GC-Biased Gene Conversion Impacts Ribosomal DNA Evolution in Vertebrates, Angiosperms, and Other Eukaryotes

Juan S. Escobar,*†,[1] Sylvain Glémin,[1] and Nicolas Galtier[1]

[1]Institut des Sciences de l'Evolution, Unité Mixte de Recherche 5554 Centre National de la Recherche Scientifique, Université Montpellier II, Montpellier, France

†Present address: Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada.

***Corresponding author:** E-mail: jsescobar2002@yahoo.fr.

Associate editor: Manolo Gouy

Research article

Abstract

Ribosomal DNA (rDNA) is one of the most conserved genes in eukaryotes. The multiples copies of rDNA in the genome evolve in a concerted manner, through unequal crossing over and/or gene conversion, two mechanisms related to homologous recombination. Recombination increases local GC content in several organisms through a process known as GC-biased gene conversion (gBGC). gBGC has been well characterized in mammals, birds, and grasses, but its phylogenetic distribution across the tree of life is poorly understood. Here, we test the hypothesis that recombination affects the evolution of base composition in 18S rDNA and examine the reliability of this thoroughly studied molecule as a marker of gBGC in eukaryotes. Phylogenetic analyses of 18S rDNA in vertebrates and angiosperms reveal significant heterogeneity in the evolution of base composition across both groups. Mammals, birds, and grasses experience increases in the GC content of the 18S rDNA, consistent with previous genome-wide analyses. In addition, we observe increased GC contents in Ostariophysi ray-finned fishes and commelinid monocots (i.e., the clade including grasses), suggesting that the genomes of these two groups have been affected by gBGC. Polymorphism analyses in rDNA confirm that gBGC, not mutation bias, is the most plausible explanation for these patterns. We also find that helix and loop sites of the secondary structure of ribosomal RNA do not evolve at the same pace: loops evolve faster than helices, whereas helices are GC richer than loops. We extend analyses to major lineages of eukaryotes and suggest that gBGC might have also affected base composition in *Giardia* (Diplomonadina), nudibranch gastropods (Mollusca), and Asterozoa (Echinodermata).

Key words: base composition, concerted evolution, isochors, recombination, ribosome, secondary structure.

Introduction

Genes encoding ribosomal RNA (rRNA) are among the most utilized and conserved genes in eukaryotes. Conservation has been demonstrated in both linear sequences and secondary structures, which suggests that rRNA is under strong purifying selective pressure. Eukaryote genomes organize the single rRNA of the small ribosomal subunit (18S) and two of the rRNAs of the large ribosomal subunit (5.8S and 28S RNA) into a shared transcription unit. Transcription units form multigene families in long tandem arrays, the ribosomal DNA (rDNA) loci, carried by one or a small number of chromosomes (Dover 1994; Gonzalez and Sylvester 2001; Eickbush TH and Eickbush DG 2007). Variable numbers of rDNA loci occur in different species (from a few tens to more than 50,000 copies), so that, depending on the considered species, rDNA can contribute a substantial percentage of the nuclear genome (Long and Dawid 1980; Rogers and Bendich 1987).

One of the most remarkable features of rDNA loci is the homogeneity in sequence among transcription units within a genome. This ability to change sequences in a highly orchestrated manner, that is, to spread or eliminate new mutations arriving in one unit to adjacent units, is known as concerted evolution (Dover 1994; Elder and Turner 1995; Liao 1999; Eickbush TH and Eickbush DG 2007). Concerted evolution of rDNA leads to high redundancy of ribosomes, which is presumably beneficial to the organism as all ribosomal subunits are equally compatible with other components of the cellular translational machinery (Averbeck and Eickbush 2005). Concerted evolution of rDNA occurs in many species, including angiosperms (Flavell and O'Dell 1976; Wendel et al. 1995; Franzke and Mummenhoff 1999; Fuertes Aguilar et al. 1999; Lim et al. 2000; Koch et al. 2003; Kovarik et al. 2004, 2005; Rauscher et al. 2004) and vertebrates (Brown et al. 1972; Arnheim et al. 1980, 1982; Hillis et al. 1991).

Two classes of mechanisms have been put forward to explain the concerted evolution of rDNA (reviewed in Kupriyanova 2000; Eickbush TH and Eickbush DG 2007): unequal crossing over between homologous rDNA units (Brown et al. 1972; Petes 1980; Szostak and Wu 1980; Endow and Komma 1986; Schlotterer and Tautz 1994) and gene conversion, that is, the copy and paste of one genomic copy onto another, whether they are orthologous or not (Hillis et al. 1991; Gangloff et al. 1996; Benevolenskaya et al. 1997; Fuertes Aguilar et al. 1999; Liao 2000; Lim et al. 2000). Today, it is widely accepted that concerted evolution in rDNA results from a combination of both these mechanisms (Eickbush TH and Eickbush DG 2007), the two being mechanistically related to the molecular process of homologous recombination.

Interestingly, recombination is associated with GC bias in a number of organisms (reviewed in Marais 2003; Duret and Galtier 2009). Such a bias may be produced by mutation, selection favoring G and C bases or GC-biased gene conversion (gBGC). gBGC is a bias in the cellular DNA repair machinery that results in a meiotic segregation distortion favoring G and C over A and T alleles, hence increasing GC content in the long term. gBGC affects single-copy genes (Romiguier et al. 2010) and multicopy genes undergoing concerted evolution (Galtier 2003; Kudla et al. 2004) in mammals (Galtier et al. 2001; Montoya-Burgos et al. 2003; Meunier and Duret 2004; Spencer 2006; Duret and Arndt 2008), birds (Webster et al. 2006), yeasts (Birdsell 2002; Mancera et al. 2008), and grasses (Glémin et al. 2006; Haudry et al. 2008; Escobar et al. 2010) and to a lower extent *Drosophila* (Galtier et al. 2006; Haddrill et al. 2007). gBGC is a recently discovered evolutionary force, which not only impacts GC-content dynamics but also affects functional components of the genome by impeding the action of natural selection (the Achilles' heel hypothesis; Galtier et al. 2009). Birdsell (2002) and Lynch (2007) suggest that gBGC could be a ubiquitous mechanism resulting from the generalized AT bias of the mutation process. The recent discovery of widespread AT-mutation bias and GC-fixation bias in bacteria (Hershberg and Petrov 2010; Hildebrand et al. 2010) somewhat confirms this view. In eukaryotes, gBGC has so far been studied in a handful of taxa for which genome-wide comparative data are available. Its phylogenetic distribution across the tree of life, therefore, is only vaguely understood despite the potential importance of this process in molecular evolution.

rDNA is an ideal target to study the impact of gBGC in the evolution of eukaryotic genomes because it has a very specific evolutionary dynamics (i.e., long-term high recombination rate thanks to concerted evolution) and because it has been sequenced in a very large number of species. We reasoned that the analysis of GC-content variations in rDNA could give insight into the prevalence of gBGC in various lineages of the eukaryotic tree. In this paper, we analyze the evolution of base composition in the 18S rDNA and examine the reliability of this molecule as a marker of gBGC. We first analyze vertebrates and angiosperms because these two groups have been thoroughly studied in terms of base composition (e.g., Eyre-Walker and Hurst 2001; Wang et al. 2004) and contain clades in which gBGC has been documented (e.g., mammals, birds, and grasses). These groups serve as positive controls in our tests, that is, if rDNA is an appropriate marker of gBGC, we expect to find signatures of GC enrichment in, at least, these three groups. Because the secondary structure of the 18S rRNA is well known, we quantify the effect of gBGC in helix and loop sites of the molecule. Single nucleotide polymorphisms (SNPs) data are used to confirm that gBGC (or selection for GC content), not mutation biases, is at work in rDNA. Finally, we extend the analysis of base composition in 18S rDNA to various groups of eukaryotes to pinpoint potential gBGC events across the eukaryotic tree.

