Syst. Biol. 51(4):599–613, 2002
DOI: 10.1080/10635150290102285

A Genomic Schism in Birds Revealed by Phylogenetic Analysis of DNA Strings

SCOTT V. EDWARDS,[1] BERNARD FERTIL,[2] ALAIN GIRON,[2] AND PATRICK J. DESCHAVANNE[2]

[1]*Department of Zoology and Burke Museum, University of Washington, Box 351800, Seattle, Washington 98195, USA; E-mail: sedwards@u.washington.edu*
[2]*INSERM U 494, 91 bd de l'Hopital, 75634 Paris, France*

Abstract.—The molecular systematics of vertebrates has been based entirely on alignments of primary structures of macromolecules; however, higher order features of DNA sequences not used in traditional studies also contain valuable phylogenetic information. Recent molecular data sets conflict over the phylogenetic placement of flightless birds (ratites - paleognaths), but placement of this clade critically influences interpretation of character change in birds. To help resolve this issue, we applied a new bioinformatics approach to the largest molecular data set currently available. We distilled nearly one megabase (1 million base pairs) of heterogeneous avian genomic DNA from 20 birds and an alligator into genomic signatures, defined as the complete set of frequencies of short sequence motifs (strings), thereby providing a way to directly compare higher order features of nonhomologous DNA sequences. Phylogenetic analysis and principal component analysis of the signatures strongly support the traditional hypothesis of basal ratites and monophyly of the nonratite birds (neognaths) and imply that ratite genomes are linguistically primitive within birds, despite their base compositional similarity to neognath genomes. Our analyses show further that the phylogenetic signal of genomic signatures are strongest among deep splits within vertebrates. Despite clear problems with phylogenetic analysis of genomic signatures, our study raises intriguing issues about the biological and genomic differences that fundamentally differentiate paleognaths and neognaths. [Bioinformatics; CpG island; genomics; isochore; ratite.]

The phylogenetic analysis of the primary structure (sequence) of DNA has matured in recent years to encompass a wide variety of techniques, including incorporation of secondary structures of RNA and proteins to improve alignment and tree building (Suyama et al., 1997; Schöniger and von Haeseler, 1999). Most of these methods rely on or produce alignments of primary structures for assigning homology to individual sites prior to or during phylogenetic analysis (Mindell and Meyer, 2001). It is less well appreciated that homology exists in DNA sequences at organizational levels higher than the individual DNA site and that "nonhomologous" DNA sequences that are not alignable by normal criteria can also contain phylogenetic information of use to systematists (Karlin and Burge, 1995; Karlin et al., 1997; Schneider, 1997). For example, distant relationships among proteins whose primary structures are unalignable can sometimes be found by examining secondary structures (Bullock et al., 1996; Matsuo et al., 1996) or hydrophobicity profiles (Leunissen and de Jong, 1986; Naylor et al., 1995; Ladunga and Smith, 1997). Here, we explore this idea with particular reference to the phylogenetic position of the major clade of flightless birds, the ratites, and the possibility that homology exists at higher order levels in DNA sequences captured in the particular DNA strings of avian species. Our analysis also suggests how large-scale bioinformatics analysis can inform phylogenetic analysis of major clades and provide new insights into genome evolution in birds and their relatives.

Previous molecular studies of higher level relationships in birds and in vertebrates generally have gleaned information from character states of homologous sites observable in primary alignable sequences of macromolecules (e.g., Cooper and Penny, 1997; Groth and Barrowclough, 1999; van Tuinen et al., 2000). However, the vast majority of DNA sequences in the databases come from studies on diverse taxa and diverse nonhomologous genes that cannot be aligned with one another. This largest source of DNA sequence data is likely to contain information of use to phylogeneticists. In addition to the information found in character states of aligned DNA sites, global sequence features and characteristics of higher order DNA sequence structure are known to

provide some phylogenetic information, particularly at very deep phylogenetic levels such as among microbial lineages (Nussinov, 1984; Beutler et al., 1989; Pietrokovski et al., 1990; Burge et al., 1992; Karlin and Ladunga, 1994; Konopka, 1994; Karlin et al., 1997; Abella et al., 1999). However, the utility of such global sequence features for vertebrate systematics is unclear. Such higher order information presumably arises from species-specific differences in genome dynamics such as genome-wide patterns of mutation, DNA repair, and selection at the molecular level.

Genomic signatures (Karlin and Burge, 1995; Deschavanne et al., 1999) are tabulations of the frequencies of short nucleotide strings and offer one way of directly comparing and deriving phylogenetic information from nonhomologous DNA sequences. These frequencies can be displayed in the form of images (Deschavanne et al., 1999), which are superior to simple lists because they provide the extra power of visual exploration of data, revealing nested patterns and interspecific string-usage differences. For many organisms, estimation of genome-wide string frequencies and signatures can be efficiently achieved with surprisingly short DNA sequences, on the order of 10^4–10^5 nucleotides (nt) long, particularly for shorter strings (Deschavanne et al., 1999, 2000; Sandberg et al., 2001). Thus, adequate DNA sequence data for addressing phylogenetic questions via genomic signatures likely exists for many clades. In this study, we sought to determine what, if any, phylogenetic information exists in the distribution of string frequencies of avian DNA sequences and over what time scales this vocabulary is phylogenetically informative.

For most of the last century, biologists have interpreted both molecular (Sibley and Ahlquist, 1972, 1990; Prager et al., 1976; Stapel et al., 1984) and morphological (Cracraft, 1988) characters as evidence that the living flightless birds (ratites and tinamous) comprise the Paleognathae, one of two primary branches in the genealogical tree for birds. The Neognathae, consisting of all other living birds, have traditionally comprised the other major branch (Cracraft, 1988). Phylogenetic analyses of slowly evolving nuclear DNA sequences support this view (Groth and Barrowclough, 1999; van Tuinen et al., 2000). In addition, recent analyses of mitochondrial DNA (mtDNA) have supported a basal position for ratites (Paton et al., 2002). However, early studies of quickly changing mitochondrial sequences consistently supported an arrangement in which ratites and basal neognaths, such as chickens and ducks, are derived within birds (Härlid and Arnason, 1999; Mindell et al., 1999; Johnson, 2001). Although many ornithologists consider the mitochondrial result a confirmed artifact of high evolutionary rates and misrooting of the tree, others do not (S.V.E., pers. obs.). Thus there is controversy over just how controversial avian relationships are. At any rate, informed reconstruction of the morphological transitions leading to flightlessness in ratites and other avian clades and of phenotypic diversity in birds generally depends critically on achieving a tree for birds supported by multiple data sets.

MATERIALS AND METHODS

Database Sequences

To maximize the signal of genomic DNA in our signatures, the highest priority for inclusion in the study was whether a sequence was genomic DNA; we tried to minimize the use of cDNA (mRNA) or mtDNA sequences, which would possess vocabularies dominated by coding regions and organelle string usage, respectively, and therefore presumably quite different from that for genomic DNA. In addition, we tried to avoid recent collections of homologous avian sequences (e.g., Groth and Barrowclough, 1999; van Tuinen et al., 2000; Haddrath and Baker, 2001) because such sequences would produce artificially close genomic signatures. In addition, we did not use any microsatellite sequences, whose repeats would possess aberrant signatures relative to the majority of coding or noncoding genomic DNA; in the one case in which a microsatellite locus was used (*Hirundo rustica*; see Appendix), we only used flanking sequence. Still, for some species we used a small number of mRNA or mtDNA sequences or sequences previously used in phylogenetic studies to increase sequence sample size from which string frequencies could be estimated; approximately 22 sequences fall into one of these classes. Thus, although our signatures represent averages of genomic DNA and mtDNA, they

are dominated by the signatures in coding and noncoding regions of genomic DNA. Our sampling of such sequences was taxonomically unbiased so as not to skew phylogenetic results. A total of 125 GenBank entries (release 119.0) from 20 bird species and the American alligator were included in the final analyses (see Appendix for accession numbers and gene descriptions): neognaths: waterfowl and gamebirds (Galloanseriformes): Mallard Duck (*Anas platyrhynchos*), 41,678 base pairs (bp); Muscovy Duck (*Cairina moschata*), 12,277 bp; Red-breasted Merganser (*Mergus serrator*), 2,703 bp; domestic chicken (*Gallus gallus*), 594,577 bp; Japanese Quail (*Coturnix coturnix*), 140,780 bp; American Turkey (*Meleagris gallopavo*), 9,429 bp; perching birds (Passeriformes): House Finch (*Carpodacus mexicanus*), 32,585 bp; Red-winged Blackbird (*Agelaius phoeniceus*), 42,265 bp; Barn Swallow (*Hirundo rustica*), 2,238 bp; Tree Sparrow (*Passer montanus*), 3,038 bp; European Starling (*Sturnus vulgaris*), 3,303 bp; Common Canary (*Serinus canaria*), 3,397 bp; other neognaths: Whooping Crane (*Grus americana*), 4,275 bp; Sandhill Crane (*Grus canadensis*), 6,393 bp; Rock Dove (*Columba livia*), 43,914 bp; Hoatzin (*Opisthocomus hoatzin*), 1,515 bp; paleognaths: Ostrich (*Struthio camelus*), 17,498 bp; Chilean Tinamou (*Nothoprocta ornata*), 4,460 bp; Rhea (*Rhea americana*), 4,434 bp; Emu (*Dromaius novaehollandiae*), 9,429 bp; outgroup: American Alligator (*Alligator mississippiensis*), 6,890 bp. A total of 987,078 nucleotides were analyzed, with a median sequence length per species of 5.2 kilobases (kb).

