Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome

Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome

Andrew M. Shedlock[†‡], Christopher W. Botka[§], Shaying Zhao[¶‖], Jyoti Shetty[¶††], Tingting Zhang[‡‡], Jun S. Liu[‡‡], Patrick J. Deschavanne[§§], and Scott V. Edwards[†]

[†]Department of Organismic and Evolutionary Biology, Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138; [§]Research Division, Joslin Diabetes Center, Harvard Medical School, One Joslin Place, Boston, MA 02215; [¶]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850; [‡‡]Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138; and [§§]Equipe de Bioinformatique Genomique et Moleculaire, Institut National de la Santé et de la Recherche Médicale (INSERM), 2 Place Jussieu, 75005 Paris, France

Edited by David B. Wake, University of California, Berkeley, CA, and approved December 26, 2006 (received for review July 24, 2006)

We report results of a megabase-scale phylogenomic analysis of the Reptilia, the sister group of mammals. Large-scale end-sequence scanning of genomic clones of a turtle, alligator, and lizard reveals diverse, mammal-like landscapes of retroelements and simple sequence repeats (SSRs) not found in the chicken. Several global genomic traits, including distinctive phylogenetic lineages of CR1-like long interspersed elements (LINEs) and a paucity of A-T rich SSRs, characterize turtles and archosaur genomes, whereas higher frequencies of tandem repeats and a lower global GC content reveal mammal-like features in *Anolis*. Nonavian reptile genomes also possess a high frequency of diverse and novel 50-bp unit tandem duplications not found in chicken or mammals. The frequency distributions of ≈65,000 8-mer oligonucleotides suggest that rates of DNA-word frequency change are an order of magnitude slower in reptiles than in mammals. These results suggest a diverse array of interspersed and SSRs in the common ancestor of amniotes and a genomic conservatism and gradual loss of retroelements in reptiles that culminated in the minimalist chicken genome.

BAC | Reptilia | retroelement | isochore | intron

Comparative genomics is a central focus of modern biology in part because it facilitates the understanding of principles of genome evolution (1–3). However, it is impractical to expect taxonomically broad comparative studies to proceed rapidly for nonmodel organisms on a whole-genome basis. A prime example of our limited understanding from the present handful of complete genomes is that we still do not know the sequence of genomic events that led to the structural diversity seen in mammalian genomes and those of their sister group, the Reptilia, which includes birds (4). The draft chicken genome (5) substantially increases our understanding of amniote comparative genomics, but evolutionary interpretation relying solely on chicken–mammal contrasts will remain difficult without new data for phylogenetically intermediate lineages. On the one hand, the common amniote ancestor may have had a small genome as in extant birds, with mammals and nonavian reptiles independently acquiring transposable elements that resulted in genome size increases in these two lineages. On the other hand, the common amniote ancestor may have had a large, repeat-rich genome as in extant mammals, with multiple sequential reductions in retroelement abundance occurring in the lineages leading to the smaller genomes of nonavian reptiles and birds (6). A third scenario might include a combination of both independent gains and reductions of specific genomic elements. Here we use a BAC- and plasmid-end sequencing approach in exemplars of three major nonavian reptile lineages, American Alligator (*Alligator mississippiensis*), Painted Turtle (*Chrysemys picta*), and the Bahamian Green Anole (*Anolis smaragdinus*), to better characterize the sequence of genomic changes underlying the diversification of amniote genomes.

Little is known about the large-scale structure of nonavian reptile genomes at the sequence level. Alligator and turtle genome sizes are ≈30% smaller than human, ≈50% larger than chicken, and only ≈12% larger than *Anolis*, whose genome size is close to the mean for nonavian reptiles. Unlike alligator genomes, the anole, painted turtle, and chicken contain a significant number of microchromosomes (7), which we expect would be gene rich as reported for chickens (8) and the soft-shelled turtle (9). In general, it is unknown how the macrochromosomes of reptiles differ from those of mammals (10) and those of the nonavian reptiles investigated here. The turtle and alligator species investigated here have environmental as opposed to genetic sex determination, and sex determination in *Anolis* is inferred to be genetic based on some karyological evidence (11). Several retroelement lineages have been characterized in turtles and other reptiles (12–15). Projects in progress will produce genome sequences for another bird, the Zebra Finch, *Taeniopygia guttata*, and a lizard, *Anolis carolinensis*. In the meantime, our goal in this project was to quickly amass a moderate database of primary sequence distributed throughout the genomes of genomically understudied lineages, which can reveal numerous genomewide trends that help characterize the most fundamental aspects of genome structure. Although the genomes we have investigated may not reflect specific changes in subclades of diverse groups such as squamates, any shortcomings of our limited taxonomic sampling are overcome by our ability to present a broad-brush window on genomic trends for nonavian reptiles, thereby quickly placing the chicken and mammal genomes in broader context.

Results and Discussion

Genome Scans and Global GC Content. Our survey includes edited, high-quality sequence reads covering 2,519,551 bp from American Alligator, 2,432,811 bp from Painted Turtle, and 1,358,158 bp from the Bahamian Green Anole, derived from a total of 8,638 nonoverlapping paired BAC- and plasmid-end reads (see supporting

Author contributions: A.M.S., J.S.L., P.J.D., and S.V.E. designed research; A.M.S., C.W.B., S.Z., and J.U.S. performed research; A.M.S., C.W.B., T.Z., J.S.L., P.J.D., and S.V.E. analyzed data; and A.M.S. and S.V.E. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS direct submission.

