wanjun Gu[a,2], Brice P. Noonan[b], Gavin Naylor[c], Zhi J. Jiang[d,e], Christopher L. Parkinson[d,e], and David D. Pollock[a,3]

[a]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045; [b]Department of Biology, University of Mississippi, Box 1848, University, MS 38677; [c]Department of Scientific Computing, Dirac Science Library, Florida State University, Tallahassee, FL 32306; [d]Center for Computational Science, University of Miami, 1120 Northwest 14th Street, Miami, FL 33136; and [e]Department of Biology, University of Central Florida, 4000 Central Florida Boulevard, Orlando, FL 32816

Edited by David M. Hillis, University of Texas, Austin, TX, and approved March 20, 2009 (received for review January 9, 2009)

Documented cases of convergent molecular evolution due to selection are fairly unusual, and examples to date have involved only a few amino acid positions. However, because convergence mimics shared ancestry and is not accommodated by current phylogenetic methods, it can strongly mislead phylogenetic inference when it does occur. Here, we present a case of extensive convergent molecular evolution between snake and agamid lizard mitochondrial genomes that overcomes an otherwise strong phylogenetic signal. Evidence from morphology, nuclear genes, and most sites in the mitochondrial genome support one phylogenetic tree, but a subset of mostly amino acid-altering substitutions (primarily at the first and second codon positions) across multiple mitochondrial genes strongly supports a radically different phylogeny. The relevant sites generally evolved slowly but converged between ancient lineages of snakes and agamids. We estimate that $\approx$44 of 113 predicted convergent changes distributed across all 13 mitochondrial protein-coding genes are expected to have arisen from nonneutral causes—a remarkably large number. Combined with strong previous evidence for adaptive evolution in snake mitochondrial proteins, it is likely that much of this convergent evolution was driven by adaptation. These results indicate that nonneutral convergent molecular evolution in mitochondria can occur at a scale and intensity far beyond what has been documented previously, and they highlight the vulnerability of standard phylogenetic methods to the presence of nonneutral convergent sequence evolution.

adaptation | convergence | phylogenetics | reptile

Convergent molecular evolution, sometimes referred to as homoplasy, can result from neutral processes or from nonneutral processes such as positive selection and adaptation (1–3). Although nonneutral convergent evolution of morphological characters has been frequently pointed to as a source of potential bias in phylogenetic inference (4–6), nonneutral convergence at the molecular-genetic level is believed to be relatively rare and limited in scope. Individual proteins placed under similar selective pressures have been observed, however, to respond with convergent molecular changes (7–12). Coordinated multigene nonneutral molecular convergence events have also been shown to occur in laboratory selection experiments (1), but have not been observed in nature. There have been few systematic screens for nonneutral molecular convergence, however, so its true frequency in nature remains largely unknown. It may be more common than widely believed but difficult to detect, or it may simply have been overlooked (11–14).

Regardless of the frequency of nonneutral convergence in nature, it is important to identify such cases to understand their impact on phylogenetic inference and to illuminate the mechanisms of functional adaptation at the molecular level. Recent convergence research has focused on identifying and avoiding the phylogenetic impacts of neutral convergence. Because both neutral and nonneutral convergence can potentially mislead phylogenetic inference in different ways, however, it is important to distinguish the evidence for both.

While analyzing complete vertebrate mitochondrial genome sequences, we discovered that the mitochondrial protein-coding genes provided strong support for phylogenetic relationships among the squamate reptiles (lizards and snakes) that were incongruent with evidence from numerous previous studies. This surprising finding led us to explore the site-specific patterns of phylogenetic signal contained in the mitochondrial data, to identify which sites provided support for the contrasting topologies, and to examine the plausibility of both neutral and nonneutral explanations for the aberrant phylogenetic signal in the mitochondrion. We examined whether the results could be explained by neutral convergence or were instead due to potentially adaptive nonneutral convergent evolution.

Results

Comparing Phylogenies from Mitochondrial and Nuclear Genes. Molecular phylogenetic estimates among 34 squamate reptile species plus 6 tetrapod outgroup species were examined based on 2 nucleotide datasets: a nuclear dataset with 2 nuclear genes (3,411 bp) and a mitochondrial dataset including all 13 protein-coding mitochondrial genes (11,727 bp). There is broad consensus that, within the squamate reptiles, the iguanas, chameleons, and agamid lizards form an exclusive clade, Iguania (5, 15–19). Analyses of the mitochondrial data, however, provided strong support for a close "sister" relationship between agamid lizards and snakes (Fig. 1; see also Fig. S1*A*). This is a radical result which, if true, would undermine the monophyly of the Iguania, and contradict our own nuclear gene analyses (Fig. 1; see also Fig. S1*B*), previous and even larger nuclear gene studies (5, 18, 19), and morphological evidence (15–17).

