Supporting Information

Harshman *et al.* 10.1073/pnas.0803242105

SI Methods

DNA Extraction, PCR Amplification, and Sequencing. We assembled a set of frozen tissue samples (Table S2) that were chosen to include all extant ratite genera, to include the basal divergences within tinamous and to represent major evolutionary lineages in neognaths (including members of both Galloanserae and Neoaves). Some analyses also included crocodilian outgroups, for which we selected *Alligator* and *Gavialis*, because these taxa are members of the two major clades within extant crocodilians (1). Because the alternative positions for the root of the avian tree are either within the passerines or between the passerines and all other extant birds (2–4) we also included representatives of the two major passerine lineages (oscines and suboscines). We selected the appropriate avian taxa using multiple lines of evidence, including surveys of the literature (5), analyses of individual loci collected for this project (6, 7), and a large-scale analysis (8). For this large-scale study of paleognaths we selected a total of 20 loci (Table S1) that map to 16 different chicken chromosomes and are therefore unlikely to exhibit linkage in any avian lineage, given the conservation of avian karyotypes (9).

DNAs were isolated by proteinase K digestion, followed by phenol-chloroform extraction according to standard protocols (10, 11) or with DNeasy kits (Qiagen) according to instructions supplied by the manufacturer. The higher yields and purer DNA provided by the first method proved preferable for the amplification of multiple loci from specific taxa. Samples were isolated in one of three laboratories [National Museum of Natural History (NMNH), Field Museum of Natural History (FMNH), and Louisiana State University (LSU)] and distributed to all five laboratories [NMNH, FMNH, LSU, University of Florida (UF), and Wayne State University (WSU)] for amplification and sequencing.

PCR amplification of selected loci was performed by using locus-specific primers designed by members of the participating laboratories or obtained from the literature (Table S1). PCR products were cycle sequenced by using BigDye 3.1 chemistry, and sequences were obtained by using capillary sequencing instruments (both from Applied Biosystems). Some individuals were heterozygous for alleles of different lengths, and the PCR products from those individuals were cloned to obtain clean sequences. Complete information regarding the PCR, cloning, and sequencing conditions for each locus is available upon request from the coauthors listed in Table S1.

Contigs were assembled by using Sequencher (Gene Codes); most (≥95%) nucleotide positions were determined on both strands of DNA. We were unable to recover sequences of certain loci for a few species: *Crypturellus* for *CLTCL1* and *NTF3*; *Nothoprocta* for *MUSK*; *Smithornis* for *TPM1*; and *Pterocnemia* for *CLTCL1*, *EEF2*, *EGR1*, *NTF3*, and *TGFB2*. Nine loci from crocodilians could be amplified, sequenced, and at least partially aligned with bird sequences (*ALDOB*, *BDNF*, *CLTC*, *HMGN2*, *IRF2*, *MYC*, *NTF3*, *PCBD1*, and *RHO*). Introns typically could not be aligned between birds and crocodilians (*IRF2* being the sole exception); the crocodilian intron sequences that could not be aligned were coded as missing data in the data matrix used for analyses.

Sequence Alignment and Phylogenetic Analyses. Preliminary alignments of finished sequences were obtained by using Clustal (12), and then these alignments were subjected to two rounds of manual refinement and cross-checking in different laboratories. Ambiguously aligned regions were identified by a committee with members from multiple laboratories (S.J.H., K.-L.H., R.T.K., B.D.M., S.R., and T.Y.). Sparsely sampled sites (those that were not present in at least four birds and at least three paleognaths) were identified by using a computer program (this program and all noncommercial computer programs and scripts used for this project were written by E.L.B.). These regions of ambiguous sequence alignment and sparsely sampled sites were excluded before analyses. Sequence alignments have been deposited in TreeBase (study accession no. S2138).

For most analyses of paleognath relationships, the four nonpasserine neognaths were used as outgroups (e.g., Fig. 1; see Table S2 for details), and the aligned 14-taxon, 20-gene dataset was 29,509 bp in length, of which 23,902 bp were used in the principal analyses. Crocodilians and passerines were included only in analyses used to test the position of the root of the avian tree (Fig. 2), with the subset of sequences for which crocodilian and avian sequences could be aligned (4,668 bp).

We used PAUP* 4.0b10 (13) to identify optimal trees using the maximum-parsimony (MP) and maximum-likelihood (ML) criteria. MP analyses used branch-and-bound searches (14), with equally weighted characters, and support was assessed by using 1,000 nonparametric bootstrap replicates. Unpartitioned ML analyses in PAUP* used heuristic searches with TBR branch swapping and 10 random addition sequence replicates. Support for ML analyses was assessed by using 100 nonparametric bootstrap replicates, with heuristic searches conducted as described above.

The Akaike information criterion was used to identify the appropriate models of nucleotide substitution for ML analyses (Table S8). The models examined for standard analyses were those used in Modeltest 3.6 (15), whereas the models examined for RY-coded data were a smaller set (eight models) based upon the two-state Neyman/Cavender-Farris (CF) model (16–18). The variants of the CF model that were examined added parameters for unequal state frequencies, Γ-distributed rates across sites, and proportion of invariant sites (each of which adds a single free parameter). We also partitioned the data in various ways to allow application of distinct models to subsets of the data. The primary partitioning scheme was by locus, in which each gene region was allowed to have a different set of evolutionary parameters. However, we also examined partitioning within loci by sequence type (coding exon, intron, or untranslated region) and found that the results were identical to a simple partitioning by locus (results not shown).

We used three different approaches to conduct ML analyses after partitioning the data. The first analytical approach used RAxML (19), which uses a distinct GTR+Γ model for each partition but links the branch length parameters, avoiding the large number of free parameters that would otherwise be necessary. Searches used the GTRMIX model in which the initial search is conducted using the GTR+CAT model, an approximate method to accommodate among-sites rate heterogeneity that is more computationally efficient than GTR+Γ (19), followed by a final optimization assuming Γ-distributed rates. Another advantage of this approach is that RAxML is fast enough to allow us to assess support using 100 nonparametric bootstrap replicates.

The other partitioned ML analyses did not link branch lengths, increasing the number of free parameters, but they were conducted because they allowed a more diverse set of nucleotide-substitution models to be used. Both analyses assumed monophyly of certain well established clades *a priori*, and focused on

a "plausible set" of 315 trees (Fig. S2). The likelihoods of all trees in the plausible set were calculated for each locus in two different programs, PAUP* and nhPhyML (20). Then the log likelihoods were summed to determine the ML tree. Analyses in PAUP* used the best fit model for each locus that was selected by using Modeltest (Table S8). Analyses in nhPhyML used the nonstationary model proposed by Galtier and Gouy (21), typically called the GG98 model, which allows GC-content to vary across the tree. Support in these analyses was assessed by using the RELL (resampling of estimated log likelihoods) bootstrap method (22) with 1,000 bootstrap replicates. The RELL bootstrap method involves bootstrapping the site likelihoods for each tree. Programs that can be used to perform this analysis are available from E.L.B. upon request.

Partitioned and unpartitioned Bayesian Markov chain Monte Carlo analyses were performed by using MrBayes 3.1.1 (23). In each analysis, four chains were run for 10 million generations each (with three chains heated), sampling each 500 generations and discarding the first 500 trees sampled, with other run parameters set at defaults. Partitioned Bayesian analyses used linked branch lengths. Convergence was assessed by using the convergence statistics presented by MrBayes (the standard deviation of partition frequencies and the potential scale reduction factor) and by examining graphs produced by AWTY (24).

The relationship between the support measures we used and the probability that a clade is correct is complex (25, 26). Indeed, a number of different criteria can be used to evaluate the performance of support measures, and the performance of those support measures will depend on the criterion used and the phylogeny being evaluated (26). Nonetheless, the nonparametric bootstrap is known to be conservative under many circumstances, and some studies indicate clades with bootstrap support $\geq$70% have a high probability of being correct (27). The RELL bootstrap has been assumed to exhibit behavior similar to the nonparametric bootstrap when compared with other support measures (e.g., refs. 28 and 29). In contrast to the bootstrap, Bayesian posterior probabilities appear to overestimate the probability that a clade is correct (e.g., refs. 26, 30, and 31). At least some of the differences between the bootstrap and Bayesian posterior probabilities reflect the philosophical framework used to interpret the support value (see ref. 26).

Support was also assessed by examining branch support and hidden support. Branch support, often called Bremer support (32) when used with the MP criterion, is the difference between the scores ($-ln$L for ML or tree length for MP) of the optimal tree with a clade of interest and the optimal tree without that clade. The contribution of individual partitions (which correspond to loci for the analyses reported here) to branch support in a combined ("total evidence") analysis can be assessed by using partitioned branch support (33, 34). Partitioned branch support is the difference in score for a data partition between the optimal tree for all data that contains a clade of interest and the optimal tree for all data without that clade. Partitioned hidden branch support (33, 35) is the difference between the partitioned branch support and the branch support (based upon a separate analysis of the partition of interest). Thus, positive partitioned hidden branch support values are associated with partitions that support specific branches more strongly in a combined analysis than in separate analyses. Hidden branch support (35) is the sum of partitioned hidden branch support values for all data partitions included in the analysis.