Construction and Interpretation of Genomic Signatures

Cosmid-scale (~30 kb) DNA sequences for birds other than chicken are increasing in number (Edwards et al., 2000; Hess et al., 2000), but because the available sequence length for several species was limited (<4 kb), the longest string length whose frequencies could reasonably be resolved for all species was 5 nt, which required frequency estimation of 4^5 (1,024) distinct strings. Frequencies of strings 2–5 nt long were counted for each species from 5′ to 3′ for both strands of DNA, moving one base at a time (for details, see Deschavanne et al., 1999). Counting string frequencies on both strands removes any possible strand biases in string frequencies that could affect analysis. Use of a single base sliding window maximizes the power of the data and provides string frequencies very similar to an abutting-window sampling scheme. Thus, the signatures represent averages of string frequencies occurring throughout exons, introns, and repetitive and noncoding sequences.

The entire set of string frequencies for a given species can be displayed under the form of a single image, where the color value of each pixel corresponds to the frequency of a specific string in the sequence (Deschavanne et al., 1999), with darker colors indicating higher frequencies (Fig. 1). A

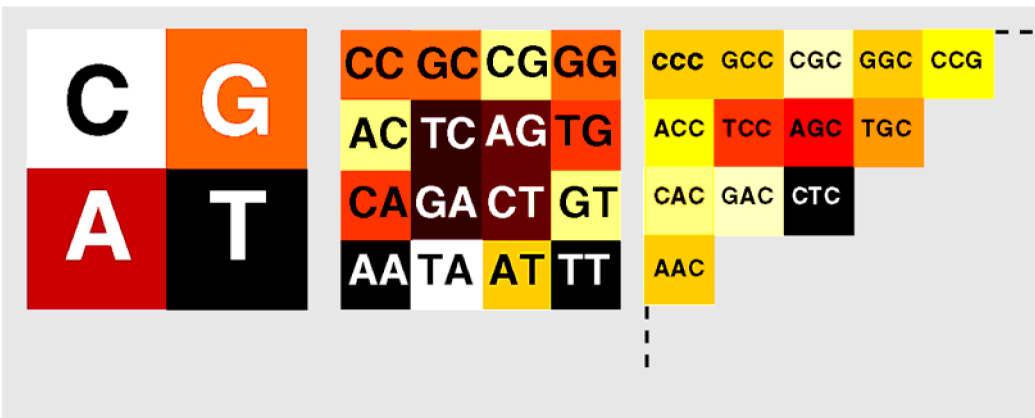

FIGURE 1. Layout of a genomic signature image. The darker the color of the pixels, the greater the frequency of the corresponding string in the DNA sequence. See text and Deschevanne et al. (1999) for details.

four-pixel image (Fig. 1, left) gives the base composition of the sequence. Each of these primary quadrants are in turn divided into four quadrants showing dinucleotide frequency (Fig. 1, center) and then trinucleotide frequency (Fig. 1, right). This recursive procedure can be iterated up to the length of strings studied. For example, the result for 5-nt strings is a 1,024-pixel image. Such a layout, referred to as chaos-game representation (Jeffrey, 1990; Deschavanne et al., 1999), has fractal properties that facilitate recognition of global vocabulary characteristics, such as prominent diagonals indicating purine and pyrimidine runs or CG scoops resulting from selection against the CG dinucleotide (Beutler et al., 1989; Deschavanne et al., 1999).

Phylogenetic Trees and Statistical Testing

A Euclidean distance, the square root of the sum of the square of the differences in frequency of strings between species, was used to generate distance matrices for phylogenetic analysis. The neighbor-joining (NJ; Saitou and Nei, 1987), Fitch–Margoliash (FM; Fitch and Margoliash, 1967), and minimum-evolution (ME; Rzhetsky and Nei, 1992) methods were employed to analyze this matrix using the PHYLIP (version 3.6) (Felsenstein, 1994) and PAUP* (Swofford, 1999) packages. In all analyses, an alligator signature was used as an outgroup. Fortuitous isolation of sequences from different isochores in the various species could introduce undesirable phylogenetic effects of global base compositional differences between species (Steel et al., 1993; Bernardi et al., 1997). To guard against these effects, the expected frequency of each string, that is, the product of the frequencies of the nucleotide composing a given string, was subtracted from the observed frequency for each species prior to distance matrix estimation.

Standard statistical comparison of phylogenetic trees with these data is problematic, primarily because there is no known estimate of the variance–covariance matrix of the distances that would permit such tests (Rzhetsky and Nei, 1992). However, we were able to make rough comparisons among four competing phylogenetic hypotheses in three ways. First, we conducted standard bootstrap analysis of the trees by resampling strings at random with replacement to create pseudosignatures for calculation of distance matrices. This process was repeated 2,000 times to make 2,000 bootstrap replicates. FM and NJ trees were subsequently constructed and summarized as a consensus tree (Felsenstein, 1985). Although strictly speaking the assumptions of the bootstrapping method are violated because strings are not independent variables, this approach was useful nonetheless. We also calculated the probability of obtaining bychance trees with two prespecified monophyletic groups from the entire universe of possible trees. This probability is $(2a - 3)!!(2b - 1)!!/(2n - 3)!!$ (M. Steel, pers. comm.), where $K!! = 3 * 5 * 7 * \ldots * K$ (for odd numbers), a and b are the number of species in the two specified clades (respectively 4 and 16), and n is the total number of species $(a + b)$. Finally, we were able to make additional comparisons among competing trees by calculating the sum of squared deviations (SSD) of the intertaxon matrix and tree distances using the Fitch–Margoliash method; the SSD is expected to be lower for trees that better fit the distance matrix. In these comparisons, trees were used that only approximated competing hypotheses because of incomplete overlap of taxa between studies.

We also conducted statistical tests of variation in string frequencies among species. Within the framework of our methods, genomic signatures are objects characterized by string frequencies, where each string adds a new level of dimensionality to interspecific comparisons. Principal components analysis (PCA) offers a handy approach for summarizing genomic signature complexity and diversity. Each species can be represented as a point in a low-dimensional space where the axes express factors of variability between species, in decreasing order of magnitude, and distances between points represent differences between species. Discriminant analysis constructs a composite function that maximizes the separability of two (or more) groups identified a priori. This function is evaluated on its ability to correctly classify species left out during the construction of the discriminant function (one species at a time). When applied to multiple-species problems such as ours, such tests are not immune to the problem of nonindependence among species inherent in phylogenetic relationships. However, such effects are expected to be small in our study because the stemminess of our trees (the ratio of lengths of

internal to external branches) was in general very low, a situation in which such statistical compromise is minimized (Purvis, 1996).

RESULTS

Distinguishing Features of Genomic Signatures in Birds

Among the 20 avian and alligator sequences, the percentage of dinucleotides G and C varies from 41% (swallow) to 58% (Whooping Crane; Fig. 2). Consistent with earlier findings (Kadi et al., 1993; Cacciò et al., 1994; Bernardi et al., 1997; Hughes et al., 1999), neither paleognaths (%GC range, 43–50%) nor alligator (46%) are base compositional outliers relative to neognaths. Of the four string lengths investigated, 5-nt string frequencies were analyzed in the greatest detail because they capture the highest genomic signature complexity. All the 5-nt avian signatures possess the major signature features of homeothermic vertebrates (Deschavanne et al., 1999; Hess et al., 2000), such as prominent diagonals, indicating a high frequency of strings consisting only of purines or pyrimidines, and light quadrants in specific areas of the signature corresponding to deficiencies of strings containing CG and TA dinucleotides (Fig. 3). Previous analysis of the first avian genomic signature, based on a ~32-kb sequence from the House Finch (Hess et al., 2000), showed that it possessed many of the features of mammalian signatures but also a deficiency of TA dinucleotides not found in mammals. This deficiency, also reported for other avian genomes (Primmer et al., 1997), is clearly present as a light region in the lower right quadrant of the lower left quadrant of all 21 signatures but less so in the alligator and ratite signatures (Fig. 3).