The mitochondrial signal favoring the radical tree is strong enough that the snake–agamid grouping was also supported in combined analysis of the 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 (5, 18, 19). Hereafter, we refer to the

Author contributions: T.A.C. and D.D.P. designed research; T.A.C., A.P.J.d.K., H.-M.K., W.G., Z.J.J. performed research; T.A.C., A.P.J.d.K., C.L.P., and D.D.P. contributed new reagents/analytic tools; T.A.C., A.P.J.d.K., H.-M.K., W.G., B.P.N., and D.D.P. analyzed data; and T.A.C., A.P.J.d.K., B.P.N., G.N., and D.D.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. FJ755180 and FJ755181).

See Commentary on page 8799.

[1]T.A.C. and A.P.J.d.K. contributed equally to this work.

[2]Present address: Key Laboratory of Child Development and Learning Science, Southeast University, Ministry of Education, Nanjing 210096, People's Republic of China.

[3]To whom correspondence should be addressed. E-mail: david.pollock@uchsc.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0900233106/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0900233106

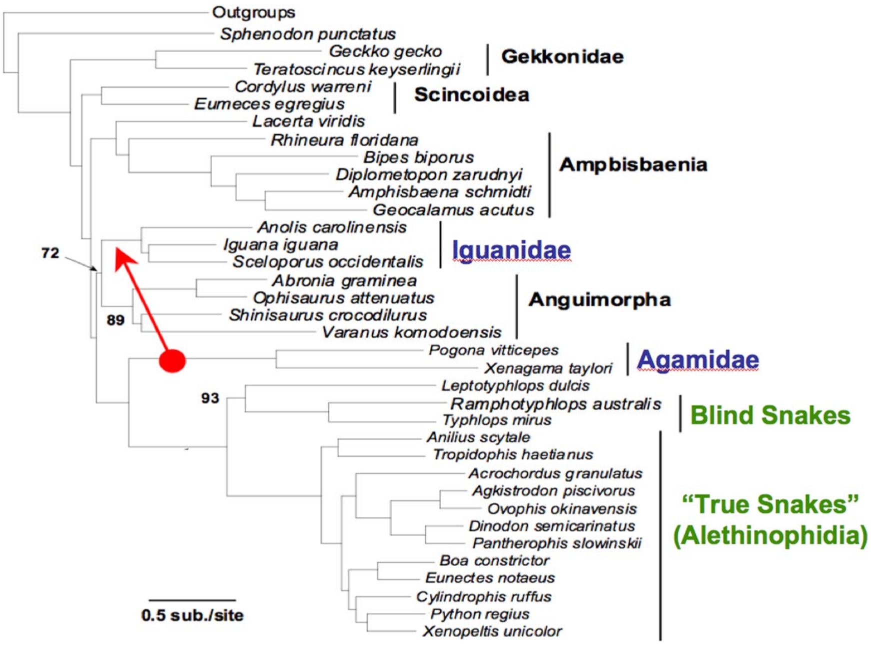

Fig. 1. Squamate reptile phylogenetic tree. This Bayesian tree was estimated by using all 13 mitochondrial protein-coding genes and 2 nuclear genes. All nodes had 100% posterior probability support, except the 3 nodes indicated. In contrast to this topology, the agamid lizards are thought to form a group 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.

tree estimated from the combined mitochondrial plus nuclear data (Fig. 1) as the "MT" topology (because it contains the snake and agamid clade of the mitochondrial tree), and the same tree but with a monophyletic Iguania as the "NUC" topology (because a monophyletic Iguania is the arrangement in the nuclear tree). The Shimodaira–Hasagawa (S-H) test (20), 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 3 codon positions separately ($P < 0.01$). Significant rejection of alternative phylogenetic hypotheses based on an S-H test is commonly accepted as conclusive. In this case, however, we questioned this result because so many independent data sources support the NUC topology.

Site-Specific Support for the 2 Topologies. To identify which nucleotide positions supported the radical MT tree, we measured the difference in site-specific log likelihood values for each of the 2 alternative topologies (ΔSSLS) across the mtDNA dataset. Most 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*A*). If only sites with strong support ($|\Delta SSLS| > 0.5$) are considered, however, the situation is reversed and approximately 5 to 9 times more sites, depending on codon position, strongly favor the MT tree over the NUC tree (Fig. S2*B*; see also Fig. 2; NUC/MT sites = 19/130).