Topology Tests and Analyses of Base Composition. We used two different topology tests to examine support for the optimal paleognath topology, both of which are appropriate when the trees compared were not specified *a priori* (36). The first test was the Shimodaira–Hasegawa (SH) test (37), which allowed us to test the null hypothesis that all trees in the plausible set are equally good explanations of the data (36). The SH test was conducted in PAUP* by using 1,000 RELL bootstrap replicates.

The second test was the Swofford–Olsen–Waddell–Hillis (SOWH) test (38), a parametric bootstrap test. Parametric bootstrap tests have been used to detect long-branch attraction in a number of studies (39). The SOWH test compares a test statistic (δ) derived from the empirical data to a null distribution for δ generated by simulation. δ is the difference in optimality scores between the optimal tree and a null-hypothesis tree for any given dataset. The test is conducted by calculating δ for the empirical data by using the optimal tree (Fig. 1) and the best tree containing a specific clade of interest that does not appear in the optimal tree. The null hypothesis for the SOWH test is that the topology used for simulations can explain the observed data as well as the optimal tree. Because the specific clade of interest for this study is the ratite clade, tree 19a, the best tree with ratite monophyly (Fig. S2), was used as the null-hypothesis topology for simulations.

Simulated data corresponding to each partition were generated with Seq-Gen 1.3.2 (40) using the best fit models (Table S8), and then the simulated sequences were concatenated. The SOWH test can be conducted by using either the MP or the ML criteria, with δ corresponding to the tree length difference in the first case and the log likelihood difference in the second case. The empirical data and 1,000 simulated datasets were analyzed with PAUP* by using the MP and ML criteria. For the empirical data, δ was 90 steps for MP and 78.47595 log-likelihood units for ML. The δ values for the empirical data were significantly greater than those observed for the simulated data regardless of whether the criterion used was MP (where the maximum δ value was 43 steps) or ML (where the maximum δ value was 2.25118 log-likelihood units).

To test the potential for long-branch attraction in a direct manner, we performed a total of 1,000 MP analyses in which the outgroup was replaced with a random sequence of similar base composition to simulate the longest possible branch. This strategy has been used in a number of studies (39, 41, 42) because it can reveal which ingroup taxa are most prone to attracting very divergent sequences. Because the individual partitions used here have different base compositions (Table S8), random outgroups with the appropriate base composition for each partition were simulated by using a program written by E.L.B., and the data were then concatenated.

Base-compositional clustering was performed by using constrained searches under the minimum evolution criterion in PAUP* using a matrix of Euclidean distances between vectors of base compositions of each taxon. Topological constraints limited searches to the plausible set (Fig. S2). Base composition was calculated only for variable sites, an approach used in other analyses of base composition (43). Relative composition variability (RCV) (43) was used as a summary statistic to describe differences in base compositional variation among loci.

Tests of Congruence Among Gene Trees. Analyses of individual partitions provide estimates of individual gene trees, which may differ from the species tree. Hypotheses regarding species trees can be examined by determining the probability of observing a specific set of gene trees given a species tree. Even in the absence of phylogenetic signal, many analyses of single genes will find a single tree. We tested whether an extreme case, a completely polytomous species tree, would be likely to yield the observed number of gene trees that show monophyly of non-ostrich paleognaths.

The branches that unite rheas, the cassowary and emu, and tinamous are all long enough (e.g., Fig. 1) for a high probability of coalescence along each of them, ensuring that virtually all loci will contain phylogenetic signal uniting each group. Thus, each of those lineages can be considered individual taxa from the

standpoint of this test. It is therefore necessary to consider only 105 possible topologies (those shown in Fig. S2*C*).

We investigated the degree of agreement among gene trees by performing binomial tests assuming one of two null models. Under the equiprobable trees null model, appropriate if data are essentially randomized and do not contain phylogenetic information, the probability of a tree showing monophyly of non-ostrich paleognaths is 1/7 because 15 of the 105 possible trees have a non-ostrich paleognath clade (shaded in Fig. S2*C*). Under the pure birth (or Yule) null model (44), appropriate when the branches in the species tree are very short and different gene trees reflect random coalescence (45), the probability of a tree with non-ostrich paleognath monophyly is 1/10. This probability reflects the fact that 12 of the 15 trees with this clade have a probability of 1/180 each and the remaining 3 have a probability of 1/90 each (45). A binomial test was used to determine whether the number of analyses in which non-ostrich paleognaths form a clade (i.e., 15 of 20 for MP and 17 of 20 for ML) is consistent with the random selection of trees under either null model.

Identification of Informative Indels. Low-homoplasy indels able to provide information about the polyphyly or monophyly of ratites were identified by coding all gap characters in a 171-taxon, 19-gene dataset (8) by using the simple gap coding method (46) as implemented in SeqState (47). By using PAUP*, these indel characters were mapped on two trees: the optimal tree (which supports monophyly of non-ostrich paleognaths) from a ML analysis of the 171-taxon dataset conducted using GARLI (48) and the same tree rearranged so that ratites are monophyletic. For each tree, all characters that mapped unambiguously on the branch of interest with a consistency index (49) of 0.5 or greater were selected. We then examined the alignments of gap characters that fit the consistency index criterion and removed those for which alignment was ambiguous. This identified the three informative indel characters described in the text. CLTCL1 was not included in the 19-gene dataset (all other genes analyzed here were), so we used a similar methodology to examine that locus independently.

