nature Vol 446|8 March 2007|doi:10.1038/nature05621

LETTERS

Origin of avian genome size and structure in non-avian dinosaurs

Chris L. Organ[1], Andrew M. Shedlock[1], Andrew Meade[2], Mark Pagel[2] & Scott V. Edwards[1]

Avian genomes are small and streamlined compared with those of other amniotes by virtue of having fewer repetitive elements and less non-coding DNA[1,2]. This condition has been suggested to represent a key adaptation for flight in birds, by reducing the metabolic costs associated with having large genome and cell sizes[3,4]. However, the evolution of genome architecture in birds, or any other lineage, is difficult to study because genomic information is often absent for long-extinct relatives. Here we use a novel bayesian comparative method to show that bone-cell size correlates well with genome size in extant vertebrates, and hence use this relationship to estimate the genome sizes of 31 species of extinct dinosaur, including several species of extinct birds. Our results indicate that the small genomes typically associated with avian flight evolved in the saurischian dinosaur lineage between 230 and 250 million years ago, long before this lineage gave rise to the first birds. By comparison, ornithischian dinosaurs are inferred to have had much larger genomes, which were probably typical for ancestral Dinosauria. Using comparative genomic data, we estimate that genome-wide interspersed mobile elements, a class of repetitive DNA, comprised 5–12% of the total genome size in the saurischian dinosaur lineage, but was 7–19% of total genome size in ornithischian dinosaurs, suggesting that repetitive elements became less active in the saurischian lineage. These genomic characteristics should be added to the list of attributes previously considered avian but now thought to have arisen in non-avian dinosaurs, such as feathers[5], pulmonary innovations[6], and parental care and nesting[7].

Birds have the smallest genomes of all amniotes, with an average haploid genome size of only 1.45 pg of DNA or roughly 1.45 billion bases[8]. Birds are therefore a useful group in which to study the causes and consequences of small genome size and the mechanisms by which genomes contract. Small genome sizes may have been favoured by the demands of flight, explaining the constricted genome sizes seen within Aves[3,4]. Consistent with this suggestion is the finding that flightless birds have larger genomes than birds that fly[4], and that bats possess smaller genomes than do mammalian sister groups[9]. Some comparative analyses suggest that genome size reduction may have begun in lineages of non-avian reptiles sometime before the origin of flight in birds[10,11], implying that small genome size evolved first but that powered flight drove further genomic contractions.

Previous investigations of genome size evolution in amniotes have largely been constrained to living species, which constitute only about 1% of all species that have ever existed in this branch of life[12]. But including long-extinct animals into comparative genomics studies has the potential to reveal the origins and macroevolutionary trends of genomic novelties, as well as the timing of major genomic changes, more accurately than by focusing solely on living species[13–15]. Here we combine information from 31 dinosaur species (some of which were extinct birds) with data on extant vertebrate species to characterize genome size and structure in extinct dinosaurs and evolutionary trends in amniote genome architecture.

In living organisms there is a well-known positive relationship between cell size and genome size[16]. Cell-size data are not generally available from extinct species, but it is possible to approximate osteocyte (bone-cell) size from fossilized bones. We sampled bone sections from 26 extant tetrapod species and fossilized bone from 31 extinct dinosaur species. We calculated osteocyte size directly from histological sections of bone by measuring the small pockets (lacunae) in the mineral matrix in which the bone cells resided during life (Fig. 1a). The distribution of osteocyte size in dinosaur bone is bimodal, with the two modes corresponding mainly to two major dinosaur groups, Ornithischia and Theropoda, with a single sauropod sample yielding an intermediate osteocyte size (Fig. 1b).

Using a phylogenetic tree and a bayesian implementation of the comparative method software Continuous[17,18], we found that osteocyte size predicted genome size in extant vertebrates well, explaining 59% of the variation (Fig. 1c) under a normal regression model and 32% of the variation under a generalized least-squares regression (GLS) model that corrects for the non-independence of data points arising from shared ancestry.