Materials and Methods

Interspecific Data Sets in Vertebrates and Angiosperms

Sequences of the 18S rDNA of vertebrates and angiosperms were obtained from the SILVA database (Pruesse et al. 2007). We downloaded 1,655 sequences (526 species) of vertebrates and 6,085 sequences (2,186 species) of angiosperms. Sequences were edited and aligned with ARB (Ludwig et al. 2004). Edition consisted in filtering out short and low-quality sequences. In the vertebrate data set, we first selected sequences ≥1,200 nt (*nuc_gene_slv* parameter) and ≥95% alignment quality (*align_quality_slv* parameter). Then, we ordered selected sequences by percentage of quality (*seq_quality_slv* and *pintail_slv* parameters), next by alignment quality, and finally by sequence length. In this way, we assembled 287 high-quality sequences (287 species). Angiosperm sequences were selected in a similar way. However, because there were much more sequences of angiosperms than vertebrates, our criteria could be more restrictive: ≥1,700 nt and ≥98% alignment quality for the first filter. Then, we ordered and selected sequences as above and randomly chose one sequence per genus. Our data set of angiosperms consisted of 1,049 high-quality sequences (1,049 species).

Selected sequences were aligned using the ARB aligner, which uses an 18S rRNA secondary structure backbone. Three alignments were obtained in each data set: 1) the complete alignment (i.e., all sites); 2) the alignment of paired double-strand regions (i.e., helix sites); and 3) the alignment of unpaired single-strand regions (i.e., loop sites). Ambiguously aligned sites, as well as sites including gaps in more than 50% of sequences, were excluded of the final alignment with Gblocks 0.91b (Castresana 2000) set to the following parameters: minimum number of sequences for a conserved position = $n/2 + 1$ (where n = number of taxa); minimum number of sequences for a flanking position = $0.85 \times n$; maximum number of contiguous nonconserved positions = 8; minimum length of a block = 2; and allowed gap positions = with half. The final alignments of all, helix, and loop sites in the vertebrate data set contained 1,342, 894, and 480 nt, respectively. In the angiosperm data set, alignments consisted of 1,400, 969, and 454 nt, respectively. Note that the sum of helix and loop sites does not correspond to the number of aligned positions at all sites because each of the three alignments was separately treated with Gblocks.

The CpG methylation-deamination process leads to hypermutability in vertebrates and angiosperms (Bird 1980; Kovarik et al. 1997; Arndt et al. 2003). To account for the potential influence of CpG sites in our results, we cleaned up the raw alignments of all sites by eliminating sites involved in CpG, TpG, or CpA pairs in at least 50% of the sequences of the alignment. We only analyzed alignments including all sites because sites in helix or loop alignments were not always physically adjacent. CpG-filtered alignments were treated with Gblocks as described above. The alignment without CpG consisted of 1,102 sites in vertebrates and 1,096 sites in angiosperms.

All alignments are available for download at http://mbb.univ-montp2.fr/MBB/subsection/data.php?section=2&from_dts=5&nb_dts=5.

Vertebrate and Angiosperm Phylogenies

We inferred phylogenetic trees of vertebrates and angiosperms with PhyML 3.0 (Guindon and Gascuel 2003) using the alignments including all sites. We used a general time reversible (GTR) + I + G model of sequence evolution, four categories for the gamma distribution, parsimony starting trees, and subtree pruning and regrafting branch swapping. Generally, species belonging to most terminal clades grouped together, although relationships among deep clades were not well resolved, consistent with previous reports (Hasegawa and Hashimoto 1993; Soltis et al. 2000; Winchell et al. 2002; Mallatt and Winchell 2007; Swalla and Smith 2008; The Angiosperm Phylogeny Group 2009). For this reason, we adjusted the trees to be congruent with the most recent and accepted phylogenies of vertebrates and angiosperms, whereas relationships among more recently derived groups were kept as inferred by PhyML. In vertebrates, we used the backbone phylogeny proposed by Alfaro et al. (2009) and manually adjusted it following other publications (van Tuinen et al. 2000; Venkatesh et al. 2001; Hudelot et al. 2003; Miya et al. 2003; Murata et al. 2003; Brinkmann et al. 2004; Douzery and Huchon 2004; Inoue et al. 2004; Townsend et al. 2004; Lavoué et al. 2005; Sullivan et al. 2006; Hugall et al. 2007; Mallatt and Winchell 2007). In angiosperms, we used the phylogeny suggested by The Angiosperm Phylogeny Group (2009) and adjusted it (Olmstead et al. 2000; Karehed 2001; Lundberg 2001; Tamura et al. 2004; Nickrent et al. 2005; González et al. 2007; Wanga et al. 2009; Worberg et al. 2009). The main analyses presented here were obtained with the modified trees. However, to test the robustness of our results, we also performed analyses using unmodified trees.

Analyses of Heterogeneity in Base Composition

We estimated the GC content at all, helix, and loop sites. We first performed analysis of variance (ANOVA) using the proportion of GC (data arcsin square root transformed) in each of these three alignments as variable and clades as factors with R 2.9.1 (R Development Core Team 2009). Clades were delimited in order to work at taxonomic levels similar to those in which gBGC has been documented (mammals, birds, and grasses). In vertebrates, clades within Actinopterygii (ray-finned fishes) generally corresponded to taxonomic orders and in Sarcopterygii (coelacanths, lungfishes, and tetrapods) to taxonomic classes (fig. 2). In angiosperms, clades were consistent with taxonomic orders or superorders of APG III (The Angiosperm Phylogeny Group 2009) (fig. 3).

We also tested heterogeneity in the GC content using nonhomogeneous models of sequence evolution and the phylogenies of vertebrates and angiosperms. These models were fitted with BPPML (Dutheil and Boussau 2008) and NHML (Galtier and Gouy 1998), which use a maximum likelihood approach and account for nonstationary (ancestral and current GC content can differ) and nonhomogeneity (branches can have distinct GC) in base composition across the phylogeny. We estimated the GC content at any node (or groups of nodes) and the equilibrium GC content (GC*) at any branch (or group of branches) of the phylogenetic tree. GC* is defined as:

$$GC* = \frac{AT \rightarrow GC}{AT \rightarrow GC + GC \rightarrow AT}, \qquad (1)$$

where AT → GC refers to the substitution rate from A or T to G or C bases, and GC → AT holds for the inverse (Sueoka 1962). GC* is a more appropriate measure of the evolutionary dynamics than the current GC (Meunier and Duret 2004; Duret and Arndt 2008).

We fitted hierarchical models of sequence evolution to test whether branches underwent similar evolution of base composition. We used likelihood-ratio tests (LRT) to assess whether more complex models provided a significantly improved fit compared with simpler models. In vertebrates, GC and GC* were estimated using eight hierarchical nested models (see table 1): 1) one estimate for all branches of the phylogeny (homogeneous); 2) cyclostomes and the remaining clades (jawed vertebrates); 3) we distinguished sharks and Euteleostomi among the jawed vertebrates; 4) we distinguished Actinopterygii and Sarcopterygii among Euteleostomi; 5) we distinguished Coelacanths, lungfishes, and tetrapods among Sarcopterygii; 6) we distinguished amphibians and amniotes among tetrapods; 7) terminal clades (as shown in fig. 2); and 8) each branch of the tree has its own GC and GC*. Note that in models 3–7, different GC and GC* were estimated in internal branches. In angiosperms, six hierarchical nested models were fitted (see table 1): 1) homogeneous; 2) Amborellales, Nymphaeales, Austrobaileyales, magnoliids–Chloranthales, monocots, Ceratophyllales, and eudicots have different GC and GC*; 3) we distinguished commelinids and non-commelinids among monocots; 4) we distinguished basal eudicots (Ranunculales, Sabiaceae, Proteales, Buxales, and Trochodendrales) and core eudicots among eudicots; 5) terminal clades (as shown in fig. 3); and 6) each branch of the tree has its own GC and GC*. As above, in models 2–5 different GC and GC* were estimated in internal branches.