1. Harshman J, Huddleston CJ, Bollback J, Parsons TM, Braun MJ (2003) True and false gharials: A nuclear gene phylogeny of Crocodylia. *Syst Biol* 52:386–402.
2. Härlid A, Arnason U (1999) Analyses of mitochondrial DNA nest ratite birds within the Neognathae: Supporting a neotenous origin of ratite morphological characters. *Proc R Soc London Ser B* 266:305–309.
3. Mindell DP, *et al.* (1999) Interordinal relationships of birds and other reptiles based on whole mitochondrial genomes. *Syst Biol* 48:138–152.
4. Johnson KP (2001) Taxon sampling and the phylogenetic position of Passeriformes: Evidence from 916 avian cytochrome *b* sequences. *Syst Biol* 50:128–136.
5. Harshman J (2007) in *Reproductive Biology and Phylogeny of Birds*, ed Jamieson BGM (Science Publishers, Enfield, NH), pp 1–35.
6. Chojnowski JL, Kimball RT, Braun EL (2008) Introns outperform exons in analyses of basal avian phylogeny using clathrin heavy chain genes. *Gene* 410:89–96.
7. Yuri T, Kimball RT, Braun EL, Braun MJ (2008) Duplication and accelerated evolution of growth hormone gene in passerine birds. *Mol Biol Evol* 25:352–361.
8. Hackett SJ, *et al.* (2008) A phylogenomic study of birds reveals their evolutionary history. *Science* 320:1763–1768.
9. Shetty S, Griffin DK, Graves JAM (1999) Comparative painting reveals strong chromosome homology over 80 million years of bird evolution. *Chromosome Res* 7:289–295.
10. Robbins MB, Braun MJ, Huddleston CJ, Finch DW, Milensky CM (2005) First Guyana records, natural history and systematics of the White-naped Seedeater *Dolospingus fringilloides*. *Ibis* 147:334–341.
11. Mariaux J, Braun MJ (1996) A molecular phylogenetic survey of the nightjars and allies (Caprimulgiformes) with special emphasis on the potoos (Nyctibiidae). *Mol Phylogenet Evol* 6:228–244.
12. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
13. Swofford DL (2003) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4* (Sinauer, Sunderland, MA).
14. Hendy MD, Penny D (1982) Branch and bound algorithms to determine minimal evolutionary trees. *Math Biosci* 60:133–142.
15. Posada D, Crandall KA (1998) MODELTEST: Testing the model of DNA substitution. *Bioinformatics* 14:817–818.
16. Neyman J (1971) in *Statistical Decision Theory and Related Topics*, eds Gupta SS, Yackel J (Academic, New York).
17. Farris JS (1973) A probability model for inferring evolutionary trees. *Syst Zool* 22:250–256.
18. Cavender JA (1978) Taxonomy with confidence. *Math Biosci* 154:1–21.
19. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
20. Boussau B, Gouy M (2006) Efficient likelihood computations with nonreversible models of evolution. *Syst Biol* 55:756–768.
21. Galtier N, Gouy M (1998) Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol* 15:871–879.
22. Kishino H, Miyata T, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol* 31:151–160.
23. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
24. Nylander JAA, Wilgenbusch JC, Warren DL, Swofford DL (2008) AWTY (Are We There Yet?): A system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics* 24:581–583.
25. Felsenstein J (2004) *Inferring Phylogenies* (Sinauer, Sunderland, MA).
26. Alfaro ME, Holder MT (2007) The posterior and the prior in Bayesian phylogenetics. *Annu Rev Ecol Evol Syst* 37:19–42.
27. Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst Biol* 42:182–192.
28. Rannala B, Yang Z (1996) Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J Mol Evol* 43:304–311.
29. Cao Y, Adachi J, Hasegawa M (1998) Comment on the quartet puzzling method for finding maximum-likelihood tree topologies. *Mol Biol Evol* 15:87–89.
30. Suzuki Y, Glazko GV, Nei M (2002) Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc Natl Acad Sci USA* 99:16138–16143.
31. Cummings MP, *et al.* (2003) Comparing bootstrap and posterior probability values in the four-taxon case. *Syst Biol* 52:477–487.
32. Bremer K (1988) The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution (Lawrence, Kans)* 42:795–803.
33. Lee MSY, Hugall AF (2003) Partitioned likelihood support and the evaluation of data set conflict. *Syst Biol* 52:15–22.
34. Baker R, DeSalle R (1997) Multiple sources of character information and the phylogeny of Hawaiian drosophilids. *Syst Biol* 46:654–673.
35. Gatesy J, O'Grady P, Baker RH (1999) Corroboration among data sets in simultaneous analysis: Hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics* 15:271–313.
36. Goldman N, Anderson JP, Rodrigo AG (2000) Likelihood-based tests of topologies in phylogenetics. *Syst Biol* 49:652–670.
37. Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* 16:1114–1116.
38. Swofford DL, Olson GJ, Waddell PJ, Hillis DM (1996) in *Molecular Systematics*, eds Hillis DM, Moritz C, Mable BK (Sinauer, Sunderland, MA), pp 407–514.
39. Bergsten J (2005) A review of long-branch attraction. *Cladistics* 21:163–193.
40. Rambaut A, Grassly NC (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *CABIOS* 13:235–238.
41. Sullivan J, Swofford DL (1997) Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J Mamm Evol* 4:77–86.
42. Graham SW, Olmstead RG, Barrett SCH (2002) A case study from the commelinoid monocots. *Mol Biol Evol* 19:1769–1781.
43. Phillips MJ, Penny D (2003) The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol Phylogenet Evol* 28:171–185.
44. Yule GU (1924) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. *Philos Trans R Soc London Ser B* 213:21–87.
45. Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. *PLoS Genet* 2:e68.
46. Simmons MP, Ochoterena H (2000) Gaps as characters in sequence-based phylogenetic analyses. *Syst Biol* 49:369–381.
47. Müller K (2006) Incorporating information from length-mutational events into phylogenetic analysis. *Mol Phylogenet Evol* 38:667–676.
48. Zwickl D (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological datasets under the maximum likelihood criterion. PhD dissertation. (Univ of Texas, Austin, TX).
49. Kluge AG, Farris JS (1969) Quantitative phyletics and the evolution of anurans. *Syst Zool* 18:1–32.

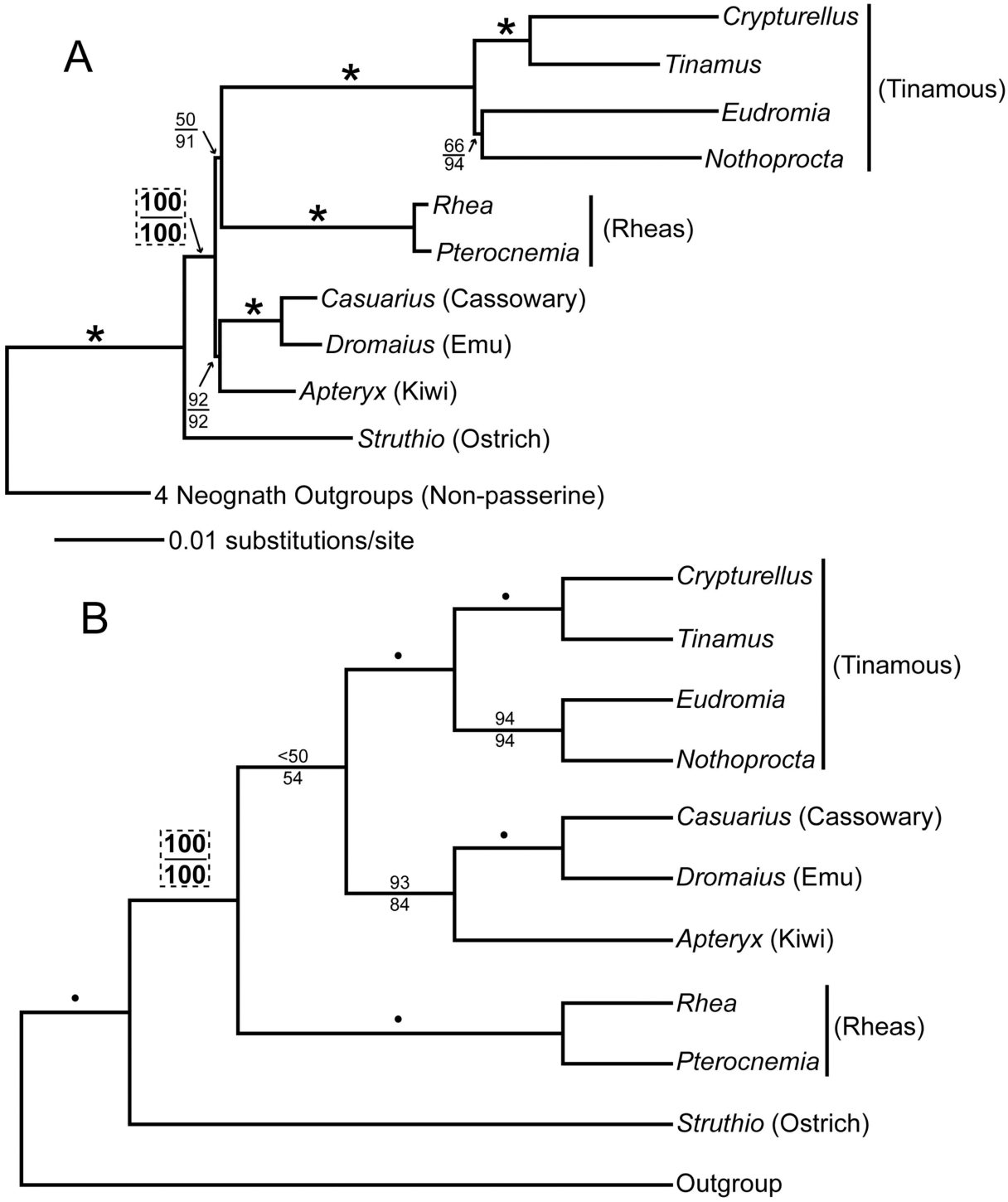

Fig. S1. Additional phylogenetic analyses of the complete nuclear dataset supporting ratite polyphyly. All analyses used *Anas*, *Gallus*, *Buteo*, and *Ciconia* as outgroups. Branches for which bootstrap support was 100% in both analyses are indicated with an asterisk; support for ratite polyphyly is highlighted. (*A*) Topology obtained by using ML with RY-coding and MP. Branch lengths were estimated in the RY-coded ML analysis. Support measures are the ML bootstrap for RY-coded data (*Upper*) and the MP bootstrap (*Lower*). (*B*) Optimal topology for partitioned ML with unlinked branch lengths (partitioned by locus and presented as a cladogram because branch length estimates vary among loci). Constrained branches (that restrict trees to the ''plausible set'' in Fig. S2) are indicated with dots above branches. The optimal topology was identical for the PAUP* analyses, that used the best fitting model for each partition (Table S8), and the nhPhyML analysis, that used the nonstationary GG98 model (1) with the ML estimates of parameters for each locus. Support measures are the percentage of 1000 RELL bootstrap replicates for the PAUP* (*Upper*) and nhPhyML (*Lower*) analyses.