One potential explanation for the conflict in signal is that different sites genuinely have different phylogenetic histories. Such a situation could conceivably have been caused by gene conversion or recombination, but this hypothesis appears unlikely 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. 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 (Fig. 2; see also Fig. S2).

Two other 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 several lines of evidence. First, there is no indication that nucleotide frequency patterns common to both agamids and snakes would tend to falsely cluster these lineages; snakes have extremely variable nucleotide frequencies that encompass the range of most other lizards in the dataset (Fig. S3). Second, log-determinant phylogenetic analyses of the mtDNA, which should reduce sensitivity to base frequency biases (21), recover the MT tree (Fig. S4). Third, amino acid sequences and second codon positions should be the least affected by mutation biases, but Bayesian phylogenetic analyses of these data both lead to trees nearly identical to the MT topology. Furthermore, site-specific support for the MT tree is less common at third codon positions than at first or second positions (Fig. 2; see also Fig. S2). Four-fold redundant third codon positions, which do not alter the amino acid sequence when they change, provide almost no differential likelihood support between the 2 topologies (Fig. 2*D*). Clearly, therefore, an amino acid based explanation of the aberrant phylogenetic signal is favored over a mutational explanation.

It has been recently pointed out (3) that failure to account for heterogeneity in functional constraint among sites can exacerbate the tendency of long branches to incorrectly cluster together in phylogenetic reconstruction (long branch attraction; LBA). An important prediction of LBA often used to diagnose its presence (2) is that fast sites will show greater neutral convergence and will have greater support for an artefactual topology than will slow sites. Contrary to this prediction, the probability that a site strongly supports the MT topology is inversely related to the relative rate of evolution at that site (Fig. 2*E*). Slowly evolving second codon position (nonsynonymous) sites, especially transversions (Fig. S5), strongly support the MT topology, whereas fast-evolving sites and synonymous sites contribute negligible support (Fig. 2). Thus, the pattern observed in this case is opposite the predictions of neutral convergence.

An amino acid mixture model that accounts for variation in fitness profiles across positions ("CAT") has been shown to be effective in overcoming LBA, partially because it better predicts the probability of neutral convergence at functionally constrained

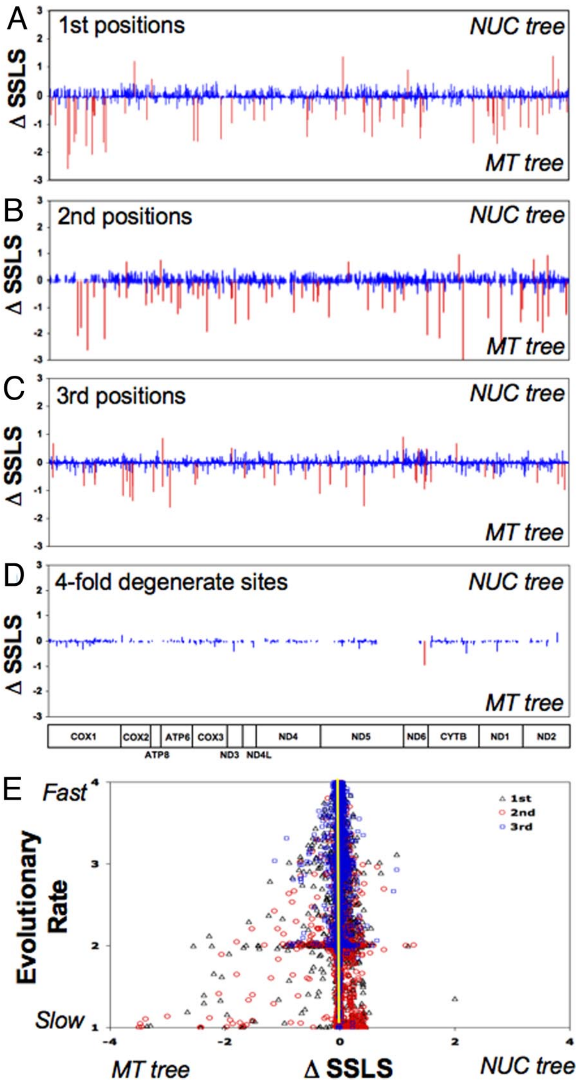

Fig. 2. Differences in site-specific likelihood support (ΔSSLS) for the MT and NUC topologies. Positive values of ΔSSLS indicate greater support for the NUC tree, and negative values indicate greater support for the MT tree. ΔSSLS across sites in all mitochondrial protein-coding genes are shown for (*A*) first codon positions, (*B*) second codon positions, (*C*) third codon positions, and (*D*) 4-fold degenerate sites. Values are shown in blue if the ΔSSLS magnitude is <0.5 and are shown in red if support levels are >0.5. This highlights strong support levels for one tree or the other. (*E*) ΔSSLS between the MT and NUC tree is broken down by relative rates of evolution for each of the 3 codon positions for all protein-coding mitochondrial genes.