Analyses in vertebrates and angiosperms were performed using all, helix, and loop sites separately. Additionally, we performed analyses using the phylogenetic trees inferred with the 18S rDNA and the alignments in which CpG sites were removed. In all cases, we analyzed alignments in which invariable sites were removed. This is because invariable sites are presumably under strong selective constraints and were useless for our analyses of the evolution of base composition. As expected, removing invariable sites affected GC but not GC* estimates.

Table 1. Hierarchical Models of Sequence Evolution of the 18S Ribosomal DNA.

Model		−lnL	Dev.	df	*P* value
Vertebrates					
1.	Homogeneous	13568.60			
2.	Cyclostoma + Jawed vertebrates	13567.59	2.01	1	0.1558
3.	Cyclostoma + Chondrichthyes + Euteleostomi	13554.27	26.64	2	1.64×10^{-6}
4.	Cyclostoma + Chondrichthyes + Actinopterygii + Sarcopterygii	13535.99	36.56	1	1.48×10^{-9}
5.	Cyclostoma + Chondrichthyes + Actinopterygii + Coelacanth + Dipnoi + Tetrapoda	13532.88	6.21	1	0.0127
6.	Cyclostoma + Chondrichthyes + Actinopterygii + Coelacanth + Dipnoi + Amphibia + Amniota	13527.21	11.35	1	0.0008
7.	Terminal clades (as shown in fig. 2)	13481.32	91.77	31	6.32×10^{-8}
8.	One GC* per branch	13202.85	556.95	533	0.2287
Angiosperms					
1.	Homogeneous	11598.04			
2.	Amborellales + Nymphaeales + Austrobaileyales + Magnolids–Chloranthales + Monocots + Ceratophyllales + Eudicots	11540.53	115.03	10	5.14×10^{-20}
3.	Amborellales + Nymphaeales + Austrobaileyales + Magnolids–Chloranthales + commelinids + non-commelinids + Ceratophyllales + Eudicots	11537.41	6.24	2	0.0441
4.	Amborellales + Nymphaeales + Austrobaileyales + Magnolids–Chloranthales + commelinids + non-commelinids + Ceratophyllales + basal eudicots + core eudicots	11537.00	0.82	1	0.3660
5.	Terminal clades (as shown in fig. 3)	11472.39	129.22	64	2.62×10^{-6}
6.	One GC* per branch	11335.82	273.13	162	1.08×10^{-7}

NOTE.—lnL: log likelihood; Dev.: residual deviance; degrees of freedom (df): residual degrees of freedom. Analyses performed in all sites of the molecule.

Because the data set in angiosperms was too big for phylogenetic analyses of base composition (1,049 sequences), we reduced the number of sequences to two per taxonomic order (according to The Angiosperm Phylogeny Group 2009) when possible. For this, we estimated the percentage of identity among sequences of each order and randomly selected pairs of sequences among those with level of identity equal to the median value of the order. We preferred to select pairs of sequences showing median rather than maximal divergence to avoid picking sequences with peculiar evolutionary dynamics (e.g., pseudogenes) that would have remained undetected along the filtering process described above. The data set of angiosperms on which we fitted the nonhomogeneous models of sequence evolution consisted of 121 sequences (121 species).

To test whether base composition evolved differently in helix and loop sites, we performed an LRT to compare the log likelihood of a model including all sites, and the sum of log likelihoods of models including helix and loop sites analyzed separately. This test uses the total number of parameters estimated in each model as the number of degrees of freedom. If the GC content evolved differently between helices and loops, we expected that the sum of log likelihoods of models separating loops and helices would significantly improve the fit relative to the model considering all sites. Analyses were performed in vertebrates and angiosperms separately using variable and invariable positions of the alignments. The alignment of all sites was obtained by concatenating helix and loop alignments treated with Gblocks. We estimated log likelihoods using a model assuming one GC* per terminal clade (models 7 in vertebrates and 5 in angiosperms) and one free GC* for each internal branch. In this analysis, branch lengths were fixed (to values estimated with PhyML using a GTR + I + G model) to be sure that the detected heterogeneity between loop and helix sites reflects variations in GC-content dynamics, not in lineage-specific evolutionary rate.

Polymorphisms in rDNA

SNPs from rDNA loci were retrieved from the Polymorphix 2 database (Bazin et al. 2005). These data contained partial or complete sequences of any of the following: 5S, 5.8S, 18S, 28S, internal transcribed spacers (ITS-1 and ITS-2), and intergenic spacers (IGS). In many cases, alignments consisted of contigs of these regions (e.g., partial 18S, completes ITS-1, 5.8S and ITS-2, and partial 28S). In angiosperms, 311 alignments (267 species) containing five sequences of the same species or more were retrieved. In vertebrates, as few as 18 alignments (15 species) containing four sequences of the same species or more were available in the Polymorphix 2 database. Because the last update of this database was in November 2004, we complemented the SNP data set in vertebrates with data from GenBank (requested on 21 May 2010). We performed searches of nuclear rDNAs using the following command line: (vertebrate* NOT mitochond*) AND (5S OR 5.8S OR 18S OR 28S OR spacer). The final data set in vertebrates consisted of 51 alignments (42 species) containing four sequences of the same species or more.

Intraspecific rDNA sequences were aligned with Prank v. 100311 (Loytynoja and Goldman 2005). The resulting alignments were filtered with Gblocks using the same parameters described above. Note that all our SNP alignments consisted of sequences from just one species, hence SNPs were not oriented. We deliberately analyzed nonoriented SNPs given the rarity of appropriate outgroups across the various alignments. Polymorphism analyses were performed using a homemade program that calculates the number of polymorphic AT ↔ GC sites and, among these sites, the number of sites with GC frequency >0.5, <0.5, and $=0.5$. SNP counts were pooled by clade (as shown in figs. 2 and 3) and two-sided binomial tests were performed to test for mutational AT ↔ GC equilibrium. At mutational equilibrium, one expects AT → GC = GC → AT, that is, equal amounts of SNPs in which GC is the most frequent allele and SNPs in which AT is the most frequent allele.

We analyzed a subset of sufficiently annotated alignments of vertebrates and angiosperms to determine whether forces shaping base composition act differently in the different fragments of rDNA loci (as documented for noncoding polymorphisms in the X chromosome of *D. simulans*; Haddrill and Charlesworth 2008). We estimated the mean per site polymorphism rate (number of SNPs/alignment length) of each alignment and compared fragments (5S, 5.8S, 18S, 28S, ITS-1, ITS-2, and IGS) using nonparametric Kruskal–Wallis tests with R 2.9.1 (R Development Core Team 2009).

Interspecific Data Set in Other Eukaryotes

We downloaded sequences of the 18S rDNA of most major lineages of eukaryotes from the SILVA database. Sequences were edited as described above using the length and quality thresholds given in table 4. The final data set of eukaryotes consisted of 11,259 sequences (6,689 genera, all lineages confounded) and included 2,506 Viridiplantae, 1,993 Fungi, and 5,211 Metazoa. We determined the distribution of GC content per genus (across species) within each eukaryotic lineage.