Abbreviations: CR1, Chicken Repeat 1; LINE, long interspersed element; MIR, Mammalian Interspersed Repeat; SSR, simple sequence repeat.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. CZ250707–CZ257443 and DX390731–DX389174).

[‡]To whom correspondence should be addressed. E-mail: shedlock@oeb.harvard.edu.

[‖]Present address: Department of Biochemistry and Molecular Biology, Institute of Bioinformatics, University of Georgia, 120 Green Street, Athens, GA 30602.

[††]Present address: J. Craig Venter Institute Joint Technology Center, 5 Research Place, Rockville, MD 20850.

This article contains supporting information online at www.pnas.org/cgi/content/full/0606204104/DC1.

EVOLUTION

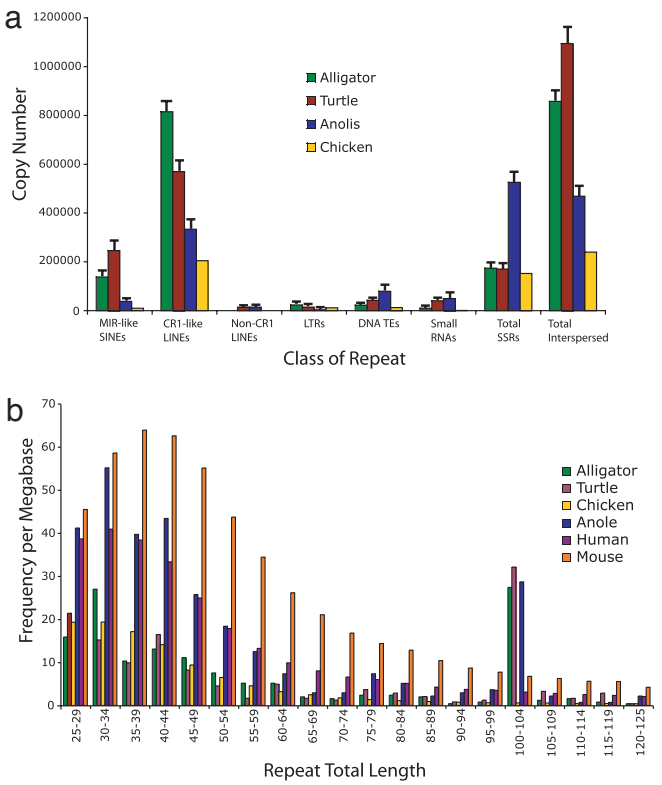

Fig. 1. Summary of interspersed and tandem repeats in nonavian reptiles. (*a*) Estimated copy number per genome of repetitive elements for four reptilian species. Estimates with error bars are based on RepeatMasker (37) queries against the chicken and primate database, summarized in SI Table 1. Error bars are 95% confidence intervals for the entire genome. Copy numbers for chicken are taken from published whole-genome assembly results (5) and targeted hybridization studies of avian microsatellites (23). DNA TE, DNA transposable element. (*b*) Histogram of frequencies of total tandem repeat array lengths, measured in base pairs, for the same sequence data examined in *a*. Details of repeat detection and analysis are presented in *Materials and Methods* and *SI Text*.

information (SI) *Text* and *Dataset 1* for details of sequence generation and GenBank accession numbers). Using a base composition model allowing for an inhomogeneous distribution of GC content among sequence reads, the estimated GC content (SI Fig. 5) for alligator and turtle agrees closely with estimates for these species based on buoyant density gradients and flow cytometry (16, 17). There are no previously published estimates of *Anolis* genomewide GC content; our results are slightly lower than those reported for other lacertid lizards and in close agreement with estimates for viperid and colubrid snakes (16, 17). Alligator, turtle, and *Anolis* GC means are significantly higher than those for whole human and chicken genomes (5, 10) and for *in silico* sampling of chicken and human BAC-end sequences (see *Materials and Methods* and *SI Text*). The distribution of GC content among all sequence reads (SI Fig. 5) shows a conspicuous tail of high GC values as in comparable human, chicken, mouse, and pufferfish genome sequence data (3, 5, 10, 18).