Phylogenetics of Genomic Signature Diversity

Phylogenetic analysis of the Euclidean distance matrix made from 5-nt signatures using all three methods placed all paleognaths (Ostrich, Emu, Rhea, and Chilean Tinamou) in a monophyletic group (bootstrap support [bss]-68%) at the base of the avian tree (Fig. 3). In addition, the analysis strongly supports monophyly of neognath sequences (bss = 100%), as implied by most morphological and molecular data (Stapel et al., 1984; Cracraft, 1988; Groth and Barrowclough, 1999; van Tuinen et al., 2000) but not by DNA hybridization (Sibley and Ahlquist, 1990) or mtDNA (Härlid and Arnason, 1999; Mindell et al., 1999; Johnson, 2001) data. Similarly, analysis of 2-nt strings placed all paleognaths at the base along with the canary signature (Fig. 4c). Analysis of 3- and 4-nt string frequencies also placed all paleognaths at the base of the tree, albeit not as a monophyletic group, and strongly supported monophyly of neognath sequences (bss > 97%; Figs. 4a, 4b). A priori, a basal paleognath/neognath split such as that detected in our analyses of

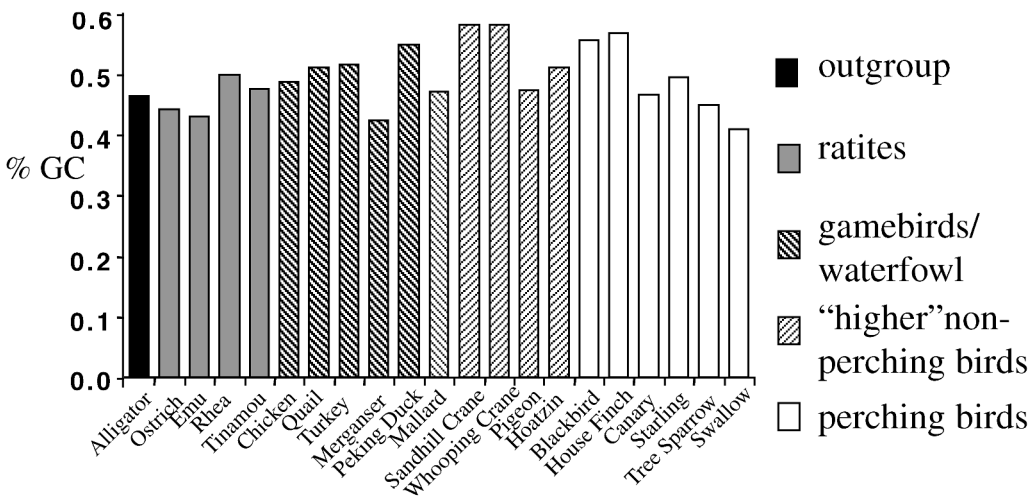

FIGURE 2. GC content of the DNA sequences from 20 avian species and an American alligator. Taxonomic categories are based on traditional ornithological groupings.

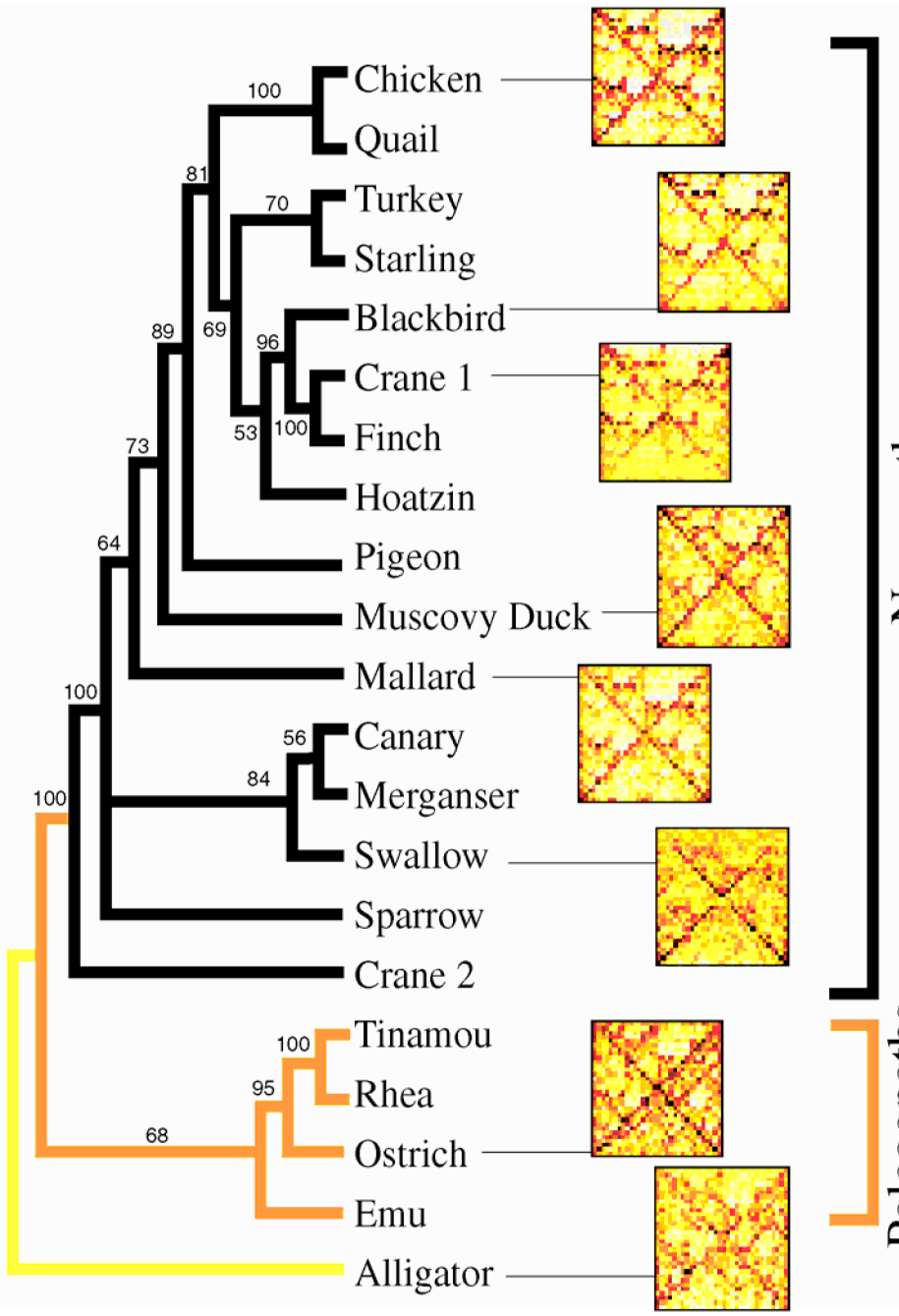

FIGURE 3. The 50% majority-rule consensus bootstrap FM tree of genomic signatures from 20 birds and an American alligator (designated outgroup). Analysis using the NJ method achieved identical results. Representative signatures depict the base compositionally corrected frequencies of 1,024 5-nt DNA strings. Black borders around each signature are for clarity and do not indicate information on word frequencies. Orange branches lead to paleognath signatures, black branches to neognaths. Branch lengths are not to scale.