Results

Heterogeneity in Base Composition of the 18S rDNA in Vertebrates and Angiosperms

Base composition of the 18S rDNA significantly varies among taxonomic groups in both vertebrates and angiosperms ($P < 2.20 \times 10^{-16}$ in all ANOVAs). This is true for alignments including all sites, helix, and loop sites. Noteworthy, helices are GC richer than loops (fig. 1), whereas loops evolve faster (vertebrate tree lengths: helix = 1.67; loop = 2.92; angiosperm tree lengths: helix = 20.60; loop = 31.02; differences in tree length between vertebrates and angiosperms essentially reflect differences in the number of analyzed species). Among vertebrates, four GC-rich clades stand out from the remaining groups: birds, Cypriniformes, mammals, and Siluriformes (fig. 1*A*). In angiosperms, heterogeneity in base composition is less remarkable than in vertebrates. However, monocots, especially commelinids, are GC richer than all other flowering plants (fig. 1*B*).

Nonhomogeneous models of sequence evolution, explicitly taking into account phylogenetic relationships, confirm heterogeneity in base composition among major clades in vertebrates and angiosperms (table 1). In vertebrates, almost all clades within Actinopterygii and Sarcopterygii have increased GC content relative to their ancestors. Ostariophysi (i.e., Siluriformes, Clupeiformes, and Cypriniformes), mammals, and birds are remarkable in this respect. GC content in the ancestor of Ostariophysi has increased ∼5% relative to the ancestor of Actinopterygii. Consistently, GC* at the base of this clade is high (0.79 overall sites), although GC* values are lower in more recently derived branches (fig. 2). Interestingly, GC* is greater in helix (0.81) than loop (0.73) sites in this branch. Mammals and birds have experienced an increase of ∼4% in the GC content relative to the ancestral Sarcopterygii. Analysis of GC* suggests that the increase took place at the base of amniotes (GC* overall sites = 0.89) and affected loop (1.00) more than helix (0.79) sites (fig. 2). Unlike Clupeiformes, Cypriniformes, and Siluriformes, GC* is high in extant mammals (0.69 overall sites), birds (0.85), and Lepidosauria (0.86), although low in the turtle (0.51) and crocodiles (0.49)—but the numbers of sequences analyzed in these two clades were low (one and three, respectively). On the other hand, a few clades of vertebrates show a decrease in GC content (low GC*), including Chondrostei, Coelacanth, and Dipnoi (lungfishes) (fig. 2). Importantly, the GC dynamics significantly vary between helix and loop sites in vertebrates (LRT: $\chi^2_{42}=185.82$, $P = 5.35 \times 10^{-20}$).

Among angiosperms, monocots, especially commelinids (grasses, palm trees, gingers, bananas, arrowroots, and allied), show the most important increase in GC content relative to the ancestor of all angiosperms but Amborellales (fig. 3). In commelinids, this increase is of about 3% for helix and 7% for loop sites. GC* are high in commelinids for all sites, helix, and loop sites and higher in helices than loops. Indeed, GC* in loop sites in this clade are the highest of all angiosperms. GC* values in the ancestor of magnoliids, Chloranthales, monocots, and Eudicots (MCME) are also high, although most lineages have apparently engaged in GC erosion (note the lower GC* values of most terminal clades relative to more internal branches; fig. 3). Indeed, most Eudicots have low GC*, especially Gunnerales, Sabiaceae, Santalales, and Trochodendrales. This suggests that an evolutionary force increasing the GC content was active at some point in the evolution of angiosperms, and is no longer active in most groups (especially Eudicots), with the notable exception of commelinids. As in vertebrates, the GC dynamics significantly varies between helix and loop sites in angiosperms (LRT: $\chi^2_{52}=299.05$, $P = 2.08 \times 10^{-36}$).

We were concerned about the effects of CpG hypermutable sites and the phylogenetic trees on our results. As expected, there is a reduction in the GC content of sequences without CpG sites. Nevertheless, the patterns observed in vertebrates and angiosperms are robust and do not depend on CpG sites (with the exception of birds; supplementary figs. S1 and S2, Supplementary Material online)

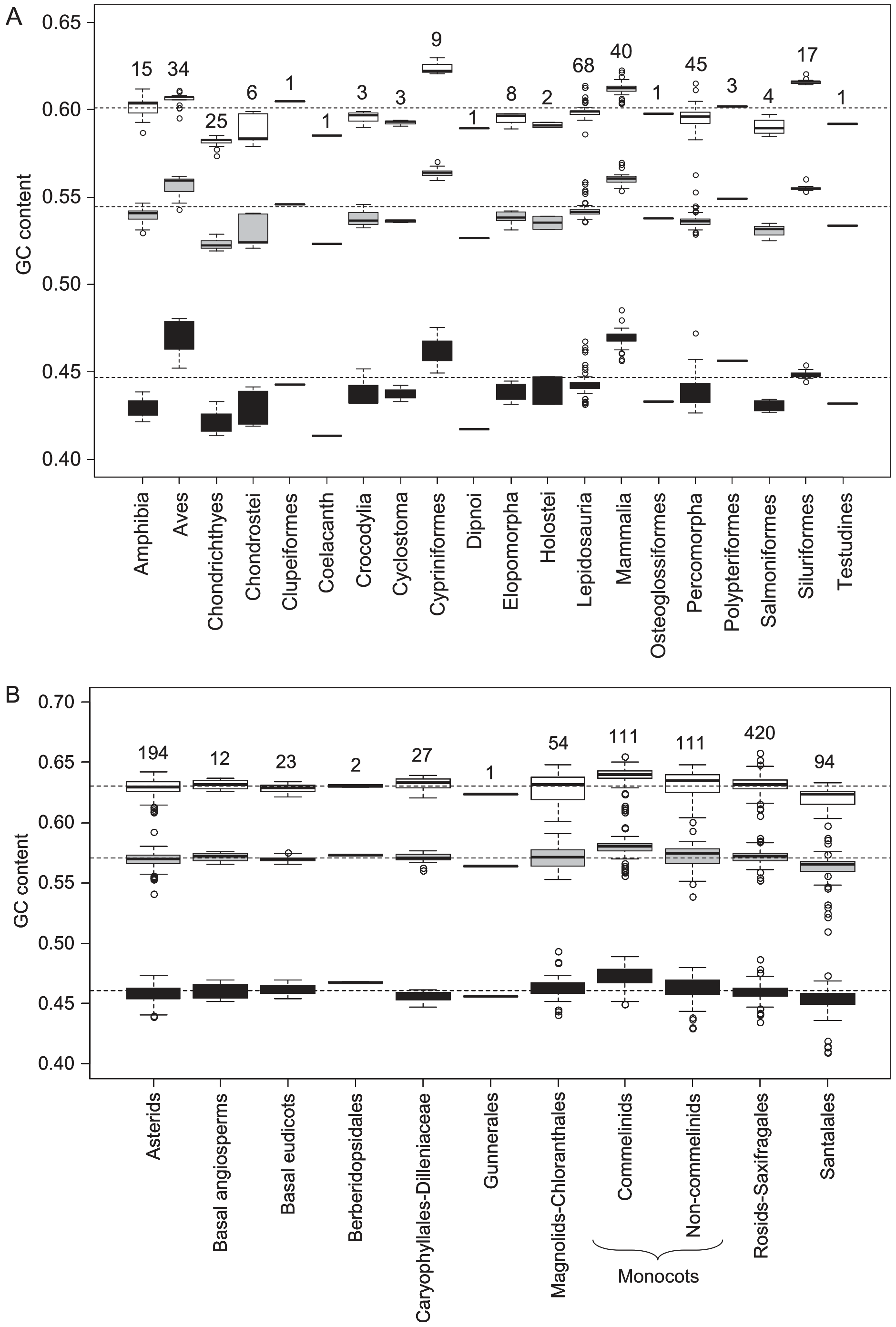

Fig. 1. Distribution of the GC content of 18S ribosomal DNA in vertebrates (*A*) and angiosperms (*B*). Numbers above boxplots are the number of analyzed genera. Open boxes: helix sites; gray boxes: all sites; and filled boxes: loop sites. Horizontal dotted lines represent means across all species for each type of sites (up: helix; middle: all; and low: loop).

or on the phylogenetic trees (supplementary figs. S3 and S4, Supplementary Material online).