We then derived the posterior predictive distributions (see Methods and Supplementary Fig. 1) of genome size for the 31 extinct species of dinosaurs from this regression model. The posterior predictive distributions account for the phylogenetic relationships among taxa and for uncertainty about the true regression model. Except for *Oviraptor*, all of the inferred genome sizes for extinct theropods (average genome size of 1.78 pg, standard deviation s.d. = 0.3) fall within the narrow range of genome sizes for living birds (0.97–2.16 pg, haploid[8]), a result that follows from the restricted size range of theropod osteocytes.

Gaps among living taxa created by extinction might give the appearance of abrupt shifts in traits. This could be the case for genome size, which displays a punctuated distribution among extant vertebrates[10], even though genome evolution in vertebrates has been hypothesized to occur gradually and continuously[19]. Our results suggest that genome size evolution in dinosaurs was indeed abrupt, with a rapid reduction in size occurring between theropod and non-theropod dinosaurs (0.7 pg, or 28% difference in average genome size; phylogenetically corrected *t*-test, *P*-value = 0.008; Fig. 2), long before the origin of Aves and powered flight. If we consider *Herrerasaurus* a basal saurischian, as a recent study suggests[20], the decline in genome size is inferred to have begun abruptly earlier, at the base of Saurischia, denoted by a shift in genome size between the two primary clades of dinosaurs, ornithischians and saurischians (phylogenetically corrected *t*-test, *P*-value = 0.007). In either scenario, the theropods are characterized by a long period of relative

[1]Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138, USA. [2]School of Biological Sciences, University of Reading, Reading, RG6 6AJ, UK.

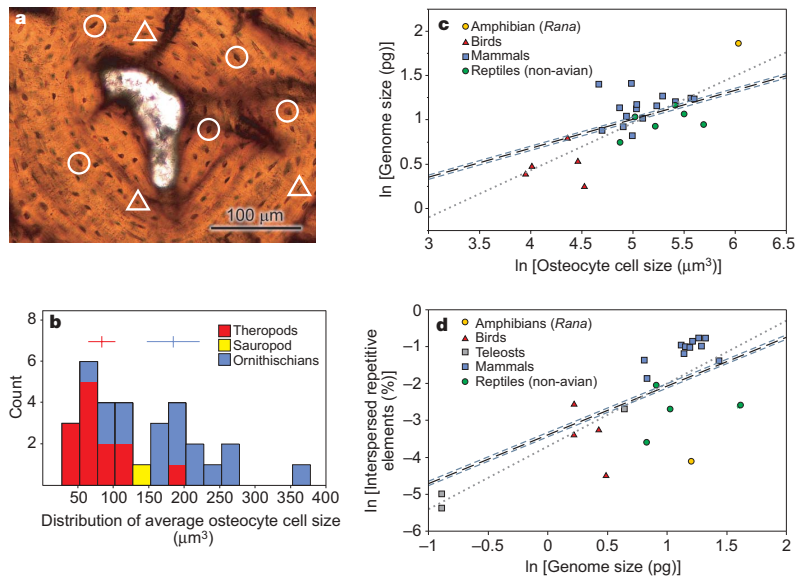

Figure 1 | Primary histological data, distribution of average osteocyte-cell size in extinct dinosaur species, and regression lines derived from data on extant animals used to infer genome size and interspersed repetitive elements. Black lines in **c** and **d** indicate the average of a bayesian posterior distribution of GLS phylogenetically corrected regression lines and grey dotted lines indicate their 95% confidence intervals (see Supplementary Information). **a**, Histological micrograph of compact cortical secondary bone from a diaphyseal cross-section of an *Allosaurus fragilis* (extinct theropod dinosaur) radius. Circles indicate the type of cells that were measured and triangles indicate the type of cells that were not measured. **b**, Distribution of average osteocyte cell sizes in 31 extinct dinosaur species. Red and blue crosses indicate the median (vertical) and standard (horizontal) deviation for theropod and ornithischian species, respectively. **c**, The regression line (grey) is ln(genome size) = −1.8 + 0.56 ln(cell size), $r^2 = 0.59$ (P-value < 0.0001, H_0, $\beta 1 = 0$). The GLS line (black) is ln(genome size) = −0.87 + 0.36 ln(cell size), $r^2 = 0.32$ (Bayes factor = 9.9, H_0, $\beta 1 = 0$). **d**, The regression line (grey) is ln(transposable elements) = −3.7 + 1.69 ln(genome size), $r^2 = 0.58$ (P-value < 0.0001, H_0, $\beta 1 = 0$). The GLS line is ln(transposable elements) = −3.39 + 1.32 ln(genome size), $r^2 = 0.34$ (Bayes factor = 9.4, H_0, $\beta 1 = 0$).