1. Galtier N, Gouy M (1998) Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol* 15:871–879.

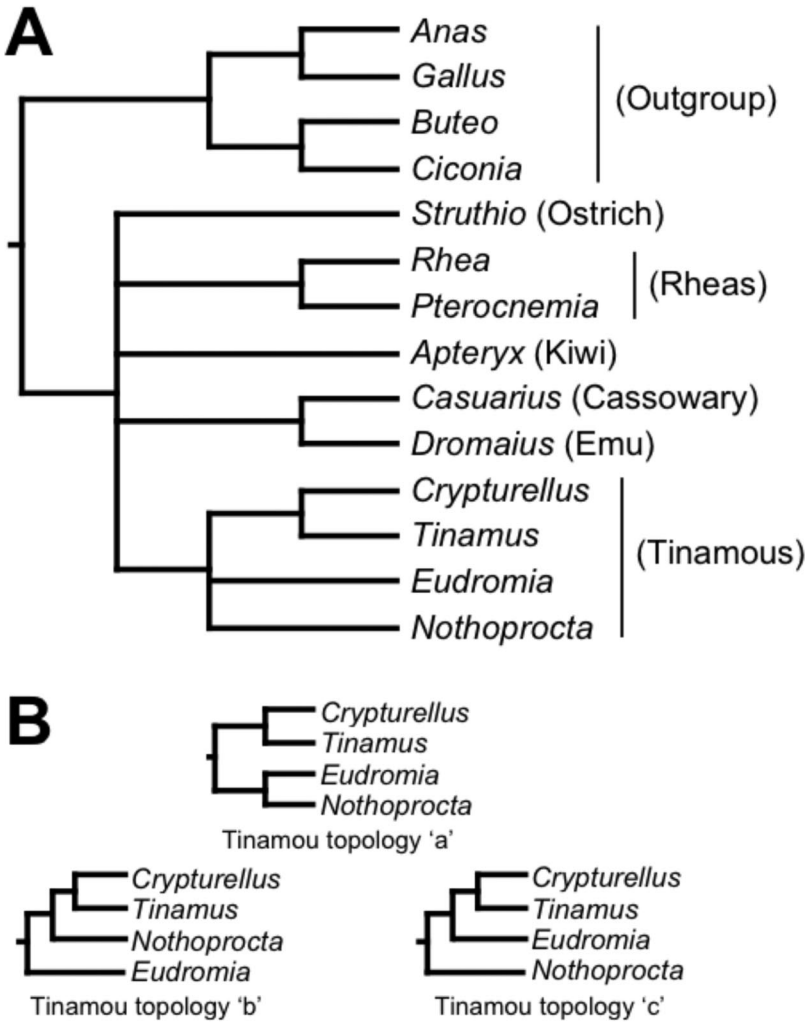

Fig. S2. The plausible set of trees and Shimodaira-Hasegawa test results. (*A*) Cladogram showing the relationships used to define the plausible set of paleognath topologies. The only relationships resolved are those that can be considered fixed *a priori*. There are five major ingroup lineages (ostrich, rheas, kiwis, cassowary + emu, and tinamous) and the plausible set includes all possible arrangements of these taxa (105 topologies). All groups assumed to be monophyletic are uncontroversial based upon a review of the literature (1), and all received strong support in our unpartitioned analyses (e.g., Fig. 1, Fig. S1). (*B*) In addition to the major ingroup lineages, three arrangements within tinamous are plausible *a priori*. All of these arrangements must be considered because all plausible arrangements within the major groups must be included in the plausible set when the SH test is conducted (2). The only other major group containing more than two members is the outgroup (neognaths), but the topology used is uncontroversial (1) so it was fixed in these analyses. Thus, we actually tested 315 trees (three sets of 105 topologies each). (*C*) The complete set of arrangements for major ingroup lineages (105 topologies). Only the 15 shaded topologies, all of which have ostrich sister to all other paleognaths, could not be rejected. *P* values for the shaded topologies are shown (using ''a'', ''b'', and ''c'' to indicate different tinamou topologies). The SH test rejected all other trees in the plausible set with $P < 0.001$, including those with ratite monophyly (topologies 16, 18, 19, 22, 23, 26, 27, 28, 29, 33, 41, 44, 45, 46, and 48). Figure continues on the next three pages.

1. Harshman J (2007) in *Reproductive Biology and Phylogeny of Birds*, ed Jamieson BGM (Science Publishers, Enfield, NH), pp 1–35.
2. Goldman N, Anderson JP, Rodrigo AG (2000) Likelihood-based tests of topologies in phylogenetics. *Syst Biol* 49:652–670.

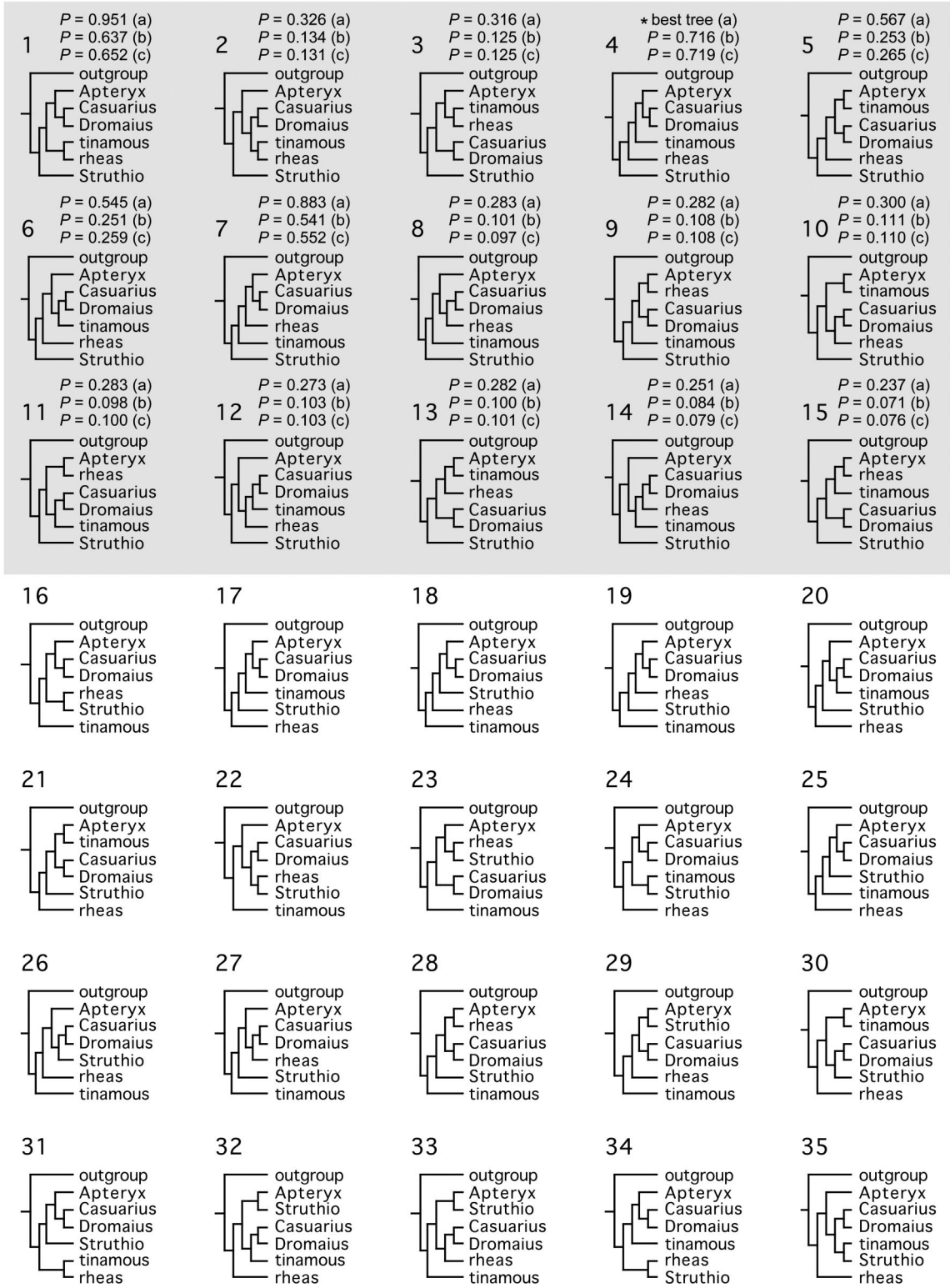

Fig. S2 (continued).