Retroelement Landscape. The repetitive landscape of nonavian reptile genomes includes a diversity of transposable elements, dominated by ancient and diverse non-long-terminal repeat (non-LTR) retrotransposons in the Chicken Repeat 1 (CR1) long interspersed element (LINE) family, Mammalian Interspersed Repeat (MIR)-like short interspersed elements (SINEs), and a low level of LTR retroelements, DNA transposons and small RNAs (Fig. 1*a* and SI Table 1). The abundance of CR1-like LINEs and MIR-like SINEs in our survey suggests that they are likely still active at a low level in nonavian reptiles; on the other hand, the drastic reduction of these elements and complete lack of full-length copies in the chicken suggest that here they may be approaching extinction (5). We estimate that the three reptile lineages investigated here possess from ≈24-fold to ≈4-fold greater numbers of MIR-like SINEs and >4-fold to 1.5-fold greater numbers of CR1-like LINEs compared with the chicken genome (Fig. 1*a*). Contrary to the findings of Lovsin *et al.* (14), we found a small number of non-CR1-like LINEs in our turtle survey. The absence of CR1 sequences >925 bp in our sampling of the 3′ termini from all three species (SI Fig. 6) is partly a consequence of our sequencing strategy but also suggests that most nonavian reptile elements are defective because of extensive 5′ truncation, a common feature of the vast majority of CR1s observed in vertebrate genomes (19). Nonetheless, we estimate that from ≈10% to 26% of the difference in genome size between birds and other reptiles derives from loss of transposable elements in birds. These results also suggest a persistence of active reptile MIRs in these lineages or their immediate ancestors and are consistent with conserved CORE SINEs giving rise to a diversity of MIR-like elements among vertebrates that have survived >550 Myr of eukaryotic genome evolution (20). By contrast, such persistence is not apparent in the chicken genome, which apparently has lost most of these elements from nonavian ancestors.

Using both Bayesian and distance methods, we evaluated the relationship between host species and CR1 element diversity by phylogenetic analysis of aligned 3′ terminal regions of 308 reptilian LINEs, including published avian CR1 subfamily sequences for chicken and other birds, and the tortoise psCR1 element (Fig. 2). In both analyses, although divergences among elements were too great to resolve relationships of many of the basal nodes in the CR1 tree, a number of well supported nodes were obtained at intermediate and shallow levels of divergence (Fig. 2*b*). Published sequences for CR1 avian subfamilies (A-E, emu, crane) and tortoise psCR1 sequences were taken from refs. 15 and 19 and clustered with chicken and turtle BAC-end sequences, respectively. The degree to which our gene tree of >300 elements reflect species phylogeny can be evaluated in terms of the degree of host-specific association among the four reptile species represented. CR1 clades represented by a single host species can be considered more evolutionarily distinct than clades that contain representatives of multiple species; in turn, reptile lineages with high CR1 specificity are likely phylogenetically divergent (21). We tested the significance of this specificity by constructing a null hypothesis for the extent of host species character change given our sampling by using a randomization test (22). The test indicated a highly significant level of clustering of CR1s by species (46 observed host changes vs. an expected mean of 157 ± 0.0216; $P < 0.001$). The frequency of species representation in nonoverlapping clades is summarized in Fig. 2*c* and indicates the greatest extent of CR1 lineage sharing occurs between alligator and turtle elements, followed by alligator–turtle–chicken combinations. By contrast, most *Anolis* CR1 lineages are phylogenetically distinct, and none clustered significantly with CR1s from other species. Significant phylogenetic clustering of CR1 elements by species suggests multiple episodes of within-lineage diversification after speciation. The high frequency of CR1 clades containing turtle, alligator, and chicken elements indirectly suggests a phylogenomic affinity of turtles and archosaur species to the exclusion of *Anolis*. The pattern of lineage-specific evolution observed is consistent with turtle and alligator possessing relics of ancient CR1s and partner MIRs that arose before the split with mammals >310 Myr ago as well as younger, active members of these repeat families that have emerged after the divergence of these species from one another.