positions (3). Analysis of the mitochondrial dataset with this model, however, returned the same radical MT topology as other Bayesian inference methods, with 100% posterior support at most nodes, including the agamid plus snake clade (Fig. S6*A*). Although it is possible that some other model not yet considered could recover the NUC topology, the failure of perhaps the best current model, the site-inhomogeneous CAT model, makes this possibility appear unlikely. Together, these results strongly support the idea that neutral convergence is not responsible for the radical phylogenetic signal in the mitochondrial data.

Convergent Evolution of Amino Acid Sequences. Given the previous results, it seems likely that support for the MT topology could be due to nonneutral convergent amino acid evolution between snakes and agamid lizards. To examine this possibility, we used maximum likelihood (ML) and Bayesian posterior approaches to estimate convergent and divergent substitutions between all pairs of independent lineages (Fig. 3 *A* and *B*), and compared their prevalence along the mitochondrial genome to site-specific support for the MT versus NUC topologies using a sliding window plot. Here, convergent changes between branches are defined as changes at the same site along both branches resulting in the same amino acid, whereas divergent changes result in different amino acids. Peaks in site-specific support for the MT topology tend to coincide with peaks in the probability of convergent substitution between the snake and agamid stem lineages (Fig. 3 *C* and *D*). Furthermore, there is a highly significant correlation ($r = -0.498, P < 2.2 \times 10^{-16}$) between the probability of convergence and likelihood support for the MT topology.

Considering all pairwise branch comparisons, 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. 3*A*) and Bayesian methods ($R^2 = 0.914$, $b = 0.17$; Fig. 3*B*). The tightness of this relationship suggests that most convergent substitutions on the tree were neutral, since they are so well predicted by the number of divergent changes. We also observed substantial differences between the estimates of convergent and divergent changes from the ML and Bayesian analyses (Fig. 3 *A* and *B*). These differences bring the accuracy of the ML results into question, since the ML reconstructions ignore error in the unknown ancestral states (22–24). We therefore primarily consider the Bayesian results hereafter.

Observed levels of convergence throughout the dataset were higher than expected (Fig. 3*B*, slope of blue versus red lines), based on standard models of protein evolution (MtRev+Γ and JTT+Γ). This difference between observed and model-based expectations of convergence is likely due to the failure of traditional models to predict the increased probability of convergence at sites under negative selection (3). Thus, most branches have more neutral convergence than expected based on standard models. We therefore focused on the observed relationship between convergence and divergence as a baseline for neutral expectations. Because convergence and divergence estimates are based on posterior distributions of ancestral states and substitutions, this relationship should be relatively robust to model violations (22).

Among all branch pairs examined, 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 divergent substitutions (Fig. 3*B*). We estimate that 113 convergent changes are distributed across all 13 mitochondrial protein-coding genes; this is ≈44 more convergent changes (presumably nonneutral) than predicted by empirically observed convergence levels between other branches over the entire tree, and ≈73 more than predicted by evolutionary model-based expectations. There were 28 readily identifiable convergent sites (>80% posterior probability of convergence between these 2 branches); these were concentrated in COX1 and ND1, but were present in other proteins as well (Fig. 3*C*). These 2 branches of 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 expectation from the orthogonal regression of divergence and convergence; Fig. 3 *A* and *B*). Using predicted levels of convergence from the orthogonal regression, 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; ref. 25). This analysis thus indicates that the amount of convergence between these 2 branches is exceptional and likely to include a major nonneutral component.

While the statistical evidence for excess, probably nonneutral, convergence between the snake–agamid branch pair was particularly strong, additional branch pairs may have experienced nonneutral convergence as well. Allowing different false-discovery rates indicates that at the peak of expected true positives (Fig. S7*A*)