GC Content and Polymorphisms in Vertebrates and Angiosperms

We performed SNP analyses to discriminate between the potential evolutionary forces underlying rDNA GC-content evolution, namely gBGC (or selection for GC content) versus mutation biases. Specifically, we aimed at testing the null hypothesis of mutational equilibrium, which predicts equal average population frequencies for AT and GC alleles, whereas directional processes (gBGC or selection) predict higher frequencies for GC alleles (Duret et al. 2002; Galtier et al. 2006). Note that nonequilibrium mutation dynamics with an increasing bias toward GC → AT mutations also predicts higher GC allele frequencies. Because the multigenic nature of rDNA, we are not sure that we analyzed true intralocus polymorphism or

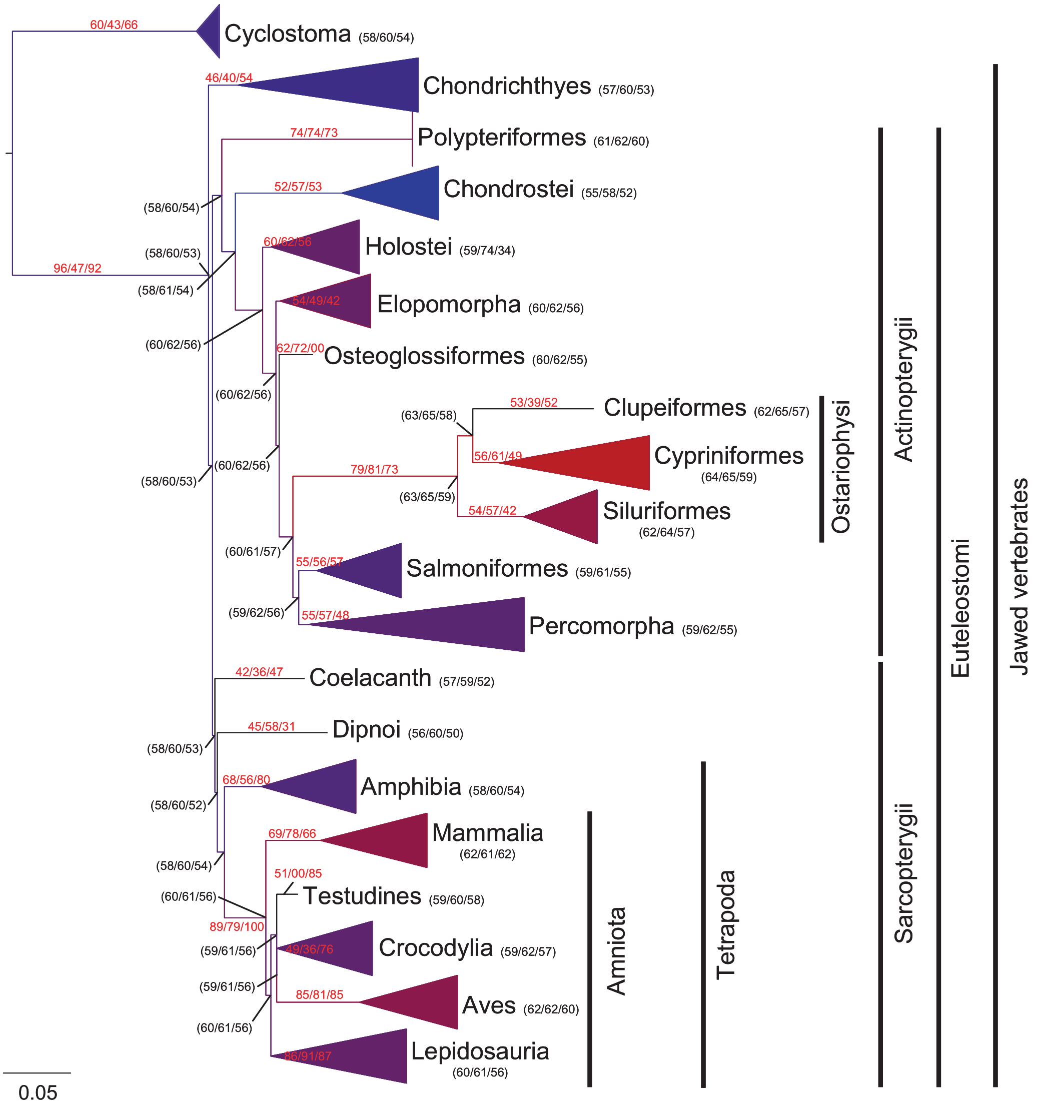

FIG. 2. Evolution of the GC content of 18S ribosomal DNA across the phylogeny of vertebrates. Only results of alignments with variable sites are shown. Values correspond to current GC (in parentheses in terminal branches), ancestral GC (in parentheses in internal nodes), or GC* (red characters along branches) at all/helix/loop sites; GC* are shown only for the most representative internal branches. Colors in terminal branches represent average GC content (blue: lowest GC; red: highest GC). Note that the color scale is relative to the data set and is not directly comparable with figure 3. Branch lengths are given in units of per site substitution rate.

between-loci differences. However, in a multigenic family, both intralocus and interlocus biased-gene conversion increases fixation rates as compared with the neutral unbiased case (Walsh 1985). Without gBGC, we also expect that, on average, the number of gene copies fixed for GC equals the number of gene copies fixed for AT. Predictions are therefore similar for intralocus and between-loci variation, so that our approach appears robust to the uncertainty about the nature of polymorphism data—in neither case does one expect asymmetrical SNP frequency spectra in absence of gBGC/selection.

In vertebrates, 673 polymorphic sites in rDNA loci were found out of a total of 60,586 aligned sites (1.11% of polymorphism). Overall clades and species, GC alleles in rDNA segregate at higher frequency than AT alleles (0.61 and 0.39, respectively; $P = 1.02 \times 10^{-8}$). GC alleles segregate at significantly higher frequencies than AT alleles in Amphibia, Chondrostei, and Cypriniformes (table 2). The average frequency of GC alleles in these clades spans 0.58–0.67. In angiosperms, the data set of polymorphisms in rDNA loci is much bigger than in vertebrates. It contained 3,425 polymorphic sites of a total of 164,127 aligned sites (2.09% of polymorphism). Analysis of this data set reveals that, overall clades and species, GC alleles in rDNA segregate at higher frequency than AT alleles (0.63 and 0.37, respectively; $P < 2.20 \times 10^{-16}$). Significant GC-biased spectra are found in asterids, Caryophyllales, monocots, Ranunculales, and rosids (table 2). The frequency of GC SNPs in these clades spans 0.62–0.71.