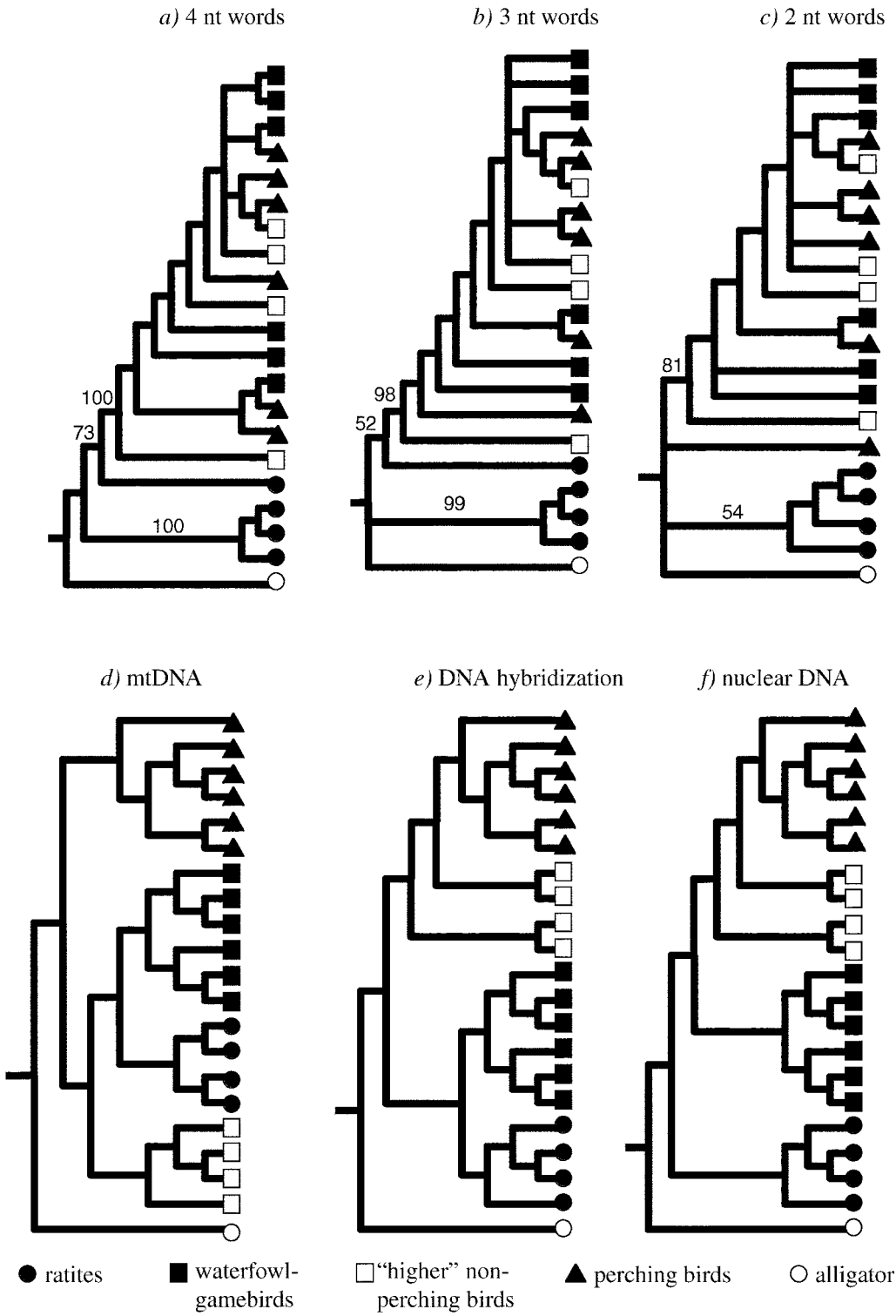

FIGURE 4. Consensus bootstrap FM trees (50% majority rule) of genomic signatures derived from frequencies of 2-, 3-, and 4-nt strings (a–c) and competing phylogenetic hypotheses to which signature distance matrices were fitted using the Fitch and Kitsch methods (d–f). In (a–c), bootstrap values are indicated above selected branches leading to ratites and to neognaths. Branch lengths are not proportional to divergence.

5-nt signatures is likely to occur by chance in only 0.035% of trees.

In addition to bootstrap analysis, inspection of SSDs shows that signature trees in which paleognaths are basal are more strongly supported than are competing trees. Whereas the best FM tree had the best fit with the lowest SSD (1.70), the trees implied by mtDNA (Fig. 4d) or DNA hybridization (Fig. 4e) data had much poorer fits with higher SSDs (8.25 and 8.28, respectively). The tree topology supported by recent analyses of nuclear DNA sequences (van Tuinen et al., 2000) has the lowest SSD (6.56) of any of the alternative hypotheses (Fig. 4f). The SSDs for all trees in which branch lengths were not constrained to a molecular clock were significantly lower than those in which branch lengths were constrained (user-defined Kitsch trees, *F*-test: $P < 0.05$; signature tree, 42.82; mtDNA tree, 43.42; DNA hybridization tree, 50.57; nuclear DNA sequence tree, 49.77), indicating that rates of signature evolution are variable in birds.

Statistical Analysis of String Frequencies

As judged by the percentage of the total variance explained by the PCA axes, the PCA of compositionally corrected signatures did a reasonable job of explaining the variation among species in string frequencies. PCA1 explained 26.6% of the variance, PCA2 explained 20.7%, and PCA3 explained 10.8%, with a total of 58% of the variance explained by these three axes (out of a total of 512 dimensions); 95% of the variance was explained in the first 12 PCA axes. Axes 1 and 2 separated all four paleognaths from the neognaths and placed paleognaths close to the alligator signature (Fig. 5). Because the alligator signature is unambiguously the outgroup, these analyses suggest that ratite genomes are linguistically primitive within birds. The first component roughly corresponds to differences in purine–pyrimidine runs, whereas the second component captures differences in string frequency of specific common strings. Discriminant analysis of the signatures correctly identified paleognaths 100% of the time (jackknife validation). The frequency of 448 of the 1,024 possible 5-nt strings was significantly different between paleognaths and neognaths at the 5% level, and 172 were different at the 1% level (nonparametric

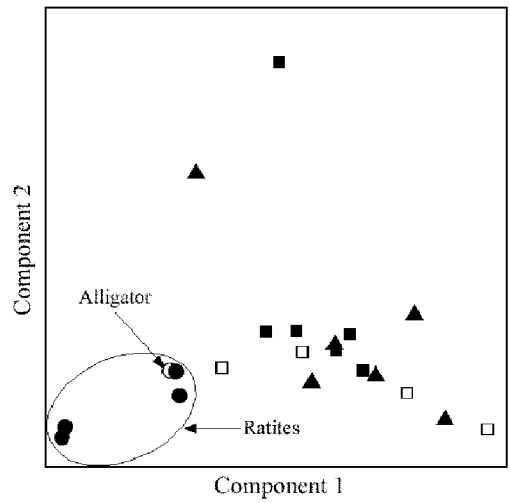

FIGURE 5. Principal component analysis of genomic signatures of 16 neognaths (■ = waterfowl/gamebirds; □ = "higher" nonperching birds; ▲ = perching birds), 4 paleognaths (●), and an alligator (○).

Kruskall–Wallis test). The frequencies of 46 strings (Table 1) show the strongest possible differences (nonoverlapping distribution) between the two groups (38 strings were found more frequently in the ratites and 8 were less frequent). These strings occur on average 2.28 times as frequently in paleognath than in neognath DNA but only 1.4 times as frequently as in the alligator, again suggesting a primitive DNA vocabulary for paleognaths. The data suggest that several 6-nt and even longer strings (7-nt strings such as ATTAGCC and CTTAACA and rare 8-nt strings) obtained by concatenation of some of the most discriminating 46 5-nt strings have frequencies that also differ drastically between the two groups.

Because we included a small number of mtDNA and mRNA sequences and sequences that were used in previous phylogenetic studies, we analyzed trees made from 5-nt string frequency tables after removal of these sequences. For this analysis, we had to add several new sequences to the outgroup, which consisted solely of such sequences. The FM tree of this reduced matrix had all ratites at the base of the tree, albeit with the Emu signature clustering with the neognath sequences, as in the 3- and 4-nt trees (Figs. 4a, 4b). The PCA plot again showed clear discrimination between neognath and paleognath sequences, with the alligator and paleognath sequences falling in the same

TABLE 1. Estimated occurrences per 10,000 bases of 46 5-nt strings that strongly distinguish paleognath and neognath DNA vocabularies.[a]

String	Alligator (A)	Paleognath (P)[b]	Neognath (N)[b]	P/N	P/A
AACCC	14.89	15.93	8.5	1.87	1.07
TAGCC	6.7	9.45	3.29	2.87	1.41
TCTCC	8.93	8.75	21.68	0.4	0.98
AGTGC	8.19	7.01	13.13	0.53	0.86
TAACC	5.95	11	4.22	2.61	1.85
TTAGC	11.16	12.89	3.73	3.45	1.15
AATCG	6.7	7.4	1.94	3.8	1.1
TCCAG	9.68	7.18	19.67	0.36	0.74
TTCTG	14.14	6.89	18.63	0.37	0.49
TCTAG	8.19	15.13	4.99	3.03	1.85
CTAAC	5.21	11.11	3.58	3.1	2.13
TAAAC	12.65	14.21	6.61	2.15	1.12
ATAAC	12.65	12.78	5.64	2.27	1.01
TTAAC	7.44	12.48	5.35	2.33	1.68
TAATC	8.93	12.22	5.68	2.15	1.37
TTAAG	15.63	19.34	8.3	2.33	1.24
ATTAG	5.21	11.22	4.17	2.69	2.15
TAATG	8.19	13.07	7.25	1.8	1.6
TATTG	9.68	13.12	6.44	2.04	1.36
TAACA	7.44	15.47	6.15	2.52	2.08
TTAGT	9.68	16.74	4.37	3.83	1.73
AACTA	9.68	17.12	5.02	3.41	1.77
AATTA	15.63	22.03	8.54	2.58	1.41

[a] Only 23 strings are shown because complementary strings have equal frequencies when both DNA strands are considered.
[b] Estimates are averaged across species within paleognaths and neognaths.

region of string frequency space. We therefore conclude that there is genuine signal in the genomic DNA data that indicates an ancestral position of ratite genomic signatures, and this signal was not compromised by the small amount of mtDNA, mRNA, or the phylogenetically sampled sequenes in the original data set.