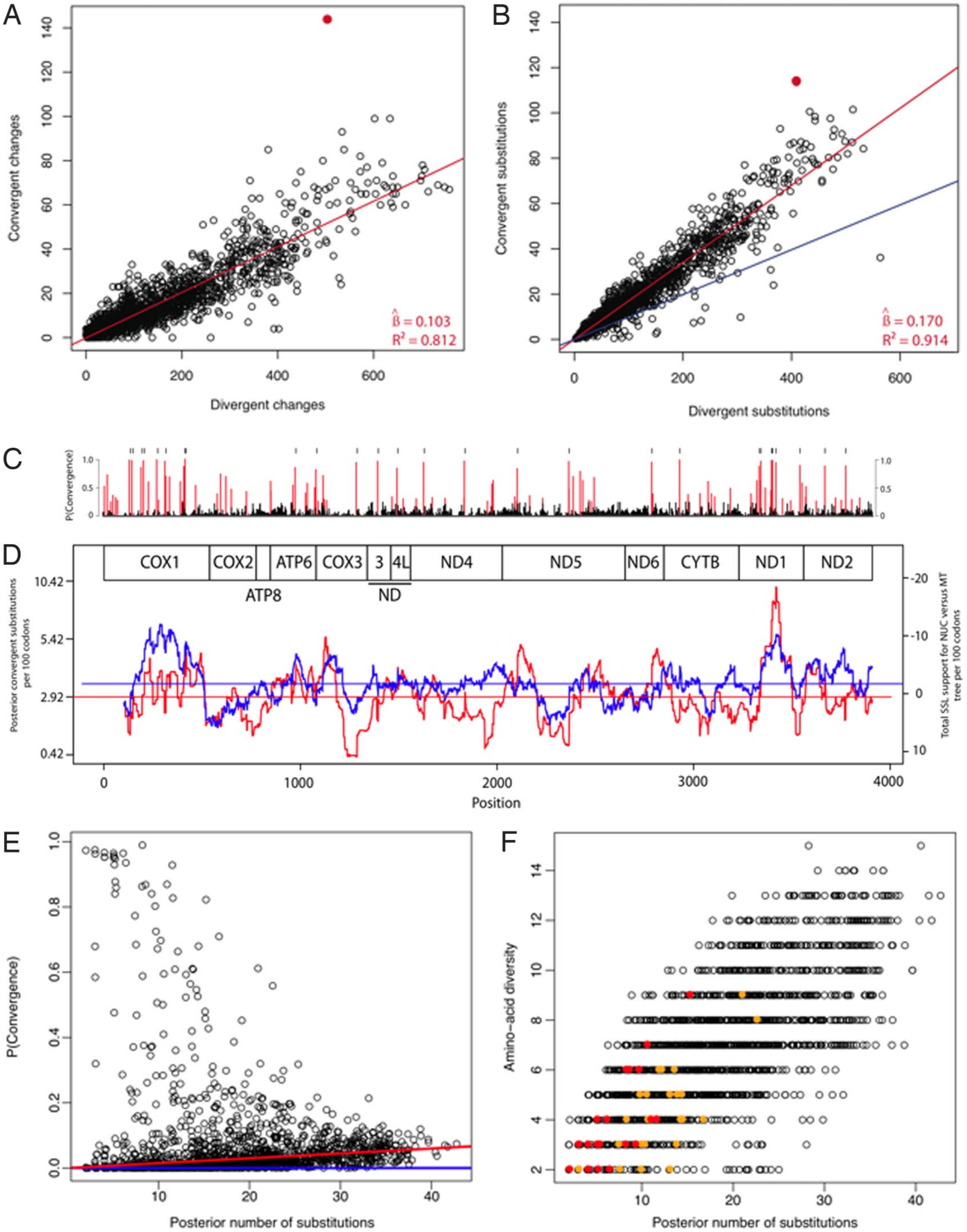

Fig. 3. Convergent evolution of protein sequences. The number of convergent and divergent substitutions in all pairs of branches along independent lines of descent was estimated (*A*) by using the ML marginal ancestral reconstructions, and (*B*) by using a Bayesian approach that calculated the posterior probability of all possible substitutions (see text). The numbers of convergent substitutions were related to the numbers of divergent substitutions by using orthogonal regressions (red line). The snake–agamid branch pair is well above the other branch pairs, regardless of the methodology used (red circles). The asymptotic calculation of the random expected fraction of neutral convergent substitutions, conditional on the ML parameter estimates from the observed data, is shown for reference (blue line in *B*; $\hat{\beta} = 0.099$). (*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). (*E*) Site-specific posterior number of substitutions versus the posterior probability of convergence per site; posterior substitutions were calculated to reduce the dependency of each site's rate estimate on the model of evolution. (*F*) Relationship between rates of evolution at a site, the probability of convergence, and the observed amino acid state space. Sites with posterior probabilities of convergence >80% are shown in red and >50% are shown in orange.

there are 11 likely excessively convergent branch pairs (Fig. S7 *B–L*), of which 6 are expected to be false positives.