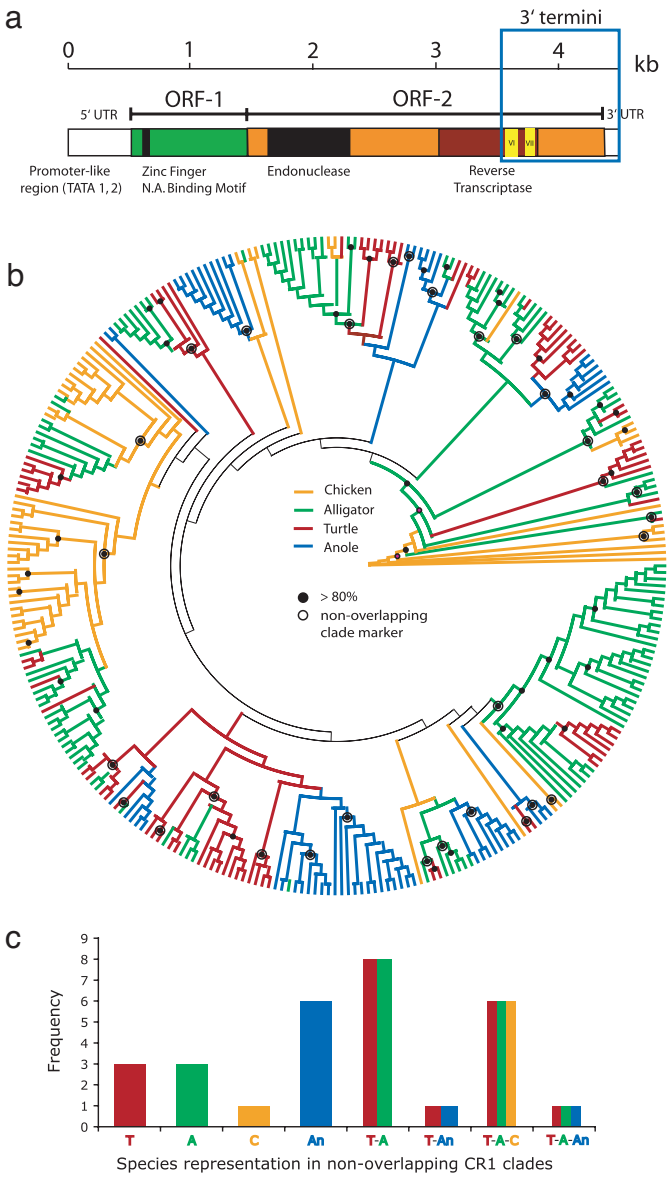

Fig. 2. Phylogenetic analysis of CR1 elements. (*a*) Diagram of the ≈4.5-kb full-length CR1-like LINE element structure. The 3′ terminal region analyzed is boxed, including the untranslated region (UTR) and conserved ORF (ORF-2) reverse-transcriptase domains. (*b*) Neighbor-joining tree of genetic distances among 308 3′ CR1 termini (alignment length = 168–976 bp) for four reptilian species with bootstrap support and host-species indicated by color. Outgroup is arbitrary and is not meant to indicate ancestral lineages. Bayesian analysis yielded similar results (see text). (*c*) Relative frequency of species representation in nonoverlapping CR1 clades with >80% bootstrap support annotated in *b*. T, turtle; A, alligator; C, chicken; An, *Anolis*. Details of sequence alignment and phylogenetic analysis are listed in *Materials and Methods*.

Diversity of Simple Sequence Repeats (SSRs) in Reptiles. For SSRs, nonavian reptiles exhibit a bimodal pattern. On the one hand, turtle and alligator exhibit distributions of repeat unit length very similar to chicken for short classes <30 bp (SI Fig. 7). By contrast, for these same short SSR classes, *Anolis* exhibits a distribution almost indistinguishable from that observed in the human genome. As expected, the highest and lowest total frequencies of SSRs across all categories for amniotes sampled are found in the mouse and chicken genomes, respectively (SI Table 2 and refs. 5, 8, 18, and 23). A bimodal pattern also holds for classes of SSR total array length (Fig. 1*b*), with *Anolis* again showing a distribution very similar to humans, and the turtle and alligator showing higher frequencies of long repeat arrays compared with chicken. The exception to the pattern of *Anolis*-human similarity in SSR distribution is our detection of high frequencies of a previously unknown yet diverse assemblage of ≈50-bp-unit tandem duplications among the three nonavian reptiles that is not apparent in the chicken, human, or mouse genome assemblies. Multiple alignment of these 50-bp repeat loci demonstrates that they do not exhibit significant sequence similarity ≥50%, nor do they share positional identity based on flanking sequence profiles (SI Fig. 9). A preliminary survey of BLAST alignments (24) for sequences from these repeat loci did not reveal any obvious patterns of functional significance. These repeats exhibit a diversity of GC contents ranging from 13% to 70% (SI Fig. 10). The presence of this anomalous repeat class suggests enzymatic mechanisms in nonavian reptiles that were presumably never present, or were active at a much lower level, in mammals.

The landscape of SSRs in the *Anolis* genome is divergent from the other reptiles and similar to those of mammals in other ways. *Anolis* exhibits a surprising 3-fold increase in predominantly short A-T rich SSRs compared with other reptilian species examined despite its relatively small genome (Fig. 1), revealing similarities to the SSR landscape of rodents, where $(AN)_n$ and $(AAN)_n$ motifs can be up to 12 times more frequent than in humans (18). Overall, our summary of SSR pattern size and array length distributions (Fig. 1*b* and SI Fig. 7) reveals that *Anolis* possesses on the order of two to three times as many SSRs per megabase than turtle and archosaurian genomes and exhibits a surprisingly mammalian-like pattern that shares aspects of both human and rodent distributions.

Genomic Signature Analysis. We sought to characterize major features of reptile genomes at the nucleotide level and further quantify genomic synapomorphies by using the clone-end sequences. We were able to directly compare our heterogeneous and unalignable reptile BAC- and plasmid-end sequences to the genomes of six other vertebrates in a 84.1 Mb phylogenomic analysis through an oligonucleotide- (DNA word-) counting approach in which the frequencies of all possible words consisting of *n*-nucleotides are counted and compared quantitatively (25, 26). The longest words for which we could reliably estimate frequencies in our reptile data set was eight nucleotides (25). These frequencies can be summarized visually as genomic signatures, consisting of pixel representations of the frequencies of all possible 65,536 (4^8) eight-nucleotide words (25). The signatures for the turtle, alligator and *Anolis* sequences exhibit the strong diagonals (indicating high frequency of homo-purine and -pyrimidine tracts) and the low frequency of motifs containing the CG dinucleotide that are found in other vertebrates (Fig. 3). The high density of AT-rich motifs in *Anolis*, as previously suggested by the summary of SSRs, is also visually apparent in the bottom corners of its signature.