DISCUSSION

Evaluation of the Phylogenetic Information Content of Genomic Signatures

Our analysis suggests that the distribution of nucleotide strings in genomic DNA sequences in birds may contain some phylogenetic information, particularly at deep levels within birds. Despite its base compositional similarity to neognath genomic DNA, paleognath genomic DNA exhibits a string usage strikingly different from that of neognaths and resembling that of alligators. This difference in string frequency is strong enough to recover one of two competing hypotheses for basal branches within birds: Our analysis favors a sister-group relationship of neognaths and paleognaths, as supported by most nuclear DNA sequence data and morphology. The monophyly of neognaths and paleognaths has been supported by analyses of morphology (Cracraft, 1988) and nuclear DNA sequence data (Prager et al., 1976; Stapel et al., 1984; van Tuinen et al., 2000) but not, for example, by DNA hybridization studies (Sibley and Ahlquist, 1990), where Galloanseriformes is the sister to paleognaths. Statistical analysis of string frequencies alone strongly contradicts the hypothesis that paleognaths are a derived clade within birds (Härlid and Arnason, 1999; Mindell et al., 1999) and supports the idea that the genomic vocabulary of ratites is primitive within birds. The morphological (De Beer, 1956; Cracraft, 1988), biogeographical (Cracraft, 1974; Houde, 1986), physiological (Dawson et al., 1996), and karyotypic (Ogawa et al., 1998) distinctness of the paleognaths has been widely discussed. In addition to these traits, our analysis predicts the existence of genomewide mutational spectra and selective constraints distinguishing the genomic vocabularies of paleognaths from those of other birds.

These results are intriguing because this information has been extracted from DNA sequences that, by normal criteria, would

clearly be considered nonhomologous and, in principle, devoid of phylogenetic information. This study is the first to demonstrate that higher order information can be gleaned from heterogeneous collections of DNA sequences and used to recover deep branches in vertebrate trees; several researchers have made similar claims for microbial systems (Pietrokovski et al., 1990; Karlin et al., 1997; Abella et al., 1999). Comparison of nonhomologous characters can frequently mislead phylogenetic analysis, but we have shown that higher order properties of "nonhomologous" DNA sequences may nonetheless be a source of phylogenetic information. However, it is less useful to describe the characters we have collected as "nonhomologous"; rather, like morphology, they likely represent traits that are affected by underlying genetic synapomorphies (such as changes in DNA repair enzymes) and are scored at hierarchical levels above the primary DNA sequence. To the extent to which similar string frequencies are due to shared developmental mechanisms, such as DNA repair, DNA–protein interactions, or mutation biases, genomic signatures can be considered to reflect homology (Mindell and Meyer, 2001). Perhaps more importantly, our results suggests that fundamentally new properties of the genomes of avian lineages can be discovered by comparison of string frequencies, new properties that need to be explained on a mechanistic basis and that could prove useful as heuristic tools in other vertebrate clades (such as marsupials and eutherians) or genomic regions (such as nonrecombining and recombining DNA). This is the largest to date on avian molecular evolution. Although many researchers have collected very large DNA sequence databases to document the taxonomic distinctness of the signatures of various microbial species (Nussinov, 1984; Karlin and Ladunga, 1994; Abella et al., 1999), our data set is the largest (~1 Mb) to be analyzed via standard tree-building methods.

The phylogenetic utility of genomic signatures in this study rests, however, on only the basal split within birds; we clearly cannot claim that genomic signatures offer phylogenetic resolution for other parts of the avian tree. Relationships within neognaths are still difficult to resolve even by standard phylogenetic analysis (Groth and Barrowclough, 1999; van Tuinen et al., 2000). For example, in the analysis presented by van Tuinen et al. (2000), the placement of the Galloanseriformes (gamebirds and waterfowl) as the sister to other neognaths showed congruence with other studies, but the remainder of their tree was poorly resolved, as indicated by low bootstrap values. Still, the tree based on genomic signatures (Fig. 3) clearly has produced results that have no previously published support (e.g., lack of monophyly of Galliformes, cranes, and Passeriformes); worse, several clearly incorrect nodes show strong bootstrap support (e.g., House Finch/Whooping Crane, Turkey/Starling, and Mallard/Pigeon)! These results raise serious issues with the use of genomic signatures as general tools for phylogenetics.

If neognath and paleognath monophyly were not correct, then we could not claim any phylogenetic utility for genomic signatures. However, we would still have discovered a fundamental genomic schism within birds, that separating paleognaths and neognaths, that is in need of explanation. The similarity of ratite signatures to one another could be a convergent result of selection on the genome imposed in independent lineages by physiological factors associated with flightlessness. This hypothesis could be tested by examining signatures of other flightless or low-metabolic-rate (but nonratite) birds, such as flightless parrots or rails. We suspect, however, that the genomic signatures have correctly recovered the basal split within birds, based on congruence with many other analyses. Even the results of the recent analysis of complete mtDNA sequences begin to converge on those of nuclear DNA studies when analyzed with attention to the details of substitution dynamics (Mindell et al., 1999), although some mitochondrial studies, particularly those on single mitochondrial genes, still recover trees in which paleognaths are derived (Johnson, 2001). Many physiological traits associated with flight and with metabolic rate probably do impose selection on avian genomes that could influence signature structure (Hughes et al., 1999; Hughes, 2000; Waltari and Edwards, 2002). Thus, it will be interesting to study the diversity of signatures across vertebrates to determine whether avian signatures cluster with mammals, possibly because of convergent homeothermy, or with reptiles, as phylogeny would sugggest.

We suspect that some of the bizarre results in our tree are the result of our data set. For example, small differences in genomic signatures can hardly be expected to be accurately detected in short DNA sequences, for which the estimation of string frequencies may be less reliable. Among recently diverged species, where the differences in signatures are expected to be small, the impact of taxon-specific physiological, mutational, or isochore biases likely mislead phylogenetic analysis (Martin, 1995; Hughes et al., 1999; Fryxell and Zuckerkandl, 2000). Extracting information from long sequences and/or long strings would permit enhanced discriminating power among signatures of closely related species (B.F., A.G., and P.J.D., unpubl.)

Still, we would be overstating the utility of genomic signatures for strictly phylogenetic analysis if we blamed our trees on an inadequate data set. Rather, we suspect that, as one might guess for comparisons of nonhomologous DNA sequences, it is simply difficult to extract detailed phylogenetic information from data such as these. We therefore promote genomic signatures not as a general tool for phylogenetics but rather as an exploratory tool for examining genome evolution, particularly for closely related groups. Our analysis suggests that differences in genomic signatures will track phylogeny primarily at deep nodes within vertebrates, as has been shown for major branches of life (Deschavanne et al., 1999). Analysis of genomic signatures of various clades could also inform traditional phylogenetic analysis. If the differences in higher order DNA sequence structure between ratites and neognaths were due to underlying differences in the mutational spectrum, directional mutation pressure, or physiological constraints on DNA sequence evolution, these differences could be accommodated into parameter values for models of DNA sequence evolution that better approximate the dynamics within each clade. The detection of nonstationarity in the substitution process, in evolutionary rates, and in base composition in different lineages is becoming more common (Hasegawa, 1990; Rzhetsky and Nei, 1995; Sanderson, 1997; Sullivan et al., 1999; Yang and Yoder, 1999; Hershkovitz and Zimmer, 2000).