genome-size stasis (phylogenetic t-test for difference between genome size of birds and non-avian theropods, P-value = 0.24; between extant birds and non-avian theropods, P-value = 0.38), and the analysis suggests that the small, presumably gene-dense genomes in this clade have not changed substantially in size for the last 230 million years (Fig. 2).

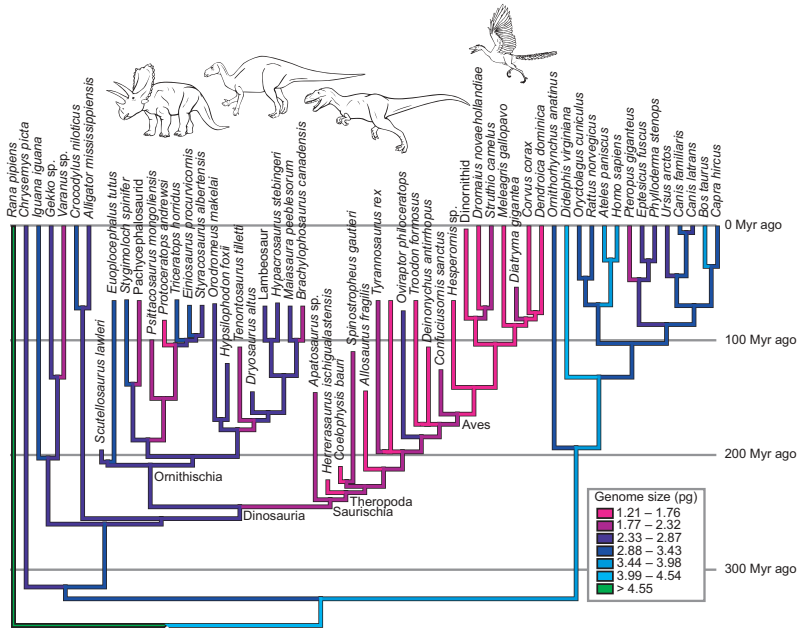

Figure 2 | Haploid genome size (mean of posterior predictive distribution) mapped onto a phylogeny shows a reduction within saurischian dinosaurs, the lineage to which birds belong. Myr, million years.

By comparison, we observe no reduction in genome size in ornithischian dinosaurs compared with non-dinosaurian relatives. We estimate that extinct ornithischians (average genome size of 2.49 pg, s.d. = 0.42) had genome sizes within the range of extant non-avian reptiles, such as monitor lizards (*Varanus*, 2.05 pg) and *Gekko* (2.56 pg), but are more comparable to the genomes of closer archosaurian relatives, such as *Crocodylus* (3.21 pg) and *Alligator* (2.49 pg). We therefore hypothesize that the ornithischian genome was structured more similarly to living basal archosaurs than to birds.

The profile of the repetitive landscape in genomes of major amniote clades can provide insight into the molecular processes that have modulated the nearly 380-fold range of genome sizes observed in living vertebrates[21]. We estimated the percentage of the genome comprised of the dominant fraction of repetitive DNA (interspersed mobile elements) in 24 extant vertebrate species using RepeatMasker and over 119 megabases of bacterial artificial chromosome (BAC)-end and scaffold DNA sampled from online databases, supplemented with data from whole-genome sequencing projects. We then generated regression models against genome size as outlined above (Fig. 1d), in a way that simultaneously accounts for the uncertainty in using an inferred variable (genome size) to infer additional variables (repeat density) and uncertainty about the true regression model. We also attempted to reconstruct genome-wide GC content in dinosaurs, but found no significant correlations of GC content with genome size or cell size in extant species.