To augment our nucleotide analyses of LBA and variable purifying selection at different sites, we considered whether some combination of variable rates and amino acid state space restriction might provide a neutral explanation for the excess convergence described above. Under neutrality and LBA, the fastest and most restricted sites should experience the most (neutral) convergence. The number of substitutions per site–integrated over the posterior distribution of ancestral states–was compared with the posterior probability of agamid-snake convergence (Fig. 3*E*), and for different observed levels of site-specific amino acid diversity (Fig. 3*F*). Many of the highly probable convergent substitution pairs occurred at constrained sites, but they tended to be slowly evolving sites rather than fast ones (Fig. 3*F*). This result is inconsistent with predictions of LBA due to neutral convergence, but is consistent with nonneutral convergence (possibly driven by positive selection for adaptation) at otherwise highly conserved, slowly evolving amino acid positions. Interestingly, there is also a small but clear increase in the probability of convergence at rapidly evolving sites (Fig. 3*E*, red line), which is consistent with a neutral LBA effect at the fastest sites. These sites have an average probability of convergence well <10%, however, and are not the ones that strongly support the MT tree. We thus conclude that while neutral convergence is present across this dataset (and in greater abundance than predicted by traditional evolutionary models), a substantial amount of nonneutral convergence has occurred between snakes and agamids, and it is this nonneutral convergence that is most likely responsible for misleading phylogenetic analyses.

Latent Phylogenetic Signal. If selection at the amino acid level has caused nonneutral convergence and resultant phylogenetic error, it is reasonable to presume that only a fraction of all sites might have been involved in those selective events, and that the majority of sites might retain a latent, but correct, phylogenetic signal. To search for such a latent phylogenetic signal, we inferred phylogenies after screening out sites having the greatest likelihood support for the MT tree and, separately, sites with the greatest convergence probability between snakes and agamids. A nucleotide-based Bayesian phylogenetic analysis excluding the 500 codons with the highest ΔSSLS supporting the MT tree (including 10,227 bp) recovered a monophyletic Iguania supported with 100% posterior probability (Fig. S8). Excluding convergent sites had similar results; as convergent sites were screened out, posterior support for the MT topology using the CAT model lessened and support for the NUC topology also increased until relative support was reversed after removal of the top 5% of convergent sites (Fig. S6 *B* and *C*). It thus appears that most of the mitochondrial data supports phylogenetic relationships consistent with morphological and nuclear data, but that when all sites are considered this signal is overwhelmed by sites corrupted by nonneutral convergence.

Discussion

This study presents evidence for nonneutral convergent molecular evolution between snake and agamid lizards in multiple mitochondrial genes at the amino acid level. The convergent sites appear to strongly mislead phylogenetic reconstruction such that snakes and agamid lizards are phylogenetically clustered together, which is inconsistent with all previous estimates of squamate reptile relationships. The degree of convergence observed is well outside both empirical and model-based expectations, and was sufficiently large to overcome the presumed true phylogenetic signal in >11 kb of sequence data. We conclude that the aberrant signal was created by episodes of nonneutral convergent molecular evolution between early lineages of snakes and a group of distantly related lizards. This case demonstrates, contrary to widespread belief, that nonneutral convergence can be a major force in molecular evolution, and that it should be considered more seriously as a cause of phylogenetic incongruence among datasets, especially in mitochondrial datasets.

We discovered this phenomenon because it was so extreme and because it severely disrupted phylogeny estimates in an obvious way. Less obvious cases are likely to exist, however, and may be mistakenly interpreted as strong evidence in favor of false phylogenetic relationships (11, 13, 26). A convergence event a fraction of the magnitude observed here could easily disrupt many topology estimates because of the relative biasing power of each convergent site. This possibility is worrisome, and phylogenetic reconstruction in the presence of nonneutral convergent evolution should therefore be taken seriously as a direction of future research and development. Improved models of sequence evolution can reduce the sensitivity of phylogenetic inference to LBA due to neutral processes (e.g., stochastic convergence and negative selection; ref. 3), and we expect that improved models could also increase the robustness of phylogenetic methods to the presence of nonneutral convergence.

An obvious potential explanation for this likely case of nonneutral convergent evolution is adaptation. It was previously shown that snake mitochondrial proteins have experienced the most extreme burst of apparently adaptive protein evolution yet observed in vertebrate mitochondria, including an astounding number of structurally linked coevolutionary events and changes likely to alter the function of important structural features in COX1 (7). The presence of so many adaptive replacements in snakes is consistent with the idea that the excess convergence levels observed here are due to adaptation as well. 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 (7). The molecular convergence between snakes and agamid lizards may thus have driven by similar adaptive pressures on metabolic function affecting both lineages. Similar to selected sites in the snakes (7), residues with the highest probability of convergence (Fig. S9) have no obvious physicochemical patterns, and a detailed structural analysis will likely be required to understand the potential functional consequences. Whatever the underlying cause, since the convergence extends across most regions of the mitochondrial genome, any common adaptive force must have been extensive and broad in scope.