The principle behind phylogenetic analysis of genomic signatures could be further compromised by the problem of nonindependence of different (adjacent) nucleotides in DNA sequences and the possibility of convergence due to selection at the level of DNA strings. For example, the data presented here and by Hess et al. (2000) suggest that the genomic signatures of birds match those of other homeotherms (mammals) quite closely. Martin (1995) and others have provided examples in which increased metabolic rates changed the quantitative dynamics of DNA damage and mismatch mutation, with increases in GC nucleotides associated with higher metabolic rates. Based on this logic, Hess et al. (2000) suggested that homeothermy and basal metabolic rate might contribute to similarities of genomic signatures independent of phylogenetic relationships. Consistent with this effect, neognaths, particularly passerines, are known to have on average higher basal metabolic rates than paleognaths (Garland and Ives, 2000). Because this scenario addresses only point mutations and because we found no strong differences in base composition between paleognaths and neognaths, it does not directly address the origin of differences in DNA strings between these groups. Nonetheless, it will be important in the future to conduct comparative analyses to determine the possible convergent impacts of physiology on genomic signature diversity.

Dynamics of Genomic Signatures

How in fact do differences in genomic signatures arise? There is a diverse and scattered literature on the molecular basis of mutation, with some attention to mutational processes governing multiple nucleotide sites. Several studies have revealed an effect of neighboring bases or local base composition on the pattern of point substitution at a focal nucleotide site (Wolfe and Sharp, 1993; Morton and Clegg, 1995), with the effect sometimes dependent on the purine–pyrimidine status of adjacent bases or on the particular nucleotide. Averof et al. (2000) recently documented a higher than expected rate of double (adjacent) substitutions in mammalian nuclear DNA. Both the process of DNA repair and the interaction of DNA and proteins have been suggested to introduce higher order mutation biases in DNA sequences. For example, the rate of repair of *N*-methylpurines is highly dependent on

the position of the damaged base (Ye et al., 1998). DNA-binding proteins also influence the spectrum and rate of mutation in the bound DNA region, with the result that DNA regions in higher order chromatin structures or targets of nucleosomal binding sites can experience retarded or elevated rates of DNA damage and mutation (Boulikas, 1992; Pfeifer et al., 1992; Holmquist, 1994). Certain DNA repair processes of experimentally damaged DNA require excisions of DNA spanning several (up to 40) nucleotides, making possible the evolution of local patches of DNA as a unit (Cleaver et al., 1991).

Some dinucleotides are particularly well studied with regard to higher order evolutionary processes. The CG dinucleotide is known to occur much less frequently than expected in vertebrate genomes, given overall base compositions (Beutler et al., 1989). This dinucleotide is of particular interest because of its well-known tendency to mutate, its association with DNA methylation and gene regulation, and its abundance in CpG islands, which occur upstream of many mammalian and avian housekeeping genes (McQueen et al., 1998). Whereas the CG dinucleotide is estimated to occur on average 379 and 344 times per 10 kb in ratite and alligator DNA, respectively, it is estimated to occur only 256 times per 10 kb in nonratite birds. This difference in the frequency of CG dinucleotides may indicate differences in CpG island frequency or length or may simply indicate that our selection of sequences contains more housekeeping genes in ratites than in nonratites. By contrast, the TA dinucleotide, which appears to occur less frequently in avian than in mammalian genomes (Primmer et al., 1997; Hess et al., 2000), occurs at an intermediate frequency in alligator DNA (695 times per 10 kb) compared with ratite (603 times) and neognath (761 times) DNA, indicating divergent shifts in frequency during avian genome evolution.

A Genomic Signature Clock?

Using a series of crude *F*-tests, we rejected a clock for genomic signature evolution. Our data suggest possible increases in rate of genomic signature evolution in perching birds (Passeriformes). First, visual inspection (data not shown) suggests longer branches along passerine lineages in unconstrained (Fitch) trees (Figs. 4d–f). Second, in constrained trees in which a molecular clock is in effect, the signature diversity fits the mtDNA tree, in which passerines are basal, better than it fits the nuclear DNA sequence or DNA hybridization trees, in which passerines are derived. For the mtDNA tree, the longer branches in the passerine clade are better accommodated at the base of the tree when branches from the root to the tips are constrained to be equal, suggesting a higher rate of signature evolution in this clade. An increase in the rate of signature evolution in passerines would be consistent with conclusions of other studies that have found increases in the rate of point mutation in this clade (Sibley and Ahlquist, 1988, 1990; Cooper and Penny, 1997). Although our analysis indicated the possible utility of genomic signatures, full realization of the potential of genomic signatures for phylogenetic reconstruction and an understanding of their evolutionary dynamics will require considerably more advanced statistical and theoretical approaches (e.g., Sandberg et al., 2001).

ACKNOWLEDGMENTS

We thank M. Steel, A. Rzhetsky, D. Mindell, and J. Felsenstein for helpful discussion. Peter Beerli implemented the bootstrapping analysis, and R. Brumfield provided assistance with minimum-evolution trees. S. Naeem, T. Price, A. J. Baker, C. Laird, J. Lyons-Weiler, C. Hess, C. Simon, H. Hoekstra, R. Brumfield, and an anonymous reviewer provided helpful comments on the manuscript. This work was supported by grants from the NSF (to S.V.E.), and we thank J. Graves and the Comparative Genomics Group, Research School of Biological Sciences, Australian National University, for providing a productive working environment during the revision of this paper.

REFERENCES

ABELLA, C. A., V. N. IVANOV, AND I. S. KIM. 1999. Taxon-specific content of oligonucleotide triplets in 16S rRNAs of anoxygenic phototropic and nitrifying bacteria. J. Theor. Biol. 196:289–296.

AVEROF, M., A. ROKAS, K. H. WOLFE, AND P. M. SHARP. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science 287:1283–1286.

BERNARDI, G., S. HUGHES, AND D. MOUCHIROUD. 1997. The major compositional transitions in the vertebrate genome. J. Mol. Evol. 44:S44–S51.

BEUTLER, E., T. GELBART, J. HAN, J. A. KOZIOL, AND B. BEUTLER. 1989. Evolution of the genome and the genetic code: Selection at the dinucleotide level by methylation and polyribonucleotide cleavage. Proc. Natl. Acad. Sci. USA 86:1992–1996.

BOULIKAS, T. 1992. Evolutionary consequences of nonrandom damage and repair of chromatin domains. J. Mol. Evol. 35:156–180.

BULLOCK, T. L., W. D. CLARKSON, H. M. KENT, AND M. STEWART. 1996. The 1.6 angstroms resolution crystal structure of nuclear transport factor 2 (NTF2). J. Mol. Biol. 260:422–431.

BURGE, C., A. M. CAMPBELL, AND S. KARLIN. 1992. Over- and under-representation of short oligonucleotide in DNA sequences. Proc. Natl. Acad. Sci. USA 89:1358–1362.

CACCIÒ, S., P. PERANI, S. SACCONE, F. KADI, AND G. BERNARDI. 1994. Single-copy sequence homology among the GC-richest isochores of the genomes from warm-blooded vertebrates. J. Mol. Evol. 39:331–339.

CLEAVER, J. E., J. JEN, W. C. CHARLES, AND D. L. MITCHELL. 1991. Cyclobutane dimers and (6-4) photoproducts in human cells are mended with the same patch sizes. Photochem. Photobiol. 54:393–402.

COOPER, A., AND D. PENNY. 1997. Mass survival of birds across the Cretaceous–Tertiary boundary: Molecular evidence. Science 275:1109–1113.

CRACRAFT, J. 1974. Phylogeny and evolution of the ratite birds. Ibis 116:494–521.

CRACRAFT, J. 1988. The major clades of birds. Pages 339–361 *in* The phylogeny and classification of the tetrapods, Volume 1. Amphibians, reptiles, birds, Volume 35A (M. J. Benton, ed.). Clarendon Press, Oxford, U.K.

DAWSON, A., D. C. DEEMING, A. C. K. DICK, AND P. J. SHARP. 1996. Plasma thyroxine concentrations in farmed ostriches in relation to age, body weight and growth hormone. Gen. Comp. Endocrinol. 103:308–315.

DE BEER, G. 1956. The evolution of ratites. Bull. B. Mus. Nat. Hist. 4:59–70.

DESCHAVANNE, P., A. GIRON, J. VILAIN, C. DUFRAIGNE, AND B. FERTIL. 2000. Genomic signature is preserved in short DNA fragments. IEEE Int. Symp. Bioinf. Biomed. Eng. 2000:161–167.

DESCHAVANNE, P. J., A. GIRON, J. VILAIN, G. FAGOT, AND B. FERTIL. 1999. Genomic signature: Characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol. 16: 1391–1399.