The estimated density of interspersed repetitive elements within the genomes of dinosaurs averages 12% (s.d. = 3.1) for ornithischian versus 8.4% (s.d. = 2.6) for non-avian theropod species. Comparisons of the inferred density of repetitive elements between theropod and non-theropod dinosaurs suggest an abrupt change in this trait along the theropod lineage (phylogenetically corrected *t*-test, $P = 0.02$). Again, the phylogenetic placement of *Herrerasaurus* does not change estimations of repeat density in dinosaurs, but when considered a basal saurischian the analysis suggests that repeat density evolved abruptly earlier with the rise of Saurischia as opposed to Theropoda (the difference in retroelement density between saurischian and ornithischian dinosaurs in this scenario is significant: phylogenetically corrected *t*-test, *P*-value = 0.02). Yet there is no apparent difference in the inferred density of interspersed repetitive elements between birds and non-avian theropods (phylogenetic *t*-test for birds and non-avian theropods, *P*-value = 0.38; between extant birds and non-avian theropods, *P*-value = 0.1).

Genome size in eukaryotes is mediated primarily by proliferating and attenuating copies of retroelements over evolutionary time[21]. Analyses of mobile elements in extant sauropsid genomes suggest that chicken repeat 1 (CR1) long interspersed nuclear elements (LINEs) and related mammalian-wide interspersed repeat (MIR)-like short interspersed nuclear elements (SINEs) dominate the repetitive landscape and were probably active in the common archosaur ancestor some 250 million years ago[22]. Our results suggest that these ancient retroelements probably underwent different rates of lineage-specific expansion within ornithischians and saurischian dinosaurs, leading to a 50% difference in this genomic compartment between them, resulting in the repetitive landscape observed in living birds. Moreover, LINE elements in avian genomes appear to be almost exclusively represented by CR1 elements[23] though their activity appears to be nearly extinct in the chicken genome, on the basis of the paucity of full-length, intact CR1 elements within this host species[1]. Our estimates of the repetitive fraction of dinosaur genomes suggest that the reduction in CR1 activity began close to the base of carnivorous theropod or saurischian dinosaurs some 230 to 250 million years ago (Fig. 3). This model of reptilian genome evolution cannot be reconstructed from available bird–mammal whole-genome comparisons and emphasizes the value of incorporating long-extinct fossil evidence into phylogenomic studies.

Besides genes themselves, the architecture of whole genomes (genome size, gene number, chromosomal synteny, GC content, and the diversity of repetitive DNA elements) vitally links genome fluidity,

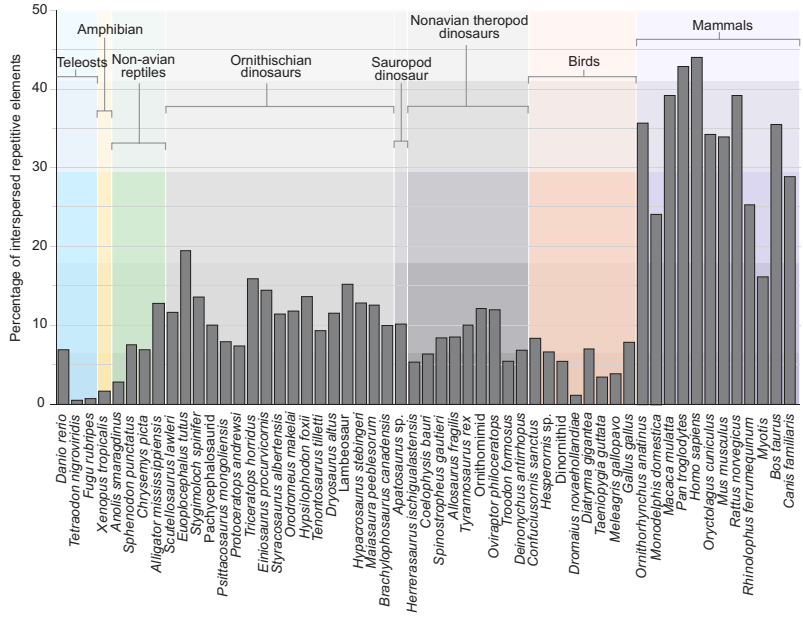

Figure 3 | Bar graph of interspersed repetitive elements in a range of extant vertebrate species and extinct dinosaur species inferred from genome size reconstructions and the correlation between repetitive element composition and genome size. Ornithischian dinosaur repetitive element densities are variable, especially within marginocephalians (*Triceratops* and relatives), but as a whole occupy a range similar to crocodilians. Conversely, the inferred percentage of repetitive elements in non-avian theropod genomes is smaller than that of ornithischian species (phylogenetic *t*-test, *P*-value < 0.02) and falls within the range of living birds.