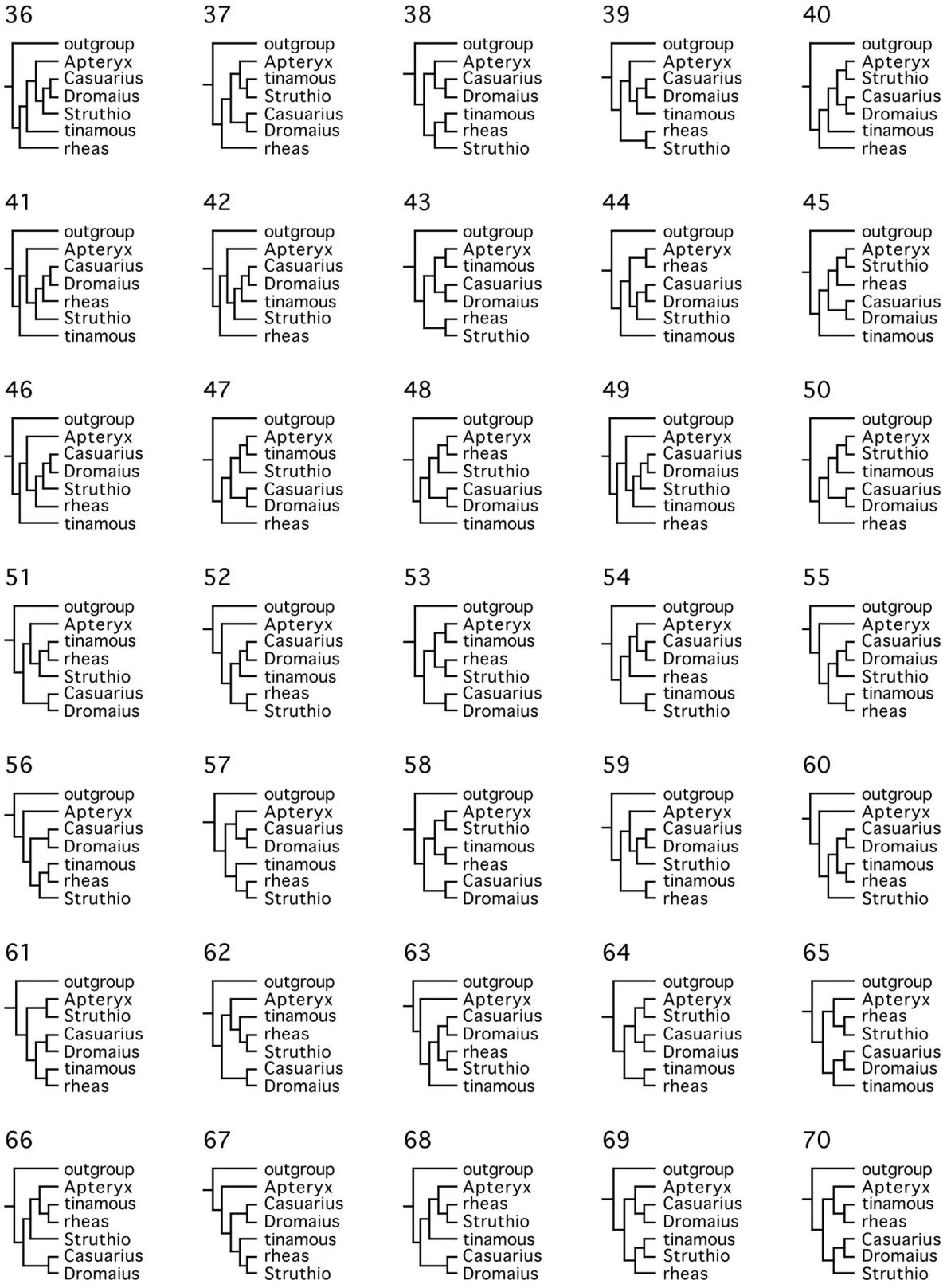

Fig. S2. (continued).

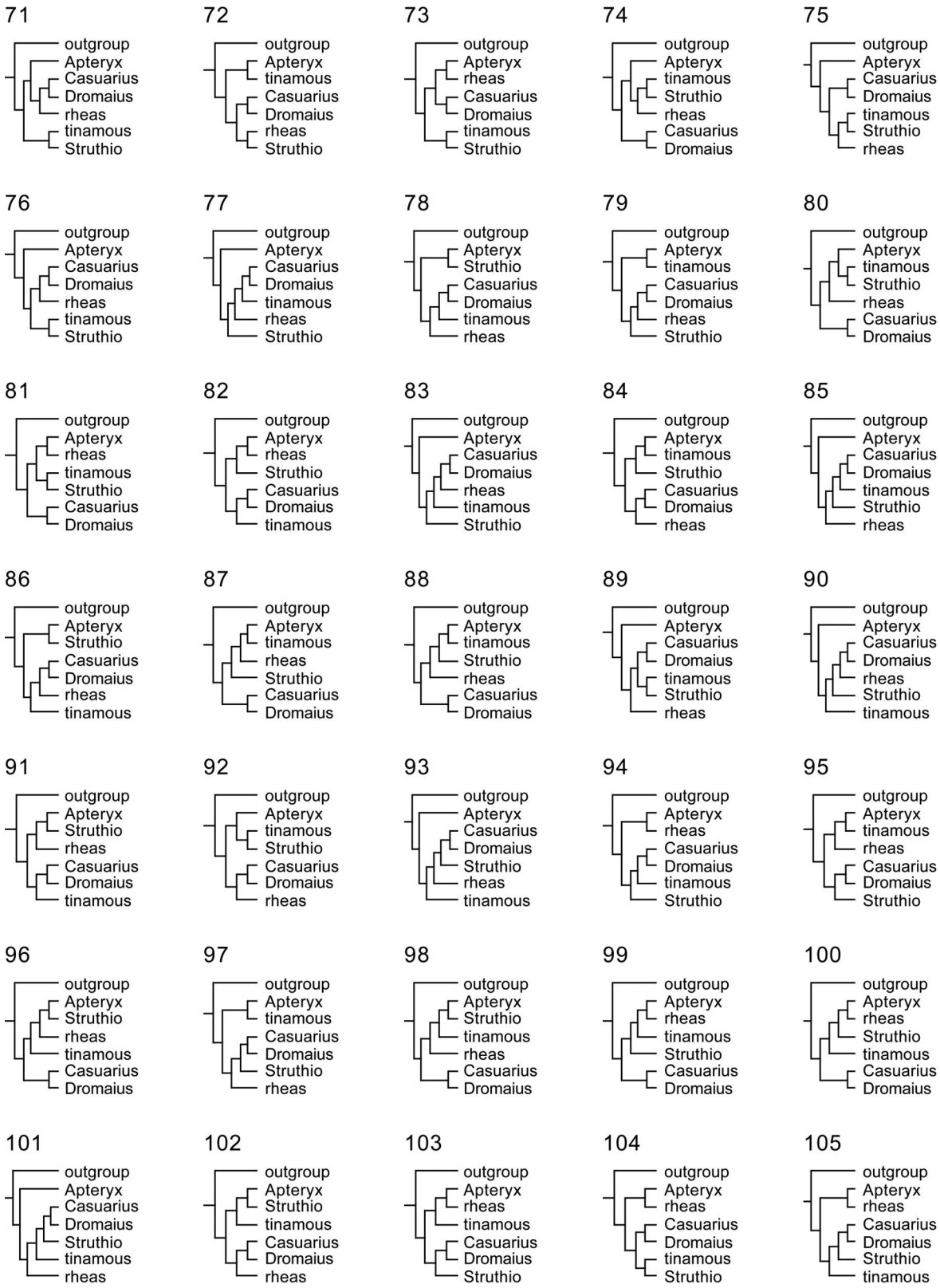

Fig. S2. (continued).

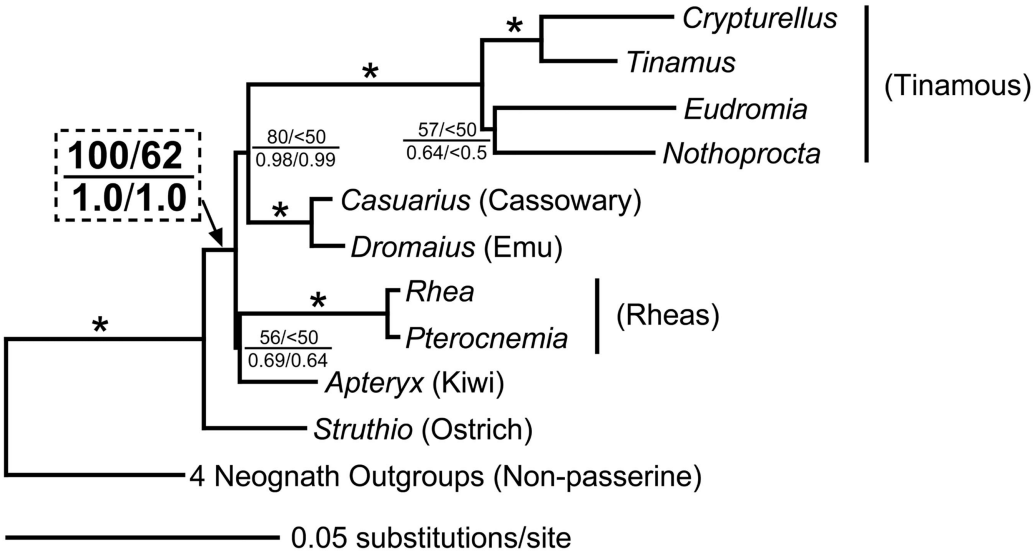

Fig. S3. Phylogenetic analyses of coding exons supporting ratite polyphyly. All analyses used *Anas*, *Gallus*, *Buteo*, and *Ciconia* as outgroups. Branches for which all support measures were 100% or 1.0 are indicated with *; support for ratite polyphyly is highlighted. Branch lengths reflect an unpartitioned ML analysis using the GTR+I+Γ model (for parameter estimates, see Table S8). Support measures are: unpartitioned ML bootstrap (*Upper Left*) ML bootstrap (*Upper Right*), unpartitioned Bayesian posterior probability (*Lower Left*), and partitioned Bayesian posterior probability (*Lower Right*).