Our estimate of phylogenomic distances between signatures of eight vertebrate species, using the zebrafish signature as an outgroup, suggests a sister relationship of the alligator and turtle to the exclusion of chicken, with 83% bootstrap support. Otherwise, the analysis supports the traditional relationships of tetrapods. These results, based on higher-order homologies embedded in word counts, agree with a growing body of molecular and fossil evidence (4, 27, 28) in placing turtles in a derived position relative to lepidosaurs (lizards + snakes) rather than in the traditional basal position within Reptilia. However, our topology conflicts with several recent analyses of aligned sequences that suggest a sister relationship of turtles to an archosaur clade (27, 29–33). Analysis of word frequencies is not expected to achieve the phylogenomic precision of aligned DNA sequences, yet we were surprised at the ability of genomic

EVOLUTION

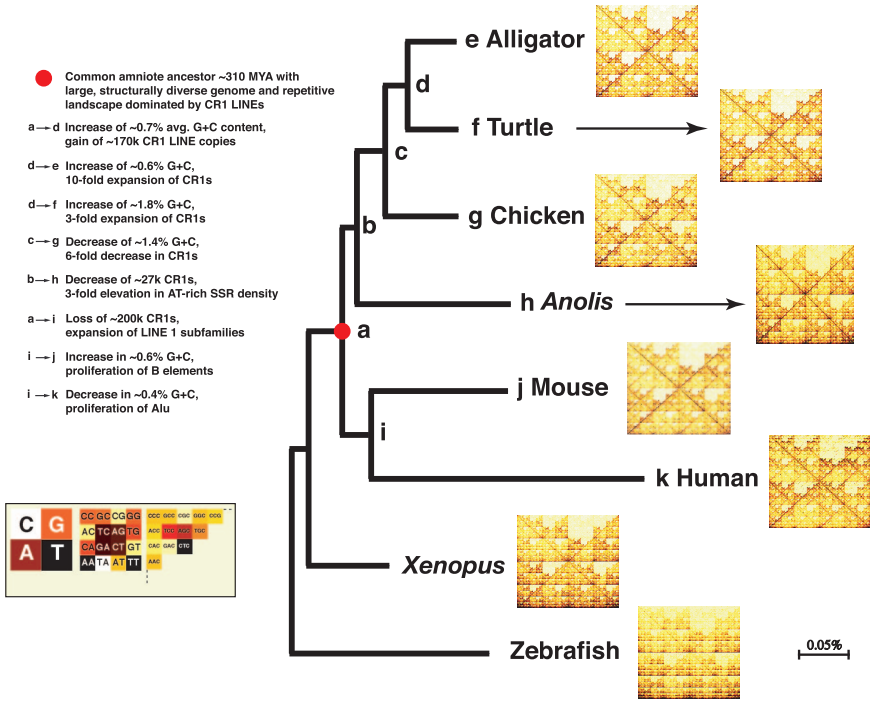

Fig. 3. History of amniote genomes and genomic signatures. Neighbor-joining tree of relationships based on Euclidean distances between signatures is shown. All nodes are resolved by >70% bootstrap support except at node b (10^3 replications). Genomic signatures are presented for eight vertebrates (zebrafish = outgroup) based on the frequency of all possible 8-nt DNA words contained in sequences analyzed. A key illustrates dark-colored (more frequent words) and light-colored (less frequent words) pixels used to construct signatures. Approximate amount of DNA sequence in megabases (and genomic source) used to construct genomic signatures are as follows: alligator, 2.4; turtle, 2.4; chicken, 6.1 (multiple chromosomes); *Anolis* (1.3); mouse, 23.7 (chromosome 17); human, 32.7 (chromosome 22); *Xenopus*, 2.6 (multiple chromosomes); Zebrafish, 14.9 (multiple chromosomes). Trends in amniote genome evolution are annotated with specific nodes and tips labeled a–k. Estimated amounts of evolutionary change indicated for CR1 LINE copies and average GC content are based on optimization of these traits across the tree using a phylogenetic generalized least squares analysis implemented in COMPARE v. 4.6 (ref. 35 and SI Table 4). Details of genome signature construction and phylogenetic analysis are presented in the text and *Materials and Methods*.

signatures to recover all but one node in the amniote tree congruent with aligned sequence analysis (27, 29–33). We suspect that homoplasy, nonindependence of word frequencies, limited taxon sampling and retention of pleisiomorphic word frequencies in turtles and alligators all contribute to elevated support for a turtle–alligator clade.

A surprising feature of our genomic signature tree is that the branch lengths within the reptiles are shorter than those between mouse and human, despite hypothesized divergences of turtles, alligators and birds >200 Mya and their common divergence from lizards >240 Mya (27, 34). To quantify rates of change of word frequencies within amniote genomes, we used comparative methods (35) to estimate amounts of change along branches of our signature tree. We find that a similar set of words, primarily from the noncoding portion of reptile genomes and including mononucleotide repeats (MNRs), low-complexity repeats (LCRs), and SSRs, comprise the fastest-changing component of the eutherian and reptilian genomes we examined (SI Table 3). No significant differences in rate estimates were obtained by using alternate phylogenetic positions for turtle. In particular, this and other surveys (23, 36) indicate that the MNR A_8/T_8 is typically the most frequent 8-mer in vertebrate genomes (dark lower left and right pixels in signatures; Fig. 3).