EDWARDS, S. V., J. GASPER, D. GARRIGAN, D. A. MARTINDALE, AND B. F. KOOP. 2000. A 39-kb sequence around a blackbird MHC class II B gene: Ghost of selection past and songbird genome architecture. Mol. Biol. Evol. 17:1384–1395.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

FELSENSTEIN, J. 1994. Phylogeny inference package (PHYLIP), version 3.6. Univ. Washington, Seattle.

FITCH, W. M., AND E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science 155:279–284.

FRYXELL, K. J., AND E. ZUCKERKANDL. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol. Biol. Evol. 17:1371–1383.

GARLAND, T., JR., AND A. R. IVES. 2000. Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. Am. Nat. 155:346–364.

GROTH, J. G., AND G. F. BARROWCLOUGH. 1999. Basal divergences in birds and the phylogenetic utility of the nuclear RAG-1 gene. Mol. Phylogenet. Evol. 12:115–123.

HADDRATH, O., AND A. J. BAKER. 2001. Complete mitochondrial DNA genome sequences of extinct birds: Ratite phylogenetics and the vicariance biogeography hypothesis. Proc. R. Soc. Lond. B 268: 939–945.

HÄRLID, A., AND U. ARNASON. 1999. Analyses of mitochondrial DNA nest ratite birds within the Neognathae: Supporting a neotenous origin of ratite morphological characters. Proc. R. Soc. Lond. B 266: 305–309.

HASEGAWA, M. 1990. Phylogeny and molecular evolution in primates. Jpn. J. Genet. 65:243–266.

HERSHKOVITZ, M. A., AND E. A. ZIMMER. 2000. Ribosomal DNA evidence and disjunctions of western American Portulacaceae. Mol. Phylogenet. Evol. 15:419–439.

HESS, C. M., J. GASPER, H. HOEKSTRA, C. HILL, AND S. V. EDWARDS. 2000. MHC class II pseudogene and genomic signature of a 32-kb cosmid in the House Finch (*Carpodacus mexicanus*). Genome Res. 10:613–623.

HOLMQUIST, G. P. 1994. Chromatin self-organization by mutation bias. J. Mol. Evol. 39:436–438.

HOUDE, P. W. 1986. Ostrich ancestors found in the Northern Hemisphere suggest new hypotheses of ratite origins. Nature 324:563–565.

HUGHES, A. L. 2000. Adaptive evolution of genes and genomes. Oxford univ. Press, Oxford, U.K.

HUGHES, S., D. ZELUS, AND D. MOUCHIROUD. 1999. Warm-blooded isochore structure in the Nile Crocodile and turtle. Mol. Biol. Evol. 16:1521–1527.

JEFFREY, H. J. 1990. Chaos game representation of gene structure. Nucleic Acids Res. 18:2163–2170.

JOHNSON, K. P. 2001. Taxon sampling and the phylogenetic position of Passeriformes: Evidence from 916 avian cytochrome *b* sequences. Syst. Biol. 50:128–136.

KADI, F., D. MOUCHIROUD, G. SABEUR, AND G. BERNARDI. 1993. The compositional patterns of the avian genomes and their evolutionary implications. J. Mol. Evol. 37:544–551.

KARLIN, S., AND C. BURGE. 1995. Dinucleotide relative abundance extremes: A genomic signature. Trends Genet. 11:283–290.

KARLIN, S., AND I. LADUNGA. 1994. Comparisons of eukaryotic genomic sequences. Proc. Natl. Acad. Sci. USA 91:12832–12836.

KARLIN, S., J. MRÁZEK, AND A. M. CAMPBELL. 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. 179:3899–3901.

KONOPKA, A. K. 1994. Sequences and codes, fundamentals of biomolecular cryptology. Pages 119–173 *in* Biocomputing: Informatics and genome projects (D. Smith, ed.). Academic Press, San Diego.

LADUNGA, I., AND R. F. SMITH. 1997. Amino acid substitutions preserve protein folding by conserving steric and hydrophobicity properties. Protein Eng. 10:187–196.

LEUNISSEN, J. A., AND W. W. DE JONG. 1986. Phylogenetic trees constructed from hydrophobicity values of protein sequences. J. Theor. Biol. 119:189–196.

MARTIN, A. P. 1995. Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. Mol. Biol. Evol. 12:1124–1131.

MATSUO, Y., A. YAMADA, K. TSUKAMOTO, H. O. TAMURA, H. IKEZAWA, H. NAKAMURA, AND K. NISHIKAWA. 1996. A distant evolutionary relationship between bacterial sphingomyelinase and mammalian DNase I. Protein Sci. 5:2459–2467.

MCQUEEN, H. A., G. SIRIACO, AND A. P. BIRD. 1998. Chicken microchromosomes are hyperacetylated, early replicating, and gene rich. Genome Res. 8:621–630.

MINDELL, D. P., AND A. MEYER. 2001. Homology evolving. Trends Ecol. Evol. 16:434–440.
MINDELL, D. P., M. D. SORENSON, D. E. DIMCHEFF, M. HASEGAWA, J. C. AST, AND T. YURI. 1999. Interordinal relationships of birds and other reptiles based on whole mitochondrial genomes. Syst. Biol. 48:138–152.
MORTON, B. R., AND M. T. CLEGG. 1995. Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. J. Mol. Evol. 41:597–603.
NAYLOR, G. J., T. M. COLLINS, AND W. M. BROWN. 1995. Hydrophobicity and phylogeny. Nature 373:565–566.
NUSSINOV, R. 1984. Doublet frequencies in evolutionary distinct groups. Nucleic Acids Res. 12:1749–1763.
OGAWA, A., K. MURATA, AND S. MIZUNO. 1998. The location of Z- and W-linked marker genes and sequence on the homomorphic sex chromosomes of the ostrich and the emu. Proc. Natl. Acad. Sci. USA 95:4415–4418.
PATON, T., O. HADDRATH, AND A. J. BAKER. 2002. Complete mitochondrial DNA genome sequences show that modern birds are not descended from transitional shorebirds. Proc. R. Soc. Lond. B 269:839–846.
PFEIFER, G. P., R. DROUIN, A. D. RIGGS, AND G. P. HOLMQUIST. 1992. Binding of transcription factors creates hot spots for UV photoproducts in vivo. Mol. Cell Biol. 12:1798–1804.
PIETROKOVSKI, S., S. HIRSHON, AND E. N. TRIFONOV. 1990. Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. J. Biol. Struct. Dynam. 7:1251–1268.
PRAGER, E. M., A. C. WILSON, D. T. OSUGA, AND R. E. FEENEY. 1976. Evolution of flightless land birds on southern continents: Transferrin comparison shows monophyletic origin of ratites. J. Mol. Evol. 8:283–294.
PRIMMER, C. R., T. RAUDSEPP, B. P. CHOWDHARY, A. P. MØLLER, AND H. ELLEGREN. 1997. Low frequency of microsatellites in the avian genome. Genome Res. 7:471–482.
PURVIS, A. 1996. Using interspecies phylogenies to test macroevolutionary hypotheses. Pages 153–168 *in* New uses for new phylogenies (P. H. Harvey, A. J. L. Brown, J. M. Smith, and S. Nee, eds.). Oxford Univ. Press, Oxford, U.K.
RZHETSKY, A., AND M. NEI. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. 9:945–967.
RZHETSKY, A., AND M. NEI. 1995. Tests of applicability of several substitution models for DNA sequence data. Mol. Biol. Evol. 12:131–151.
SAITOU, N., AND M. NEI. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.
SANDBERG, R., G. WINBERG, C.-I. BRÄNDEN, A. KASKE, I. ERNBERG, AND J. CÖSTER. 2001. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Res. 11:1404–1409.
SANDERSON, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14:1218–1231.
SCHNEIDER, T. D. 1997. Information content of individual genetic sequences. J. Theor. Biol. 189:427–441.
SCHÖNIGER, M., AND A. VON HAESELER. 1999. Toward assigning helical regions in alignments of ribosomal RNA and testing the appropriateness of evolutionary models. J. Mol. Evol. 49:691–698.
SIBLEY, C. G., AND J. E. AHLQUIST. 1972. A comparative study of the egg white proteins of non-passerine birds. Peabody Mus. Nat. Hist. Bull. 39:1–276.
SIBLEY, C. G., AND J. E. AHLQUIST. 1988. A classification of the living birds of the world based on DNA–DNA hybridization studies. Auk 105:409–423.
SIBLEY, C. G., AND J. E. AHLQUIST. 1990. The phylogeny and classification of birds: A study in molecular evolution. Yale Univ. Press, New Haven, Connecticut.
STAPEL, S. O., J. A. M. LEUNISSEN, M. VERSTEEG, J. WATTEL, AND W. W. DE JONG. 1984. Ratites as oldest offshoot of avian stem—Evidence from α crystallin sequences. Nature 311:257–259.
STEEL, M. A., P. J. LOCKHART, AND D. PENNY. 1993. Confidence in evolutionary trees from biological sequence data. Nature 364:440–442.
SULLIVAN, J., D. L. SWOFFORD, AND G. J. P. NAYLOR. 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol. Biol. Evol. 16:1347–1356.
SUYAMA, M., Y. MATSUO, AND K. NISHIKAWA. 1997. Comparison of protein structures using 3D profile alignment. J. Mol. Evol. 44:S163–S173.
SWOFFORD, D. L. 1999. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b. Sinauer, Sunderland, Massachusetts.
VAN TUINEN, M., C. G. SIBLEY, AND S. B. HEDGES. 2000. The early history of modern birds inferred from DNA sequences of nuclear and mitochondrial ribosomal genes. Mol. Biol. Evol. 17:451–457.
WALTARI, E., AND S. V. EDWARDS. 2002. The evolutionary dynamics of intron size, genome size, and physiological correlates in archosaurs. Am. Nat. (in press).
WOLFE, K. H., AND P. M. SHARP. 1993. Mammalian gene evolution: Nucleotide sequence divergence between mouse and rat. J. Mol. Evol. 37:441–456.
YANG, Z. H., AND A. D. YODER. 1999. Estimation of the transition/transversion rate bias and species sampling. J. Mol. Evol. 48:274–283.
YE, N., G. P. HOLMQUIST, AND T. R. O'CONNOR. 1998. Heterogeneous repair of *N*-methylpurines at the nucleotide level in normal human cells. J. Mol. Biol. 284:269–285.