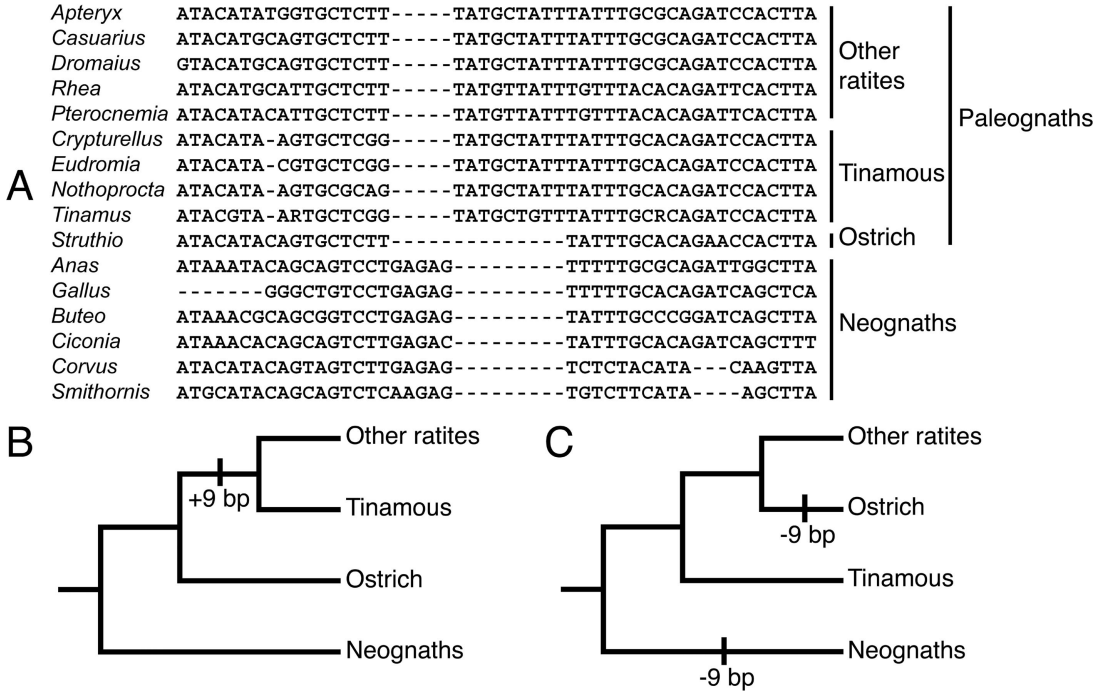

Fig. S4. A 9-bp insertion in *MYC* supporting ratite polyphyly. (*A*) Alignment of region around the informative insertion in *MYC*, positions 298–306 in the *MYC* alignment (available in TreeBase under study accession number S2138). The ostrich (*Struthio*) shares its character state (−9 bp) with neognaths, whereas tinamous share the character state of all other ratites (+9 bp). (*B*) The distribution of character states can be mapped as a single insertion on the optimal topology found in our study. (*C*) The distribution requires at least two steps on the more conventional topology (one possible reconstruction shown). An adjacent 5-bp indel can be mapped as an insertion in neognaths or a deletion in paleognaths but is equally parsimonious on either tree.

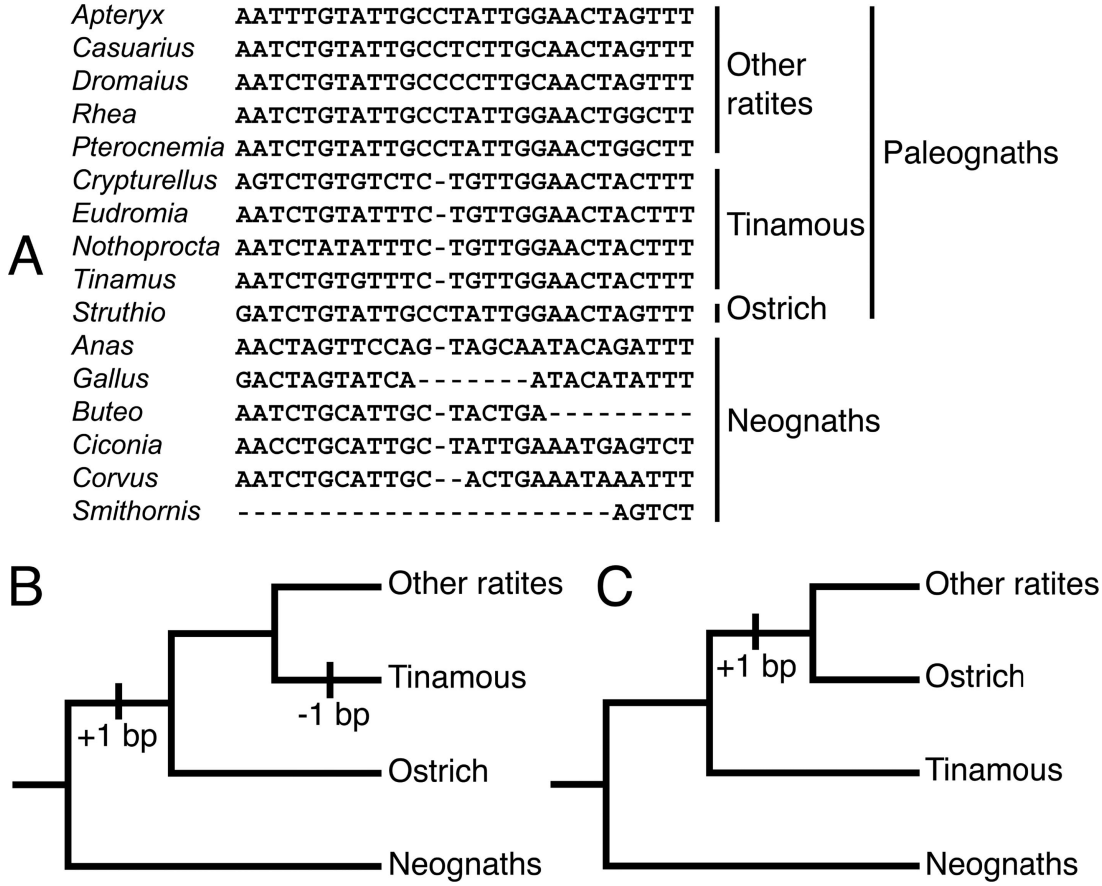

Fig. S5. A 1-bp insertion in *CLTC* supporting ratite monophyly. (*A*) Alignment of region around the informative 1-bp indel in *CLTC*, at position 1909 of the *CLTC* alignment (available in TreeBase under study accession number S2138). The ostrich shares its character state (+1 bp) with other ratites, whereas tinamous share the character state of many neognaths (−1 bp). (*B*) The distribution of character states requires at least two steps on the optimal topology found in our study (one possible reconstruction shown). (*C*) The distribution can be mapped as a single insertion on the more conventional topology.