For the 50 words changing most in frequency among amniote genomes the rate of change along mammalian lineages is on average an order of magnitude higher than the rate found in bird and reptile lineages (Fig. 4). This high rate is particularly pronounced among dinucleotide microsatellites where mammal frequencies appear to be changing between ten and 25 times faster than in nonavian reptiles. Words changing faster in reptiles than in mammals tend to have several orders of magnitude smaller absolute differences in rates as compared with words changing faster in mammals (SI Table 3). Although our analysis is based on heterogeneous, unaligned and nonhomologous se-

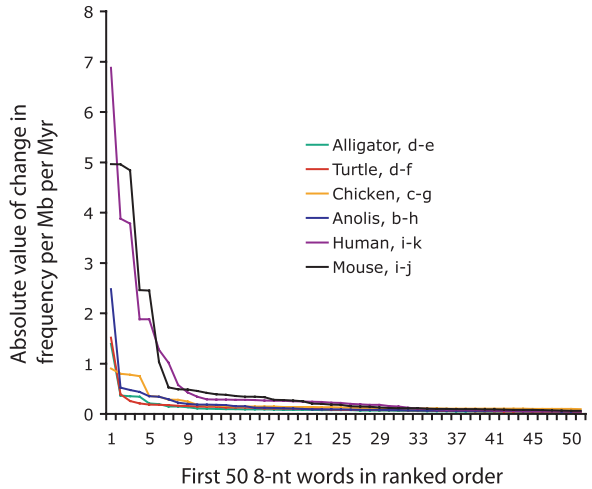

Fig. 4. Rapid evolution of genomic word-frequency change in mammals. Estimates of amounts of lineage-specific change are based on a phylogenetic generalized least-squares analysis implemented in COMPARE v. 4.6 (35). Rates and standard errors for a subset of the most rapidly evolving words analyzed are listed in SI Table 3. Word rank order plotted for each lineage is determined by rank order amount of change within each lineage and similar but not identical between lineages. Divergence times used for rate estimations are listed in *SI Text*.

quences, this result is nonetheless consistent with the unexpectedly low proportion of alignable sequence observed in targeted genomic regions of, for example, humans and rodents (2). In fact, we find that rates of word frequency change between homologous regions of mammals, such as the *CFTR* region of Thomas *et al.* (2), often exceed those for unalignable BAC- and plasmid-end sequence of reptiles in our data set. Although the molecular and selective basis of global word-frequency spectra in genomes is unclear (25, 36), sequence slippage, and to a lesser extent biased patterns of point mutation, gene conversion and repair, are all likely important for modulating word frequencies. By counting differences in word frequency between chicken BAC and EST sequences, as well as between BAC sequences with and without high-confidence protein BLAST hits in the alligator and turtle sequences, we confirmed the expectation that the majority of the largest genomewide word frequency shifts in reptiles occur in noncoding regions (SI Fig. 8). Overall our analysis indicates that the large frequency change of specific simple and low-complexity repeats dominate evolution of genomic language in amniotes and reveals a slowdown in relation to the generation of higher-order complexity in reptile genomes (Figs. 3 and 4).

Optimizing GC content and CR1 LINE copy number quantitatively across the genomic signature tree suggests that the ancestral amniote genome had a GC content just over 41% and a CR1 density on the order of 260,000 copies (Fig. 3 and SI Table 4). By contrast, optimizing these genomic traits on a consensus tree of published data placing turtle as sister to an alligator–chicken clade produced two minor differences in the estimated amount of character change along the tree: (i) a larger increase in GC% (0.81 and 1.99) and CR1 copies (408,800 and 173,200) along branches leading to alligator and turtle, respectively; and (*ii*) a smaller increase in GC% (0.03) and CR1 copies (9,400) during the 10-Myr period between the divergence of turtles and the most recent archosaur common ancestor (SI Table 4). All other estimates of character change were identical or nearly so for both trees; consequently, none of our conclusions regarding trends in GC content or CR1 copy number were changed by considering an alternative phylogenetic position for turtle.

GC levels are inferred to have increased by 0.7% at the base of the alligator-turtle-chicken clade, followed by convergent increases of 0.6% and 1.8% during the past 207 Myrs in lineages leading to alligator and turtle, respectively. This contrasts with a marked decrease of 1.4% in the branch leading to chicken during roughly the same timeframe. A 10-fold increase in the number of CR1 copies in the branch leading to alligator, 3-fold expansion in the turtle lineage, and 6-fold reduction along the avian branch leading to chicken comprise the most significant events in the dynamics of amniote CR1 amplification. This pattern suggests that whereas active CR1s have nearly gone extinct in the chicken (5) they have undergone substantial recent diversification in nonavian reptiles. Moreover, a drastic loss of ≈200,000 CR1s occurred in the ancestral lineage in a span of only 65 Myrs before the divergence of rodents and primates. A preliminary survey of LINE densities in a monotreme (duck-billed platypus; *Ornithorhynchus anatinus*) and a marsupial (South American opossum; *Monodelphis domestica*) revealed 27- and 3-fold greater incidence of non-CR1 vs. CR1 elements, respectively, per megabase of BAC clone sequence examined in each of these two species (A.M.S., unpublished data). These patterns and our estimates of ancestral states support the hypothesis that CR1s began declining early in mammalian evolutionary history and were displaced by younger LINE-1 elements which have since proliferated to high copy number in eutherians, for example as in mouse and human where LINE-1 comprises ≈18% of the genome relative to <1% for CR1 (18).