Assuming an adaptive explanation is correct, the extreme scale and breadth of this event sets a new precedent for the extent to which natural selection can drive large-scale coordinated changes during protein evolution. This example involves many more convergent changes than in other known cases. The implication that this extreme event may have been caused by adaptive pressure on protein function suggests that further study may reveal valuable insight into the structure and function of mitochondrial proteins. The tendency for convergent amino acid substitutions to occur at otherwise conserved positions also implies that many of these convergent changes may have important structural and functional effects (7).

The data presented raise the question, is this phenomenon restricted to the mitochondrial genome or might it also affect nuclear datasets? The mitochondrial genome encodes metabolic genes that are functionally related. Directional selection may thus tend to affect many mitochondrial genes at once, leading to large-scale convergence if similar selective events occur in different lineages. Such multigenic directional selection may occur in nuclear genes as well, but it is reasonable to expect that in large nuclear datasets only a small proportion of genes will have similar selective pressures, and convergent events between lineages will involve a similarly small proportion of genes. Nevertheless, the potential for problems arising from nonneutral convergence in nuclear genes will depend on the strength of the true phylogenetic signal. Ancient divergences and rapid radiations can be difficult to infer with even a large amount of data. In such cases, and especially when nonneutral convergent events occur at otherwise well conserved sites, a single nonneutral convergent site can incorrectly appear highly informative and can potentially outweigh true phylogenetic signal at large numbers of neutral sites.

Based on our results, we have some recommendations on how the phylogenetic problem might be addressed effectively. Since the sites most likely to have undergone nonneutral convergence are those that are subject to strong selection, the simplest solution is to use sites that are likely to be nearly neutral (e.g., synonymous sites or introns) if such sites are abundantly available and the phylogeny is not too deep. If selected sites must be used, one should sample multiple nuclear loci, especially functionally unrelated ones, and determine if different genes have different phylogenetic signal (i.e., there is strong support for alternative gene trees). If so, one should use methods similar to those in this study to dissect whether some genes have sites that differentially support the different topologies and whether that support can be attributed to neutral or nonneutral convergence. Neutral convergence may be accounted for by using site-heterogeneous models, but for nonneutral convergence those sites should be removed until appropriate phylogenetic inference methods are developed that are robust to their presence. Ideally, the bulk of sites remaining will produce a consistent topology (or set of closely related topologies) that predicts the underlying species tree. Thus, when phylogenetic signal is variable among subsets of sites or data partitions, methods such as those used here can be used to infer the causes for variable phylogenetic signals among sites.

Materials and Methods

Mitochondrial Genome Sequencing, Alignment, and Phylogeny Inference. To increase sampling at the base of snake phylogeny, mitochondrial genomes were Sanger dideoxy sequenced and annotated for 2 snake species, *Anilius scytale* and *Tropidophis haetianus* (see *SI Methods*). All 13 mitochondrial protein-coding genes ($\approx$11,700 bp) from complete mitochondrial genomes of squamates available at the time of study, plus the 2 new species, were aligned using ClustalX (27) based on translated amino acid sequences; multiple species per genus were excluded (Table S1). Representatives of major tetrapod lineages were included to root the tree. Nucleotide sequences of 2 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 (*SI Methods*). Bayesian phylogenetic trees were estimated in MrBayes 3.0b4 (28) 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 (29); where relevant, *P* values $<$0.05 were considered significant. Support for alternative topologies was evaluated using the S-H test (20). Site-specific likelihood support (SSLS) was estimated using ML and a GTR+ΓI model per codon position. Bayesian analyses with PhyloBayes were performed using the CAT model with an unrestricted number of site profiles and a Dirichlet process prior, and discrete gamma-distributed rate variation across sites.

Analysis of Convergent Evolution. We used *PAML* (30) 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 2 branches) for all pairwise 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 2 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.

For our empirical Bayesian approach, we modified the *codeml* program of *PAML* (30) 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 2 branches in question. For analyses using the posterior number of substitutions per site, the posterior substitution probabilities were used to calculate the expected number of substitutions of each type, and were summed over branches. The details of these calculations are given in the *SI Methods*.

Using the posterior expected number of convergent substitutions with predicted levels of convergence (from orthogonal linear regressions), we performed 1-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 assumed each site was a drawn from a binomial distribution with a probability of convergence (*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.