Table S1. Gene regions used for this study

Gene*	Description	Chromosome†	Median length	Target segment‡	Source§
ALDOB	Fructose-bisphosphate Aldolase B	Z	2180	Introns 3–7	R.T.K. and E.L.B.; (1)
BDNF	Brain-derived Neurotrophic Factor	5	690	Exon 1	(2)
CLTC	Clathrin, Heavy Polypeptide	19	1850	Introns 6–7	(3)
CLTCL1	Clathrin, Heavy-Polypeptide-Like 1	15	770	Intron 7	(3)
CRYAA	α-A-Crystallin	1	1210	Intron 1	R.T.K. and E.L.B.
EEF2	Eukaryotic Translation Elongation Factor 2	28	3340	Introns 4–7	R.T.K. and E.L.B.
EGR1¶	Early Growth Response 1	13	1710	Exon, 3′ UTR	(4)
FGB¶	Fibrinogen Beta Chain	4	2720	Introns 4–7	R.C.K.B.; K.J.M.; (5)
GH1	Growth Hormone 1 (Somatotropin)	27	1240	Introns 2–3	(6)
HMGN2	Nonhistone Chromosomal Protein HMG-17	23	1730	Introns 2–5	R.T.K. and E.L.B.; (1)
IRF2	Interferon Regulatory Factor 2	4	610	Intron 2	K.J.M.
MB	Myoglobin	1	720	Intron 2	(7, 8)
MUSK	Skeletal Muscle Receptor Tyrosine Kinase	Z	590	Intron 3	F. K. Barker
MYC	c-Myc Proto-oncogene Homolog	2	1250	Intron 2, 3′ UTR	W. Holznagel; (9)
NGF	Nerve Growth Factor, β polypeptide	26	740	Exon 4	(2)
NTF3	Neurotrophin 3	1	730	Exon 3	(2)
PCBD1	Pterin-4α-Carbinolamine Dehydratase	6	1020	Introns 2–3	R.T.K. and E.L.B.
RHO	Rhodopsin	12	1890	Intron 1–3	R.T.K. and E.L.B.; (1)
TGFB2	Transforming Growth Factor β2	3	580	Intron 5	(10)
TPM1	Tropomyosin 1 (alpha)	10	460	Intron 6	(11)
Total			26030		

*Gene symbol used in Entrez, which is identical to the HUGO gene symbols (12, 13). Many genes have additional symbols (e.g., *ZENK* for *EGR1*, *DCOH* for *PCBD1*, and *β-fibint7* for *FGB* intron 7).

†Chromosomal location in chickens.

‡Major elements included in the fragment; many primer pairs amplify small amounts of additional sequence (e.g., amplicons generated to sequence individual introns typically include a small part of flanking exons).

§Designers of the PCR primers or the publications describing them. Complete information regarding the positions of all primers and their robustness in a variety of avian lineages has been assembled (R.T.K., unpublished work); additional information is available from R.T.K. (rkimball@ufl.edu).

¶Several nonoverlapping amplicons were generated for *EGR1* and *FGB*; all other gene regions could be assembled into a single contig that includes all amplified segments.

1. Cox WA, Kimball RT, Braun EL (2007) Phylogenetic position of the New World quail (Odontophoridae): Eight nuclear loci and three mitochondrial regions contradict morphology and the Sibley–Ahlquist Tapestry. *Auk* 124:71–84.
2. Sehgal RNM, Lovette IJ (2003) Molecular evolution of three avian neurotrophin genes: Implications for proregion functional constraints. *J Mol Evol* 57:335–342.
3. Chojnowski JL, Kimball RT, Braun EL (2008) Introns outperform exons in analyses of basal avian phylogeny using clathrin heavy chain genes. *Gene* 410:89–96.
4. Chubb AL (2004) New nuclear evidence for the oldest divergence among neognath birds: The phylogenetic utility of *ZENK* (i). *Mol Phylogenet Evol* 30:140–151.
5. Driskell AC, Christidis L (2004) Phylogeny and evolution of the Australo–Papuan honeyeaters (Passeriformes, Meliphagidae). *Mol Phylogenet Evol* 31:942–960.
6. Yuri T, Kimball RT, Braun EL, Braun MJ (2008) Duplication and accelerated evolution of growth hormone gene in passerine birds. *Mol Biol Evol* 25:352–361.
7. Heslewood MM, Elphinstone MS, Tidemann SC, Baverstock PR (1998) Myoglobin intron variation in the Gouldian Finch *Erythrura gouldiae* assessed by temperature gradient gel electrophoresis. *Electrophoresis* 19:142–151.
8. Slade RW, Moritz C, Heideman A, Hale PT (1993) Rapid assessment of single-copy nuclear DNA variation in diverse species. *Mol Ecol* 2:359–373.
9. Harshman J, Huddleston CJ, Bollback J, Parsons TM, Braun MJ (2003) True and false gharials: A nuclear gene phylogeny of Crocodylia. *Syst Biol* 52:386–402.
10. Primmer CR, Borge T, Lindell J, Saetre G-P (2002) Single-nucleotide polymorphism (SNP) characterization in species with limited available sequence information: High nucleotide diversity revealed in the avian genome. *Mol Ecol* 11:603–612.
11. Friesen VL, Congdon BC, Kidd MG, Birt TP (1999) Polymerase chain reaction (PCR) primers for the amplification of five nuclear introns in vertebrates. *Mol Ecol* 8:2141–2152.
12. Maglott D, Ostell J, Pruitt KD, Tatusova T (2005) Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res* 33:D54–D58.
13. Bruford EA, *et al.* (2008) The HGNC database in 2008: A resource for the human genome. *Nucleic Acids Res* 36:D445–D448.

Table S2. Taxa and DNA samples

Group	Species	Common name*	Institution†	Voucher or tissue number	Collector
Paleognaths					
	Apteryx australis	Southern brown kiwi	LSUMNS	B8606	Captive
	Casuarius casuarius	Southern cassowary	LSUMNS	B10202	Captive
	Dromaius novaehollandiae	Emu	LSUMNS	B5895	Captive
	Rhea americana	Greater rhea	USNM	541231	D. Johnston
	Pterocnemia pennata	Darwin's rhea	USNM	620827	Captive
	Struthio camelus	Common ostrich	LSUMNS	B1526	Captive
	Crypturellus soui	Little tinamou	USNM	586295	K. S. Bostwick
	Eudromia elegans	Elegant crested tinamou	LSUMNS	B5893	Captive
	Nothoprocta perdicaria	Chilean tinamou	LSUMNS	B23841	Captive
	Tinamus guttatus	White-throated tinamou	FMNH	389673	B. D. Patterson
Nonpasserine neognaths					
	Anas platyrhynchos	Mallard	FMNH	398624	T. Valentino
	Gallus gallus	Red junglefowl	LSUMNS	B19438	Captive
	Buteo jamaicensis	Red-tailed hawk	LSUMNS	B33264	A. Aleixo
	Ciconia ciconia	White stork	KU	90088	Captive
Passerine neognaths					
	Corvus corone	Carrion crow	USNM	612224	B. K. Schmidt
	Smithornis rufolateralis	Rufous-sided broadbill	FMNH	429425	D. E. Willard
Crocodilians					
	Alligator mississippiensis	American alligator	LSU HSC	none	H. Dessauer
	Gavialis gangeticus	Gharial	LDD	1001871	L. Densmore

*Common names of avian species are taken from ref. 1.

†FMNH, Field Museum of Natural History; KU, University of Kansas; LDD, Collection of Llewellyn D. Densmore III, Texas Tech University; LSU HSC, Louisiana State University Health Science Center; LSUMNS, Louisiana State University Museum of Natural Science; USNM, United States National Museum.

1. Gill F, Wright M (2006) *Birds of the world: Recommended English Names* (Princeton Univ Press, Princeton).

Table S3. Branch support and hidden support for ratite polyphyly

	Maximum likelihood (ML)			Maximum parsimony (MP)		
Locus	Branch support*	Partitioned branch support†	Partitioned hidden branch support‡	Branch support*	Partitioned branch support†	Partitioned hidden branch support‡
ALDOB	2.33584	4.91546	2.57962	3	0	−3
BDNF	−1.56771	0	1.56771	−6	−7	−1
CLTC	0.15142	0.15936	0.00794	−3	12	15
CLTCL1	3.73463	5.22235	1.48772	0	0	0
CRYAA	2.24095	2.28447	0.04352	3	−2	−5
EEF2	11.32733	11.31118	−0.01615	8	7	−1
EGR1	4.21178	4.83997	0.62819	1	4	3
FGB	−0.14579	0.11616	0.26195	3	5	2
GH1	1.18155	2.31122	1.12967	2	7	5
HMGN2	4.03957	4.03258	−0.00699	1	9	8
IRF2	1.25499	2.52369	1.2687	3	6	3
MB	7.73518	9.73489	1.99971	3	5	2
MUSK	9.38987	9.78853	0.39866	6	9	3
MYC	5.53331	5.66907	0.13576	0	1	1
NGF	2.56123	3.05734	0.49611	2	2	0
NTF3	5.63971	6.11189	0.47218	2	5	3
PCBD1	7.21264	9.61173	2.39909	4	10	6
RHO	2.96495	2.92965	−0.0353	5	8	3
TGFB2	−0.98119	−0.47666	0.50453	0	0	0
TPM1	0.35757	1.15056	0.79299	1	2	1
Hidden branch support§			16.11561			45

*Branch support is the difference in score (for separate analyses of individual loci) between the optimal tree containing the branch of interest and the optimal tree without that branch. The branch of interest here unites non-ostrich paleognaths. For MP, branch support is typically called Bremer support. ML scores were taken from Table S9.

†Partitioned branch support is the difference in score, for that locus, between the optimal tree for all data containing the branch of interest and the optimal tree for all data without that branch.

‡Partitioned hidden branch support is partitioned branch support minus branch support .

§Hidden branch support is the sum of partitioned hidden branch support values for all loci.