Conclusion

In summary, our analysis suggests that the ancestral amniote genome featured a relatively low global GC content as in mammals and a rich repetitive landscape dominated by CR1 and MIR retroelements and an abundance of AT-rich SSRs. Our finding of diverse CR1 lineages in nonavian reptiles qualifies a model in which a chicken-like streamlined ancestral amniote genome underwent expansion in mammals and nonavian reptiles independently (6). Rather it implies a complex scenario in which the diversity of CR1 elements in the ancestral amniote underwent a wholesale replacement by L1 and related mobile elements in mammals, and in which multiple sequential reductions in diversity occurred in the lineages leading to nonavian reptiles and birds. We expect that further genomic scans in additional reptile species, as well as further whole-genome sequencing projects, will considerably refine the major features in reptile genome evolution that we have outlined here.

Materials and Methods

Calculating Genomewide GC Content and Confidence Limits. When estimating genomewide GC content, we first checked for autocorrelation of bases up to 50 nt away from a focal base; finding none, we assumed a model in which the GC value for each read, y_i follows the binomial distribution with n_i trials and probability p_i. Because of inhomogeneous distribution of the GCs throughout the whole genome, each read may have a different p_i. To accommodate this feature, we further assume that the p_i are independently and identically distributed with an unknown density $f(p)$. The whole-genome GC content thus corresponds to $\theta \equiv E(p)$ under this unknown distribution [see complete formulas for mean and variance of $E(p)$ in *SI Text*).

Retroelement and Tandem Repeat Copy Number Estimation. Details of BAC sequence generation and informatics of data assembly and repeat detection are summarized in *SI Text*. We used RepeatMasker (37) to identify and summarize repeat content in our BAC and plasmid sequences. The informatics tools available in the online resource Tandem Repeat Database (TRDB; ref. 38) were used to detect and summarize distributions, and to align tandem repeats. To detect repeats in original nonavian reptile sequence data, we used default alignment parameters that were directly comparable to summary statistics available through TRDB for the most recent chicken, human, and mouse whole-genome assemblies.

CR1 Element Sequence Phylogeny. RepeatMasker output files were used to compile nucleotide sequences from 3′ termini of CR1 LINE elements (Fig. 2*a*), and data were aligned and edited for gap and terminal length variation by using ClustalW (39) to produce a multiple sequence alignment for 307 CR1 sequences across 1,477 sites. Neighbor-joining trees of genetic distances from aligned sequences were generated by using PAUP* (40) under the HKY85 substitution model and evaluated for bootstrap support (10^3 replications) and rooted arbitrarily as shown in Fig. 2*b* (midpoint rooting did not identify any sequences for obvious root selection). Tree-length frequency distributions calculated over 1,000 random equiprobable trees were evaluated for statistically significant levels of taxon-specific character change by using MacClade (Version 4.0; ref. 41). The phylogeny of retroelements was also evaluated by Bayesian analysis of the data matrix by using MrBayes (42) with 10^7 generations (first 10% as burn-in) run under a General Time Reversible model including an estimated proportion of invariable sites and gamma-shaped distribution of rate variation across sites. The alignment of 3′ CR1 termini is available from the authors upon request.

Genomic Signature Analysis. Genomic signatures were produced for seven vertebrate species according to the methods of Karlin and Ladunga (26) and Deschavanne *et al.* (25), the latter of which used a 1-bp sliding-window approach when counting words. The signatures were analyzed for phylogenetic structure by calculating Euclidean distances (the square root of the sum of the square of the differences in frequency of motifs) between signatures. In some analyses, genomic signatures derived from five sets of alternate chromosomal locations (see *SI Text*) and were normalized for genomewide base compositional differences before generating distance matrices by subtracting the expected frequency of each motif based on overall base composition from the observed frequency for each species. Bootstrapping was applied by random resampling of 8-nt words with replacement to create pseudosignatures for distance estimations, although we recognize that, like many types of molecular characters, DNA words are not independent variables. The program PAUP* (Version 4.0; ref. 40) was used for a neighbor-joining analysis of Euclidean distance matrices and to calculate bootstrap consensus results.

Evolutionary rates of word frequency changes, and their standard errors were estimated for specific lineages by using the Phylogenetic Generalized Least Squares approach described by Martins and Hansen (43) as implemented by using default settings in the PGLS-ancestor module of COMPARE (Version 4.6; ref. 35). Divergence times used to calibrate the neighbor-joining tree are listed in Supporting Information. The same methods were used to track changes in estimated global GC content and CR1 copy number along branches of the signature and consensus trees.

We thank C. Amemiya, J. Froula, J. R. Macey, and Z. Wang for technical assistance and BAC library construction; J. Losos, P. Minx, and W. Warren for providing plasmid sequences of *Anolis*; A. F. A. Smit, H. Ellegren, M. Gerstein, M. Lee, D. Zheng, Z.-X. Luo, and D. Burt for helpful discussion; C. Chapus, C. Moreau, G. Benson, and the Computational Biology Group at Harvard's Bauer Center for Genomics Research for technical support; and D. Burt, M. Long, and H. Wichman for helpful comments on the manuscript. This research was supported in part by National Science Foundation Grant IBN-0431717 (to S.V.E., Chris Amemiya, and J. R. Macey) and by Harvard University.