ACKNOWLEDGMENTS. We thank Jimmy McGuire, Juan Daza, Alex Poole, Josh Reece, and 2 anonymous reviewers for many helpful suggestions and comments. We acknowledge the support of National Institutes of Health (NIH) Grants GM065612, GM065580, and GM083127 (to D.D.P.); the National Science Foundation (DEB-0416000) and a University of Central Florida startup package (to C.L.P.); and NIH Training Grant LM009451 (to T.A.C.).

1. Bull JJ, et al. (1997) Exceptional convergent evolution in a virus. *Genetics* 147:1497–1507.
2. Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol* 54:743–757.
3. Lartillot N, Brinkmann H, Philippe H (2007) Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol* 7(Suppl 1):S4.
4. Harmon LJ, Kolbe JJ, Cheverud JM, Losos JB (2005) Convergence and the multidimensional niche. *Evolution* 59:409–421.
5. Lee MSY (1998) Convergent evolution and character correlation in burrowing reptiles: Towards a resolution of squamate relationships. *Biol J Linn Soc* 65:369–453.
6. Wiens JJ, Chippindale PT, Hillis DM (2003) When are phylogenetic analyses misled by convergence? A case study in texas cave salamanders. *Syst Biol* 52:501–514.
7. Castoe TA, Jiang ZJ, Gu W, Wang ZO, Pollock DD (2008) Adaptive evolution and functional redesign of core metabolic proteins in snakes. *PLoS ONE* 3:e2201.
8. Jost MC, et al. (2008) Toxin-resistant sodium channels: Parallel adaptive evolution across a complete gene family. *Mol Biol Evol* 25:1016–1024.
9. Zakon HH, Lu Y, Zwickl DJ, Hillis DM (2006) Sodium channel genes and the evolution of diversity in communication signals of electric fishes: Convergent molecular evolution. *Proc Natl Acad Sci USA* 103:3675–3680.
10. Kornegay JR, Schilling JW, Wilson AC (1994) Molecular adaptation of a leaf-eating bird: Stomach lysozyme of the hoatzin. *Mol Biol Evol* 11:921–928.
11. Stewart CB, Schilling JW, Wilson AC (1987) Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* 330:401–404.
12. Zhang J, Kumar S (1997) Detection of convergent and parallel evolution at the amino acid sequence level. *Mol Biol Evol* 14:527–536.
13. Kitazoe Y, et al. (2005) Multidimensional vector space representation for convergent evolution and molecular phylogeny. *Mol Biol Evol* 22:704–715.
14. Rokas A, Carroll SB (2008) Frequent and widespread parallel evolution of protein sequences. *Mol Biol Evol* 25:1943–1953.
15. Estes R, De Queiroz K, Gauthier JA (1988) in *Phylogenetic Relationships of the Lizard Families, Essays Commemorating Charles L. Camp*, ed Estes R (Stanford Univ Press, Stanford, CA), pp 119–281.
16. Frost DR, Etheridge R (1989) A phylogenetic analysis and taxonomy of iguanian lizards (Reptilia: Squamata). *Misc Publ Mus Nat Hist, Univ Kansas* 81:1–65.
17. Fry BG, et al. (2006) Early evolution of the venom system in lizards and snakes. *Nature* 439:584–588.
18. Townsend TM, Larson A, Louis E, Macey JR (2004) Molecular phylogenetics of squamata: The position of snakes, amphisbaenians, and dibamids, and the root of the squamate tree. *Syst Biol* 53:735–757.
19. Vidal N, Hedges SB (2005) The phylogeny of squamate reptiles (lizards, snakes, and amphisbaenians) inferred from nine nuclear protein-coding genes. *Comptes Rendus Biologies* 328:1000–1008.
20. Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* 16:1114–1116.
21. Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence. *Mol Biol Evol* 11:605–612.
22. Krishnan NM, Seligmann H, Stewart CB, de Koning APJ, Pollock DD (2004) Ancestral sequence reconstruction in primate mitochondrial DNA: Compositional bias and effect on functional inference. *Mol Biol Evol* 21:1871–1883.
23. Williams PD, Pollock DD, Blackburne BP, Goldstein RA (2006) Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comp Biol* 2:e69.
24. Yang Z (2003) in *Handbook of Statistical Genetics*, eds Balding D, Bishop M, Cannings C (Wiley, New York), 2nd ed, pp 229–254.
25. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J R Stat Soc B Met* 57:289–300.
26. Stewart CB, Wilson AC (1987) Sequence convergence and functional adaptation of stomach lysozymes from foregut fermenters. *Cold Spring Harb Symp Quant Biol* 52:891–899.
27. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The clustal_x windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25:4876–4882.
28. Ronquist F, Huelsenbeck JP (2003) Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
29. Swofford DL (2001) *PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods)* (Sinauer Associates, Sunderland, MA).
30. Yang ZH (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–556.