Table S4. Relative compositional variability (RCV) and evolutionary rates within paleognaths

Locus	RCV*	Tinamou Rate†	
		Ostrich	Non-ostrich ratites
ALDOB	0.057	2.12	2.91
BDNF	0.355	5.93	10.68
CLTC	0.046	2.01	2.23
CLTCL1	0.065	4.65	2.91
CRYAA	0.125	3.19	2.77
EEF2	0.056	2.72	2.59
EGR1	0.115	3.67	1.52
FGB	0.036	1.97	2.29
GH1	0.071	2.65	2.69
HMGN2	0.068	2.29	1.78
IRF2	0.076	1.84	2.91
MB	0.071	2.74	2.49
MUSK	0.067	2.37	1.91
MYC	0.102	3.29	2.54
NGF	0.139	11.39	4.51
NTF3	0.080	3.83	2.91
PCBD1	0.110	4.31	3.02
RHO	0.086	2.94	1.95
TGFB2	0.045	2.71	2.50
TPM1	0.075	4.70	2.73
Minimum	0.036	1.84	1.52
Median	0.073	2.84	2.64
Maximum	0.355	11.39	10.68

*Relative composition variability (RCV) reflects the average differences among taxa in base composition. Higher values indicate greater variation in base composition.

†Rates for tinamous are presented relative to the ostrich and to non-ostrich ratites. The values reported are the mean patristic distance from the base of the paleognaths to each of the tinamous divided by (*i*) the distance from the base of the paleognaths to ostrich or (*ii*) the mean distance from the base of the paleognaths to each non-ostrich ratite.

Table S5. Clustering of taxa by base compositional similarity

Gene	Non-ostrich paleognaths clustered?	Ratites clustered?	Topology*
ALDOB			64a
BDNF		Yes	23a
CLTC			64a
CLTCL1		Yes	22c
CRYAA		Yes	22c
EEF2	Yes		15a
EGR1		Yes	33b
FGB		Yes	29b
GH1			68c
HMGN2		Yes	18b
IRF2			57a
MB	Yes		15a
MUSK		Yes	22
MYC		Yes	45b
NGF			94c
NTF3		Yes	22a
PCBD1	Yes		14b
RHO			63b
TGFB2		Yes	28a
TPM1	Yes		8a
Number of loci	4†	10‡	

If our result (monophyly of non-ostrich paleognaths) were an artifact of base compositional convergence, we would expect non-ostrich paleognaths to cluster together more often than predicted by chance. Instead, ratites clustered together more often than expected by chance given the equiprobable trees null model.
*Topology numbers refer to the 105 arrangements of major paleognath lineages (Fig. S2*C*). Letters refer to the topology within the tinamous (Fig. S2*B*). The topology within tinamous is not provided for *MUSK* because this locus was not sequenced from *Nothoprocta*.
†Binomial test, $P = 0.318$ under the equiprobable model, not significant.
‡Binomial test, $P = 0.000164$ under the equiprobable model.

Table S6. ML analyses using a method that accommodates variable base composition (the GG98 model)

Gene	Non-ostrich paleognaths monophyletic?	Ratites monophyletic?	Topology*
ALDOB	Yes		9a
BDNF		Yes	33a
CLTC			20a
CLTCL1	Yes		12c
CRYAA	Yes		6a
EEF2	Yes		4c
EGR1	Yes		6c
FGB	Yes		5a
GH1	Yes		7a
HMGN2	Yes		4c
IRF2			102b
MB	Yes		8b
MUSK	Yes		10
MYC	Yes		6a
NGF	Yes		11c
NTF3	Yes		1c
PCBD1	Yes		3b
RHO	Yes		4c
TGFB2			85a
TPM1	Yes		13a
Number of loci	16[†]	1	

*Topology numbers refer to the 105 arrangements of major paleognath lineages (Fig. S2*C*). Letters refer to the topology within the tinamous (Fig. S2*B*). The topology within tinamous is not provided for *MUSK* because this locus was not sequenced from *Nothoprocta*.

[†]Binomial test, $P = 3 \times 10^{-12}$ under the equiprobable trees model.

Table S8. ML parameter estimates

	Base frequencies†				Substitution rates (G-T = 1)‡					Rate heterogeneity§		
Partition*	A	C	G	T	A-C	A-G	A-T	C-G	C-T	Proportion invariant	Gamma shape	Model¶
ALDOB	0.2812	0.2316	0.2044	0.2828	1.0850	3.6773	0.7244	1.3099	4.3239	0	1.5031	GTR+Γ
BDNF	0.2544	0.2758	0.2900	0.1798	1.0131	4.8637	0.1840	1.7550	4.8637	0	0.1986	TVM+Γ
CLTC	0.2902	0.1941	0.2133	0.3025	1.0489	6.1262	0.8733	1.6294	5.2612	0.2025	4.9477	GTR+ I+Γ
CLTCL1	0.2630	0.2129	0.2375	0.2866	1	3.8110	1	1	3.8110	0.3318	—	HKY+ I
CRYAA	0.25	0.25	0.25	0.25	1	5.8274	1	1	5.8274	0	1.4295	K80+Γ
EEF2	0.2353	0.2189	0.2695	0.2764	1.1205	5.3185	0.7969	1.4086	5.3185	0.2540	4.9219	TVM+ I+Γ
EGR1	0.2573	0.2670	0.1909	0.2848	1	7.0056	1	1	4.8892	0	0.4924	TrN+Γ
FGB	0.3166	0.1755	0.1964	0.3114	1.1390	4.2738	0.7481	1.4706	4.2738	0.1338	4.9335	TVM+ I+Γ
GH1	0.25	0.25	0.25	0.25	1	4.1428	1	5.1094	1	0.1703	4.6444	TrNef+ I+Γ
HMGN2	0.2756	0.1731	0.2277	0.3235	1.2069	4.9597	0.4812	1.7360	4.9597	0	1.3422	TVM+Γ
IRF2	0.2539	0.1997	0.2041	0.3423	1	6.7780	0.6785	0.6785	4.6683	0	1.2063	TIM+Γ
MB	0.2905	0.2273	0.2301	0.2521	1	4.2549	1	1	5.5920	0	1.8262	TrN+Γ
MUSK	0.3065	0.1822	0.1963	0.3150	1.3217	3.8674	0.6343	1.5679	3.8674	0	1.9896	TVM+Γ
MYC	0.2823	0.2405	0.2412	0.2361	1	7.0537	1	1	7.0537	0.3566	0.5575	HKY+ I+Γ
NGF	0.2511	0.3002	0.2588	0.1899	1.8674	9.0033	1.4265	2.7837	9.0033	0	0.4273	TVM+Γ
NTF3	0.3330	0.2084	0.2162	0.2423	1	6.9305	1	1	6.9305	0	0.2681	HKY+Γ
PCBD1	0.2558	0.2721	0.2812	0.1910	0.7719	3.3646	0.6402	0.9722	4.5610	0	1.4515	GTR+Γ
RHO	0.1829	0.2858	0.2998	0.2315	1.2416	5.0932	0.7122	1.2601	5.0932	0.1935	1.9785	TVM+ I+Γ
TGFB2	0.2529	0.2110	0.2272	0.3089	1.1232	4.8241	0.5933	1.3873	3.8128	0	4.2516	GTR+Γ
TPM1	0.2265	0.2399	0.1618	0.3718	1	9.3507	0.5512	0.5512	2.9721	0	0.6509	TIM+Γ
Combined (Fig. 1)	0.2683	0.2252	0.2319	0.2746	1.0429	4.7124	0.7337	1.3376	4.7124	0.1977	2.0923	TVM+ I+Γ
Exons (Fig. S3)	0.2811	0.2531	0.2510	0.2148	1.3307	6.8480	1.0232	1.6381	10.0596	0.4690	0.4690	GTR+ I+Γ
Crocodilians (Fig. 2)	0.2823	0.2363	0.2427	0.2387	1.1772	6.2171	0.9567	1.4486	8.8047	0.3475	0.7377	GTR+ I+Γ

Models were chosen using ModelTest using the Akaike information criterion

*The names of the individual loci; ''combined'' refers to the concatenation of all loci, ''exons'' refers to the exons only dataset, and ''crocodilians'' refers to the subset of the combined dataset that could be reliably aligned with the crocodilian outgroup.

†Estimated base frequency parameters, which may differ from empirical base frequencies.

‡Substitution rates for each pair of bases, relative to the G-T rate.

§Rate heterogeneity parameters. Proportion invariant, the estimated proportion of invariant sites; Gamma shape, estimated value of α, the shape parameter of the Γ distribution used to model among-sites rate variation.

¶Name of the best-fit model using the terminology of ModelTest.

Other Supporting Information Files

Table S7 (XLS)
Table S9 (XLS)