1. Pollock DD, Eisen JA, Doggett NA, Cummings MP (2000) *Mol Biol Evol* 17:1776–1788.
2. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, *et al.* (2003) *Nature* 424:788–793.
3. Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, *et al.* (2004) *Nature* 431:946–957.
4. Meyer A, Zardoya R (2003) *Annual Rev Ecol Systematics* 34:311–338.
5. International Chicken Genome Sequencing Consortium (2004) *Nature* 432:695–716.
6. Waltari E, Edwards SV (2002) *Am Naturalist* 160:539–552.
7. Burt DW, Bruley C, Dunn IC, Jones CT, Ramage A, Law AS, Morrice DR, Paton IR, Smith J, Windsor D, *et al.* (1999) *Nature* 402:411–413.
8. Ellegren H (2005) *Trends Ecol Evol* 20:180–186.
9. Kuraku S, Ishijima J, Nishida-Umehara C, Agata K, Kuratani S, Matsuda Y (2006) *Chromosome Res* 14:187–202.
10. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.* (2001) *Nature* 409:860–921.
11. Olmo E (1986) *Animal Cytogenetics: Chordata: Reptilia* (Gebrüder Borntraeger, Berlin), Vol 4, No 3A.
12. Chen Z-Q, Ritzel RG, Lin CC, Hodgetts RB (1991) *Proc Natl Acad Sci USA* 88:5814–5818.
13. Kajikawa M, Ohshima K, Okada N (1997) *Mol Biol Evol* 14:1206–1217.
14. Lovsin N, Gubensek F, Kordis D (2001) *Mol Biol Evol* 18:2213–2224.
15. Sasaki T, Takahashi K, Nikaido M, Miura S, Yasukawa Y, Okada N (2004) *Mol Biol Evol* 21:705–715.
16. Hughes S, Clay O, Bernardi G (2002) *Gene* 295:323–329.
17. Vinogradov AE (1998) *Cytometry* 31:100–109.
18. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al.* (2002) *Nature* 420:520–562.
19. Vandergon TL, Reitman M (1994) *Mol Biol Evol* 11:886–898.
20. Gilbert N, Labuda D (1999) *Proc Natl Acad Sci USA* 96:2869–2874.
21. Page RM, Charleston MA (1997) *Mol Phylogenetics Evol* 7:231–240.
22. Maddison WP, Slatkin M (1991) *Evolution (Lawrence, Kans)* 45:1184–1197.
23. Primmer CR, Raudsepp T, Chowdhary BP, Moller AP, Ellegren H (1997) *Genome Res* 7:471–482.
24. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D (1997) *Nucleic Acids Res* 25:3389–3402.
25. Deschavanne P, Giron A, Vilain J, Fagot G, Fertil B (1999) *Mol Biol Evol* 16:1391–1399.
26. Karlin S, Ladunga I (1994) *Proc Natl Acad Sci USA* 91:12832–12836.
27. Hedges SB, Poling LL (1999) *Science* 283:998–1001.
28. Reipell O, deBraga M (1996) *Nature* 384:453–455.
29. Matsuda Y, Nishida-Umehara C, Tarui H, Kuroiwa A, Yamada K, Isobe T, Ando J, Fujiwara A, Hirao Y, Nishimura O, *et al.* (2005) *Chromosome Res* V13:601–615.
30. Zardoya R, Meyer A (1998) *Proc Natl Acad Sci USA* 95:14226–14231.
31. Cao Y, Sorenson MD, Kumazawa Y, Mindell DP, Hasegawa M (2000) *Gene* 259:139–148.
32. Iwabe N, Hara Y, Kumazawa Y, Shibamoto K, Saito Y, Miyata T, Katoh K (2005) *Mol Biol Evol* 22:810–813.
33. Rest JR, Ast JC, Austin CA, Waddell PJ, Tibbets EA, Hay JM, Mindell DP (200) *Mol Phylogenetics Evol* 29:289–297.
34. Kumar S, Hedges B (1998) *Nature* 392:917–920.
35. Martins EP (2004) COMPARE (Department of Biology, Indiana Univ, Bloomington, IN), Version 4.6b. Available at: http://compare.bio.indiana.edu.
36. Toth G, Gaspari Z, Jurka J (2000) *Genome Res* 10:967–981.
37. Smit AFA, Hubley R, Green P (2004) RepeatMasker Open-3.0.5. Available at: www.repeatmasker.org.
38. Benson G (2006) Tandem Repeats Database (Laboratory for Biocomputing and Informatics, Boston University, Boston), Version 2.16. Available at: http://tandem.bu.edu/cgi-bin/trdb/trdb.exe?taskid=0.
39. Thompson J, Higgins D, Gibson T (1994) *Nucleic Acids Res* 22:4673–4680.
40. Swofford DL (1999) PAUP*: Phylogenetic analysis using parsimony (*and other methods) (Sinauer Associates, Sunderland, MA), Version 4.0b.
41. Maddison WP, Maddison DR (2000) *MacClade 4: Inter-active Analysis of Phylogeny and Character Evolution* (Sinauer, Sunderland MA).
42. Huelsenbeck JP, Ronquist F (2001) *Bioinformatics* 17:754–755.
43. Martins EP, Hansen TF (1997) *Am Naturalist* 149:646–667.