Not surprisingly, we found significant differences in the per site polymorphism rate among fragments of the rDNA loci in vertebrates (Kruskal–Wallis test: χ^2_6=27.51, $P <$ 0.0001) and angiosperms (χ^2_6=64.80, $P <$ 0.0001) (table 3). As expected, regions that are not directly involved in

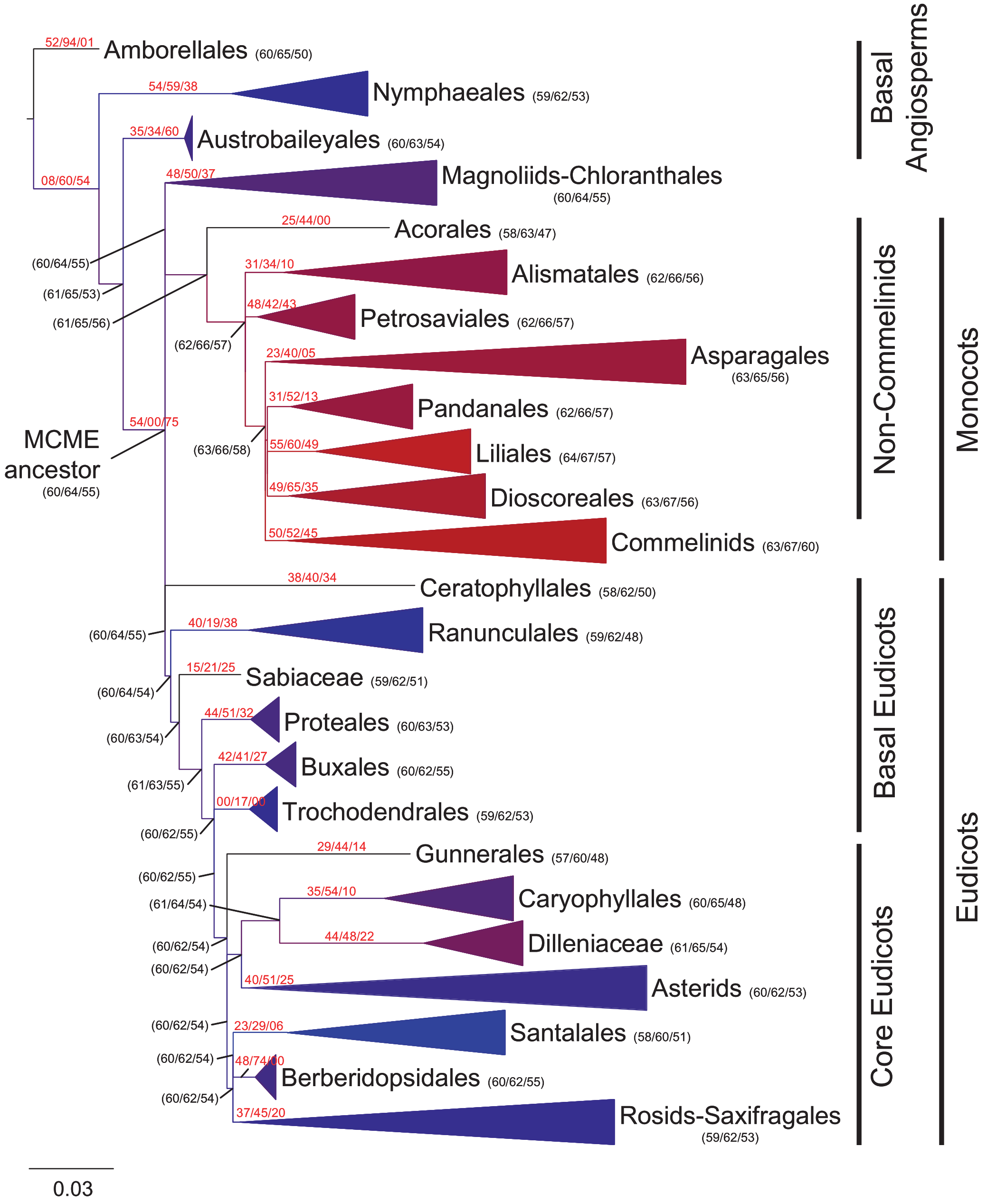

FIG. 3. Evolution of the GC content of 18S ribosomal DNA across the phylogeny of angiosperms. Only results of alignments with variable sites are shown. Values correspond to current GC (in parentheses in terminal branches), ancestral GC (in parentheses in internal nodes), or GC* (red characters along branches) at all/helix/loop sites; GC* are shown only for the most representative internal branches. Colors in terminal branches represent average GC content (blue: lowest GC; red: highest GC). Note that the color scale is relative to the data set and is not directly comparable with figure 2. Branch lengths are given in units of per site substitution rate. MCME: Magnoliids–Chloranthales–Monocots–Eudicots.

the ribosome structure (IGS, ITS-1, and ITS-2) are probably less constrained by selection, hence more likely to accumulate polymorphisms. Polymorphism rate is high in 5S sequences too, although this gene codes for a functional RNA. This is possibly because 5S sequences are usually poorly annotated and frequently correspond to the actual 5S fragment as well as part of the untranscribed spacer.

Heterogeneity in Base Composition of rDNA Among Eukaryotes

The GC content in rDNA varies significantly among major lineages of eukaryotes ($F_{11,6677} = 302.47$, $P < 2.20 \times 10^{-16}$). Differences are highly significant even after excluding a potential outlier (Diplomonadina; table 4, fig. 4A). We obtained the distribution of GC content in the 18S rDNA across eukaryotic genera and selected the 5% exhibiting the highest values (336 genera). These extreme values correspond to 1 Diplomonadina (*Giardia*), 330 Metazoa (2 annelids, 30 arthropods, 1 cephalochordate, 190 vertebrates, 56 echinoderms, 3 hemichordates, 47 mollusks, and 1 nematode), and 5 Viridiplantae (2 Ulvophyceae, 1 bryophyte, and 2 commelinid monocots) (supplementary table S1, Supplementary Material online).

Table 2. Polymorphism in Ribosomal DNA in Vertebrates and Angiosperms.

Clade	N_a	N_{sp}	AT > 0.5	GC > 0.5	P value	GC/(GC + AT)
Vertebrates						
Chondrostei	14	12	75	154	1.95×10^{-7}	0.67
Cypriniformes	16	11	76	103	0.05	0.58
Percomorpha	14	13	58	51	0.56	0.47
Amphibia	7	6	53	103	7.67×10^{-5}	0.66
Angiosperms						
Monocots	62	52	344	640	2.20×10^{-16}	0.65
Ranunculales	9	9	20	50	4.40×10^{-4}	0.71
Trochodendrales	1	1	29	25	0.68	0.46
Asterids	96	80	230	454	2.62×10^{-16}	0.66
Caryophyllales	10	9	32	69	2.96×10^{-4}	0.68
Rosids	128	111	581	898	2.20×10^{-16}	0.61
Saxifragales	5	5	29	24	0.58	0.45

NOTE.—N_a: number of alignments; N_{sp}: number of species; AT > 0.5: SNPs in which A or T is the majority allele; GC > 0.5: SNPs in which G or C is the majority allele. Clades are sorted by phylogenetic proximity.

We also performed separate analyses in Metazoa, Viridiplantae, and Fungi. In Metazoa, genera showing the 5% highest GC content in 18S rDNA are represented by 1 Polychaeta (Annelida), 3 insects, 3 Hemichordata, 13 Asteroidea echinoderms, 35 mollusks (1 Aplacophora, 2 Bivalvia, and 32 Doridina nudibranch), and 123 vertebrates (1 Elopomorpha, 7 Percomorpha, 18 Ostariophysi, 2 Polypteriformes, 2 Amphibia, 34 birds, 20 Lepidosauria, and 39 mammals) (fig. 4*B*). In Viridiplantae, the highest GC content corresponds to 2 genera of Ulvophyceae, 1 of Chlorophyceae, 6 of Trebouxiophyceae, 2 of Briophyta, and 82 of Tracheophyta, including 1 fern, 1 conifer, 9 rosids, 3 magnolids, 6 non-commelinid monocots, and 62 commelinids (supplementary fig. S5, Supplementary Material online). In fungi, the highest GC content corresponds to 38 Saccharomyceta (2 Dothideomycetes, 20 Eurotiomycetes, 1 Lecanoromycetes, 1 Lichinomycetes, and 14 Sordariomycetes) and 4 Agaromyceta (2 Boletales, 1 Hymenochaetales, and 1 Polyporales) (supplementary fig. S6, Supplementary Material online).