life history and reproduction[24]. In addition, because genome size in part affects cell size, it has direct consequences for the rate of cell division, transcriptional processes, and cellular respiration[25]. Consequently, it is thought that physiological demands may have constrained the evolution of genome size in endothermic vertebrates[10,26] by favouring smaller red blood cells that increase surface area to volume ratios, and therefore their ability to facilitate gas exchange (a constraint that mammals may have circumvented with enucleate red blood cells)[4,27]. Our results suggest that this component of endothermy in living birds may have originated early in the saurischian/theropod lineage with commensurate changes in genome size, a conjecture consistent with studies of dinosaurian growth physiology using bone palaeohistology[28,29]. The later secondary expansion of genome size in flightless birds[3] suggests that, even though flight and genome size may not have arisen together, the two may be functionally related, perhaps at a physiological level.

METHODS

See Supplementary Information for additional details.

Histology and genome size data. Histological slides from extant and extinct adult, sub-adult and juvenile animals were obtained from museum collections. Digital micrograph images were randomly assigned labels and cell size (volume) data were measured blind for a total of 1,423 lacunae (osteocytes). Only larger cells from compact bone tissue, both primary and secondary, were sampled to help ensure that cells were measured near their mid-axis, thereby minimizing variation caused by slicing cells on different planes during sectioning. Data on genome size were obtained from the Animal Genome Size Database[8].

Repetitive element data. Human, dog, rat, chicken and mouse repetitive genome data were taken from published genome assemblies. For species for which published data were not available, we used the software program RepeatMasker (http://www.repeatmasker.org) to summarize repeat density and GC content in BAC clone sequences covering over 119 Mb downloaded from GenBank and random regions along scaffolds from published genomes on the UCSC genome browser (http://genome.ucsc.edu/).

Trees and character matrices. Mesquite version 1.11 (http://mesquiteproject.org/) was used to create character matrices and phylogenetic topologies. Branch lengths were estimated from the fossil record supplemented with molecular studies. Various tree topologies (*Chrysemys* as a basal amniote and as the sister group to archosaurs; *Herrerasaurus* as a basal saurischian and as a basal theropod) and branch-length scaling were used to assess the robustness of our approach. As mentioned in the text, most results were robust to these changes in topology and branch lengths.

Regression model. The dependent variable y and an independent variable x measured on a sample of n species are assumed to be related by the regression model $y_i = \beta_0 + \beta_1 x_i + e_i$, where β_0 is the y-axis intercept and β_1 is the slope of the line relating the y_i to the x_i. The e_i are the random errors and the residual variance of the regression is given by the variance of the $e_i = \sigma^2$.

Represented as vectors, $\mathbf{y}$ and $\mathbf{x}$ contain the observed data across species and are studied in a GLS framework such that $p(\mathbf{y}|\mathbf{x},m) \propto \exp\left\{-[\mathbf{y}-(\beta_0+\beta_1\mathbf{x})]' \mathbf{V}^{-1}[\mathbf{y}-(\beta_0+\beta_1\mathbf{x})]\right\}$, where m denotes the regression model, and $\mathbf{V}$ is the expected variance–covariance matrix of the residual errors given by the phylogenetic tree describing the relationships among species[17,18].

We account for the uncertainty about the true regression model $p(\mathbf{y})$ by defining $\mathbf{y}$ as given above, and the integral is approximated by Markov chain Monte Carlo (MCMC) methods[30]. A Markov chain is constructed in which new values of the parameters of the regression model are proposed on successive iterations of the Markov chain. At each step in the chain, models are accepted or rejected by the Metropolis–Hastings algorithm[30]. The chain is allowed to run to convergence, after which it samples from the posterior distribution of $p(\mathbf{y}|\mathbf{x})$ and the posterior distribution of m. All MCMC chains ran for 5,010,000 iterations with a burn-in of 1,000.