First submitted 1 August 2001; reviews returned 10 January 2002; final acceptance 14 March 2002
Associate Editor: Chris Simon

APPENDIX

GenBank accession numbers used in this study. *Coturnix coturnix*: QULTROPIA, fast skeletal muscle troponin I gene; AB007195, TAP2; AB005533, major histocompatibility complex (MHC) class I antigen; QULAB, aromatase; AB005532, MHC class I antigen; AB024279, tyrosinase; AB005529, MHC class I antigen; AB000967, acid α glucosidase; AB005530, MHC class I antigen; AF231111, GLI3 gene N- and C-terminal fragments; AB005531, MHC class I antigen; CCJ002238, qMEF2D gene; AF139128, troponin T isoform; CJU56840, BKJ gene; CCTPMY01, α-tropomyosin; QULNFLW, neurofilament-L; CCU53861, myosin heavy chain 3; AF189778, bone morphogenetic protein receptor IB; QUCSRC3, c-src; AB006754, acid α glucosidase; QULQUOX7, homeobox protein; CCNR13, NR-13 gene; CCQR1G, QR1 gene; CCT64CLU, clusterin gene. *Gallus gallus*: GGVITIIG, vitellogenin II gene; GDLIPLIP, ipoprotein lipase gene; GDCOL6A2G, type VI collagen subunit α2; GGU83833, T-cell receptor

α chain gene; GDCOL6A1A, collagen α 1 type VI; GGD390, connectin/titin; GGDEF1C, transcriptional repressor delta EF1; GGDYS, dystrophin; GGLRPA2MR, LRP/α-2-macroglobulin receptor; AF143730, recombination activating protein 1 (RAG-1): GGA012570, sequence downstream of β-globin locus; CHKHBBRE, rho-globin, β-H globin, β-A globin, ε-globin, and olfactory receptor-like protein COR3′β genes; AB019555, pro-opiomelanocortin; AF062636, collagen type XII α-1; CHKFRA2A1, fra-2 oncogene; U67275, grf/pacap gene; AF077830, nuclear factor CTCF gene; SEG_AB050938S, transcription factor Foxa2; CHKPRO, p20K gene; AF246975, Mox-1 gene; AB030749, prolactin receptor; AF173612, 18S ribosomal RNA (rRNA) gene; GGCALB, conalbumin (ovotransferrin); AB029075, immunoglobulin mu heavy chain; GGBLOCUS, contig of 3 MHC cosmids cB12, c4.5, and cBF23; AB017036, skeletal muscle troponin T; GGRYR3, ryanodine receptor type 3; GGA246055, cosmid mapping to chromosome 1; GGAJ5158, DCoH gene; GGEAP300, EAP-300 gene; CHKMYHE, embryonic myosin heavy chain gene; CHKCRYD, delta-1 and delta-2 crystallin genes; GGA224516, interleukin (IL) 2 gene; GGU77715, POU gene; AF082667, class II cytokine receptor gene cluster; GGMY05, myosin light chain; GGCERBA2, c-erb A gene; GGA9800, putative IL-8 gene; AB022344, riboflavin binding protein; AF105022, glutamine synthetase gene; GGY18681, genomic DNA, 13.8 kb upstream of the α-globin gene; CHKOVAL, ovalbumin gene. *Agelaius phoeniceus*: AF030997, MHC class II B gene; AF170972, MHC-bearing cosmid, complete sequence. *Anas platyrhynchos*: APHISH1, 18S RNA gene, complete sequence; DUKMTRGTGN, 16S rRNA gene 3′ end, RNA-Leu gene, and ND1 gene 5′ end; APU06050, delta-1 crystalline gene 3′ region, delta-2 crystalline gene 5′ region, and intragenic spacer with CR1 repetitive element; APHISH1, histone H1 gene; S73733, acyl CoA-binding protein/diazepam-binding inhibitor-endozepine homolog; APU64985, serum amyloid A gene; APU60144, replication factor C large subunit; AF039749, carboxypeptidase D mRNA; AF137264, glycine decarboxylase p protein mRNA; DUKFASA, S-acyl fatty acid synthase thioesterase gene; APIFNG, interferon gene. *Cairina moschata*: CIIIGLVA6, Ig germline light chain J-region gene; CIIIGLVA5, Ig germline light chain V1-region gene; CIIHGAP, embryonic α-globin pi gene; CMBGA2B2, rearranged β-globin gene; CMEGA2E2, ε-globin gene; CMHIST34, H3 and H4 histone genes. *Columba livia*: AF173630, 18S RNA gene; CLRHII1, RH2 opsin gene; AB001981, α-D globin, α-A globin; CLU50598, pineal organ–specific opsin gene; PGNANXN01, annexin I (cp35) gene; AF018267, nucleoside diphosphate kinase (NDPK) gene; AF018266, NDPK gene; AB017906, gene for feather keratin; PGNCP37, annexin I (cp37) gene. *Dromaius novaehollandiae*: AF173610, 18S RNA gene; DNAJ2924, 12S rRNA, tRNA-Val, and 16S rRNA genes; AB006694, iron responsive element binding protein; AB006695, ZOV3 gene. *Grus americana*: AF033107, B-G-like protein gene. *Grus canadensis*: AF173632, 18S RNA gene; AF033106, MHC class I heavy chain (f51) mRNA; AF143732, (RAG-1) *Hirundo rustica*: HRU9MICST, microsatellite HrU9. *Carpodacus mexicanus*: AF205032, cosmid containing MHC *CameDAB1* and serine–threonine kinase genes. *Meleagris gallopavo*: MGU13978, β-4C-adrenergic receptor (ADRB4C) gene; MGPROLAC1, prolactin gene; AF006002, subgroup E ALV receptor mRNA. *Mergus serrator*: MRGRBMII, retropseudogene-like repetitive element I (RBMI). *Nothoprocta ornata*: AF173606, 18S rRNA gene; TTAJ2921, 12S rRNA, tRNA-Val, and 16S rRNA genes. *Opisthocomus hoatzin*: OPMLYSAH, lysozyme gene fragment. *Passer montanus*: AF143738, recombination activating protein 1 (RAG-1). *Rhea americana*: AF173608, 18S RNA gene; RAAJ2923, 12S rRNA, tRNA-Val, and 16S rRNA genes. *Serinus canaria*: SEINMYC1, N-myc gene. *Struthio camelus*: AF173607, 18S rRNA gene; AB005912, ZOV3 gene; AF143727, recombination activating protein 1 (RAG-1); SCMITSEQ1, tRNA-Phe gene partial sequence, 12S rRNA gene. *Sturnus vulgaris*: AF113513, estrogen receptor β mRNA. *Alligator mississippiensis*: ALLRRTRA, 12S rRNA gene; AF173605, 18S rRNA gene; AF143724, recombination activating protein 1 (RAG-1).