Discussion

Heterogeneity in Base Composition of the 18S rDNA

We show significant heterogeneity in base composition in rDNA across the vertebrate and angiosperm phylogenies. The average GC level observed among vertebrates in this study is 54.0%, very close to previous reports (55.4% in Xia et al. 2003; 54.3% in Wang et al. 2006; and 54.8% in Varriale et al. 2008). In angiosperms, base composition in rDNA has been less considered, and there is no study to compare with.

Table 3. Average Per site Polymorphism Rate (and the corresponding standard error mean) in Different Fragments of the Ribosomal DNA Loci.

Fragment	Vertebrates	Angiosperms
5S	0.0601 (0.0097)	0.0396 (0.0028)
IGS	0.0686 (0.0496)	0.0247 (0.0074)
ITS-1	0.0362 (0.0207)	0.0250 (0.0026)
ITS-2	0.0156 (0.0081)	0.0227 (0.0023)
18S	0.0085 (0.0069)	0.0183 (0.0085)
5.8S	0.0071 (0.0055)	0.0124 (0.0023)
28S	0.0021 (0.0015)	0.0075 (0.0028)

Within vertebrates, the highest GC content in rDNA was found in mammals, birds, and Ostariophysi (ray-finned fishes). In the latter, Siluriformes, Cypriniformes, and Clupeiformes exhibit high current GC content, although they do not display high GC*. However, GC* is high in the ancestral branch leading to Ostariophysi, suggesting that an episode increasing the GC content took place in the ancestry of this group. This would explain the high GC observed in current sequences. Among angiosperms, monocots, and especially commelinids, Dioscoreales, Liliales, and Pandanales, display the highest GC content (and high GC*) of all flowering plants. Interestingly, high GC* was inferred in the branch grouping magnoliids–Chloranthales and monocots and in the branch at the base of Eudicots.

Evolution of Base Composition in the rRNA Secondary Structure

According to our analyses, rDNA loops evolve 1.75 times faster than helices in vertebrates and 1.50 in angiosperms. Our estimates are quantitatively consistent with previous results in eukaryotes (e.g., 1.37-fold faster in Smit et al. 2007) and qualitatively with past observations in angiosperms (Soltis et al. 1997; Soltis PS and Soltis DE 1998). It has been suggested that rates of evolution in rDNA vary with the distance from functionally important parts of the ribosome, such as the transfer RNA path and the peptidyltransferase center (Smit et al. 2007). Although loops may engage in tertiary interactions (Dutheil et al. 2010) or junctions that link several helices together (Smit et al. 2006), they appear generally less constrained than helices. This is likely because loops are free to evolve by substitutions that do not change the secondary structure, hence may be neutral and fix by drift, whereas helices need much rarer compensatory mutations to maintain stability and high fitness (Gavrilets 2004; Meer et al. 2010).

On the other hand, helices were GC richer than loops (respectively 59.8% and 43.9% in vertebrates; 63.1% and 46.1% in angiosperms), whereas loops displayed a very high

Table 4. GC Content in 18S Ribosomal DNA in Major Lineages of Eukaryotes.

Lineage	*N*	Length (nt)	Quality (%)	Mean GC ± SD (range)
Alveolata	371	1,700	98	0.46 ± 0.013 (0.41–0.49)
Amoebozoa	45	1,200	95	0.52 ± 0.004 (0.50–0.52)
Choanoflagellida	5	1,600	85	0.46 ± 0.004 (0.45–0.47)
Cryptophyta	75	1,200	95	0.46 ± 0.008 (0.45–0.48)
Diplomonadina	6	1,300	80	0.71 ± 0.023 (0.68–0.75)
Euglenozoa	46	1,700	98	0.50 ± 0.003 (0.49–0.51)
Fungi	1,993	1,700	98	0.47 ± 0.013 (0.41–0.52)
Haptophyceae	117	1,700	95	0.49 ± 0.005 (0.48–0.50)
Heterokont	463	1,700	98	0.46 ± 0.015 (0.41–0.50)
Metazoa	5,211	1,700	95	0.50 ± 0.024 (0.40–0.60)
Rhodophyta	421	1,200	98	0.50 ± 0.011 (0.47–0.52)
Viridiplantae	2,506	1,700	98	0.49 ± 0.013 (0.44–0.54)
All eukaryotes	11,259	—	—	0.49 ± 0.025 (0.40–0.75)

NOTE.—*N*: number of sequences; length and quality refer to the thresholds used to filter short and low-quality sequences; nt: nucleotides; SD: standard deviation calculated across sequences.

content of adenine relative to uracil (respectively, 39.6% and 7.0% in vertebrates; 43.3% and 10.7% in angiosperms). This pattern of base composition in helix and loop regions seems to be universal: it has been observed in prokaryotes (Galtier and Lobry 1997; Wang and Hickey 2002; Wang et al. 2006) and eukaryotes (Wang et al. 2006; Smit et al. 2006, 2007). There is evidence that adenine contributes to the stability of the single-stranded regions of the RNA (Gutell et al. 2000). Also, because there are three hydrogen bonds in GC pairs and two in AT pairs, it has been suggested that GC pairs stabilize double-helix structures (Marmur and Doty 1959; Wada and Suyama 1986; Galtier and Lobry 1997).

Faster loop evolution seemed associated with enrichment in G and C bases in several clades, especially Polypteriformes and tetrapods among vertebrates, and Austrobaileyales among angiosperms (see the higher GC* relative to current GC in loop sites in figs. 2 and 3). This difference in the GC enrichment of the secondary structure, which was found highly significant by LRT, is likely due to their baseline difference in GC content: because loops are AT richer than helices, AT → GC substitutions are more likely to occur in the former than the latter. Nevertheless, GC enrichment in helix sites was observed in Polypteriformes, Holostei, Osteoglossiformes, mammals, birds, and Lepidosauria among vertebrates and Berberidopsidales among angiosperms (see the higher GC* relative to current GC in helix sites in figs. 2 and 3).

Mutation Bias or gBGC/Selection for GC?

rDNA sequences have one characteristic that makes them particular relative to most genes in the genome and that makes them especially interesting for this broad phylogenetic study: besides they are under strong purifying selective pressure (which guarantee that sequences are not saturated), the different paralogous copies of the rDNA undergo more frequent recombination events than most other genes in the genome thanks to concerted evolution (either through gene conversion, unequal crossing over or both). In addition, concerted evolution has been shown to enhance selection efficacy (Mano and Innan 2008). Concerted evolution should enhance gBGC too because it is formally equivalent to selection (Nagylaki 1983). All in all, the existence of multiple rDNA loci evolving in a concerted way presumably exacerbates the effects of recombination, including gBGC, as compared with a single-locus sequence.