Posterior predictive distributions. We obtain estimates for unknown values of the dependent variable $\tilde{\mathbf{y}}$ in a new sample for which $\mathbf{x}$ is known by using the posterior distribution of m from $p(\mathbf{y}|\mathbf{x})$. The probability of the unknown $\tilde{\mathbf{y}}$ can be written as $p(\tilde{\mathbf{y}}|\mathbf{x}) = \int_m p(\tilde{\mathbf{y}}|\mathbf{x},m)p(m)\mathrm{d}m$, where the $\mathbf{x}$ values here correspond to species for which we wish to estimate the unknown $\tilde{\mathbf{y}}$. Our MCMC methods sample m and $\mathbf{e}$ together from their joint posterior distribution as derived in the initial regression modelling step, and assume that the e_i follow a multivariate normal distribution. These are combined to produce new values of $\tilde{\mathbf{y}} = \beta_0 + \beta_1\mathbf{x} + \mathbf{e}$. The posterior predictive distribution of genome sizes considers uncertainty in the parameters of the regression model as well as the inherent uncertainty of prediction summarized in σ^2.

Posterior predictive distributions of repetitive genetic elements are produced as above; let $\tilde{\mathbf{y}}$ denote the unknown values of these variables. The probability of these unknown values is $p(\tilde{\mathbf{y}}|\mathbf{x}) = \int_{\tilde{\mathbf{x}}}\int_m p(\tilde{\mathbf{y}}|\mathbf{x},\tilde{\mathbf{x}},m)p(m)p(\tilde{\mathbf{x}})\mathrm{d}m\mathrm{d}\tilde{\mathbf{x}}$. The Markov chain samples from posterior distributions of the model parameters (including residual error) derived from known data, and from the posterior distribution of the genome sizes of extinct dinosaurs. The posterior distribution of repetitive elements therefore considers uncertainty in the regression model and in the values of some of the predictors.

Model/method checking. We evaluated the adequacy of our regression models by generating simulated species data from a multivariate normal distribution of about MN(μ,$\mathbf{V}$). We also individually removed several extant species (*Meleagris*, *Capra*, *Iguana* and *Alligator*); we then inferred their genome size from osteocyte data, and found them to be within 2% to 15% of the reported average values (see Supplementary Information for more details). These tests show that by adding back phylogenetic information, initially 'lost' during GLS phylogenetic correction, our inferences of genome size are actually more accurate than a standard regression model, despite reduced r^2 values for phylogenetically corrected correlations.

Received 31 July 2006; accepted 25 January 2007.

1. Hillier, L. W. *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432,** 695–716 (2004).
2. Ellengren, H. The avian genome uncovered. *Trends Ecol. Evol.* **20,** 180–186 (2005).
3. Hughes, A. L. *Adaptive Evolution of Genes and Genomes* (Oxford Univ. Press, Oxford, 1999).
4. Hughes, A. L. & Hughes, M. K. Small genomes for better flyers. *Nature* **377,** 391 (1995).
5. Xu, X., Zhou, Z. & Prum, R. O. Branched integumental structures in *Sinornithosaurus* and the origin of feathers. *Nature* **410,** 200–204 (2001).
6. O'Connor, P. M. & Claessens, P. A. M. Basic avian pulmonary design and flow-through ventilation in non-avian theropod dinosaurs. *Nature* **436,** 253–256 (2005).
7. Horner, J. R. & Makela, R. Nest of juveniles provides evidence of family structure among dinosaurs. *Nature* **282,** 296–298 (1979).
8. Gregory, T. R. *Animal Genome Size Database* ⟨http://www.genomesize.com/⟩ (2005).
9. Van den Bussche, R. A. How bats achieve a small C-value: frequency of repetitive DNA in *Macrotus*. *Mamm. Genome* **6,** 521–525 (1995).
10. Waltari, E. & Edwards, S. V. Evolutionary dynamics of intron size, genome size, and physiological correlates in archosaurs. *Am. Nat.* **160,** 539–552 (2002).
11. Tiersch, T. R. & Wachtel, S. S. On the evolution of genome size of birds. *J. Hered.* **82,** 363–368 (1991).
12. Raup, D. M. *Extinction: Bad Genes or Bad Luck?* 3–21 (W. W. Norton & Company, New York, 1992).
13. Conway Morris, S. & Harper, E. Genome size in conodonts (Chordata): inferred variations during 270 million years. *Science* **241,** 1230–1232 (1988).
14. Masterson, J. Stomatal size in fossil plants: Evidence for polyploidy in majority of angiosperms. *Science* **264,** 421–423 (1994).
15. Thomson, K. S. & Muraszko, K. Estimation of cell size and DNA content in fossil fishes and amphibians. *J. Exp. Zool.* **205,** 315–320 (1978).
16. Gregory, T. R. The bigger the C-value, the larger the cell: Genome size and red blood cell size in vertebrates. *Blood Cells Mol. Dis.* **27,** 830–843 (2001).
17. Pagel, M. D. Inferring evolutionary processes from phylogenies. *Zool. Scr.* **26,** 331–348 (1997).
18. Pagel, M. D. Inferring the historical patterns of biological evolution. *Nature* **401,** 877–884 (1999).
19. Gregory, T. R. & Hebert, P. D. N. The modulation of DNA content: proximate causes and ultimate consequences. *Genome Res.* **9,** 317–324 (1999).
20. Langer, M. C. in *The Dinosauria* (eds Weishampel, D. B., Dodson, P. & Osmólska, H.) 25–46 (Univ. California Press, Berkeley, 2004).
21. Kazazian, H. H. Jr. Mobile elements: Drivers of genome evolution. *Science* **303,** 1626–1632 (2004).
22. Shedlock, A. M. *et al.* Phylogenomics of non-avian reptiles and the structure of the ancestral amniote genome. *Proc. Natl Acad. Sci. USA* **104,** 2767–2772 (2007).
23. Shedlock, A. M. Phylogenomic investigation of CR1 LINE diversity in reptiles. *Syst. Biol.* **55,** 902–911 (2006).
24. Petrov, D. A. Evolution of genome size: new approaches to an old problem. *Trends Genet.* **17,** 23–28 (2001).
25. Kozlowski, J., Konarzewski, M. & Gawelczyk, A. T. Cell size as a link between noncoding DNA and metabolic rate scaling. *Proc. Natl Acad. Sci. USA* **100,** 14080–14085 (2003).
26. Cavalier-Smith, T. in *The Evolution of Genome Size* (ed. Cavalier-Smith, T.) 104–184 (John Wiley & Sons, Chichester, 1985).
27. Szarski, H. Cell size and the concept of wasteful and frugal evolutionary strategies. *J. Theor. Biol.* **105,** 201–209 (1983).

28. Erickson, G. M., Curry-Rogers, K. & Yerby, S. A. Dinosaurian growth patterns and rapid avian growth rates. *Nature* **412,** 429–433 (2001).
29. Padian, K., Horner, J. R. & de Ricqlès, A. J. Growth in small dinosaurs and pterosaurs: the evolution of archosaurian growth strategies. *J. Vert. Paleontol.* **24,** 555–571 (2004).
30. Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. in *Markov Chain Monte Carlo in Practice* (eds Gilks, W. R., Richardson, S. & Spiegelhalter, D. J.) 1–19 (Chapman and Hall, London, 1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the Museum of Comparative Zoology at Harvard University and the Gabriel Laboratory for Cellular and Molecular Paleontology at the Museum of the Rockies for access to histology sections. We also thank D. Smith at the Imaging Center in the Department of Cellular and Molecular Biology, Harvard University for facilitating microscopy, A. Crompton and J. Horner for offering materials, laboratory space, and discussions on palaeohistology, and D. Jablonski and T. Garland for discussions. We are grateful for comments from B. Jennings, N. Hobbs and M. Laurin, which have improved this manuscript. This research was supported by an NIH Postdoctoral Fellowship granted to C.L.O., an NSF grant to S.V.E. and a NERC grant to M.P.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.L.O. (corgan@oeb.harvard.edu).