We hypothesized that variations in the GC content of rDNA could serve as an indicator of the gBGC (or selection) process because this molecule is submitted to high recombination. Two aspects of our results are in agreement with this hypothesis. First, we find that the heterogeneity in base composition of rDNA coincide with the documented occurrence of gBGC: mammals, birds, and grasses, in which significant gBGC has been detected from genome-wide analyses (Glémin et al. 2006; Webster et al. 2006; Haudry et al. 2008; Duret and Galtier 2009), show distinctively high rDNA GC content and GC* (grasses make the majority of commelinid species in this data set). However, comprehensive analyses of base composition have only been made in a few species, and we lack information on, for example, sharks, ray-finned fishes, non-grass monocots, or eudicots. For this reason, results in Ostariophysi and commelinids cannot be compared with previous reports because these groups were overlooked in the past.

The second line of evidence comes from polymorphism data analyses. Allele frequency spectra were found significantly GC biased in 8 of 11 clades of vertebrates or angiosperms, and unbiased in three, for which limited amount of data was available. Hence, these data reveal a strong excess, among AT versus GC polymorphic sites, of SNPs for which G or C is the majority allele (table 2). Remarkably, this excess is detected in groups showing increasing (e.g., monocots, Siluriformes), stable (e.g., asterids, rosids), or decreasing (Chondrostei) rDNA GC content. We argue that such a pattern is consistent with the hypothesis of gBGC/selection-driven evolution of rDNA GC content and rejects the hypothesis of a GC-biased mutation process.

Because we analyzed nonoriented polymorphisms, the SNP frequency spectrum has a U form: SNPs in which AT is the majority allele are found at one extreme and at the other extreme are GC alleles. In the absence of

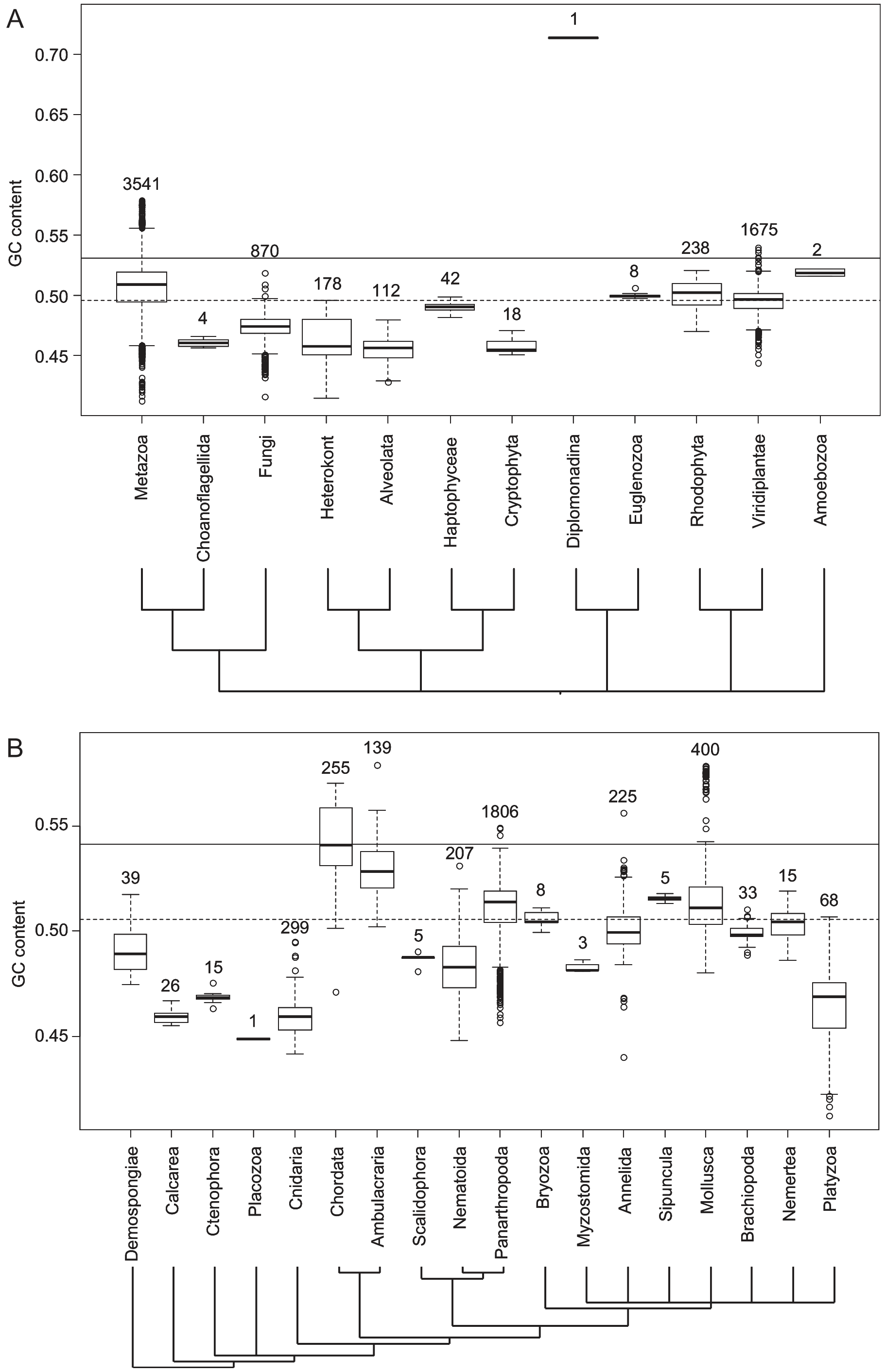

FIG. 4. Distribution of the GC content of 18S ribosomal DNA among major lineages of eukaryotes (*A*) and metazoans (*B*) for all sites of the molecule. Numbers above boxplots are the number of analyzed genera; horizontal lines represent the median across all genera (dotted) and the highest 95% percentile (continuous). Trees in the lower part of the figures represent the phylogeny of Eukaryotes, in *A* (Delsuc et al. 2005), or the animal phylogeny, in *B* (Halanych 2004).

any fixation bias (gBGC or selection), all classes of mutations have equal probability of reaching a given population frequency. Assuming that GC content is at equilibrium, symmetric AT versus GC allele-frequency patterns would be expected (i.e., AT and GC polymorphisms reach the same frequency). Observing a higher average allele frequency of GC alleles in species for which rDNA GC content is at equilibrium is therefore indicative of a fixation bias, such that G and C alleles have a higher probability to reach high frequencies than A and T alleles (Webster and Smith 2004; Galtier et al. 2006). SNP data, therefore, support the existence of a GC-biased fixation process (gBGC or selection) in groups showing equilibrium rDNA GC content (e.g., asterids, rosids).

The situation is more complex when GC content is not at equilibrium. In this case, asymmetric allele-frequency patterns are expected even under the neutral hypothesis because the numbers of AT → GC and GC → AT mutations arising at each generation differ from each other (Galtier et al. 2006). First, consider the case of increasing GC content (as in mammals, birds, and commelinids). If a GC-biased mutation process was the cause of GC increase, then the number of AT → GC mutations would be greater than the number of GC → AT mutations, and we would expect to observe higher frequencies of AT than GC SNPs. We found the opposite pattern, rejecting the mutation bias hypothesis as far as GC-increasing species are concerned. The observed pattern, however, would be expected if GC-increase was caused by an increased intensity of gBGC/selection. Now consider the case of GC-decreasing sequences, as in Chondrostei. In this case, the hypothesis of an AT-biased mutation process would predict a majority of GC → AT mutations and therefore of high GC-frequency SNPs.

Taken together, polymorphism data reject the hypothesis of mutation-driven GC content increase in rDNA. Rather, they support a model in which rDNA GC content is governed by two opposing forces, namely gBGC/selection, which tends to increase GC content, and AT-biased mutation pressure, which tends to decrease it. Depending on the relative importance of these forces, the GC content in rDNA will increase when gBGC (or selection) is switched on or strengthened (as in mammals, birds, or commelinids), decrease when gBGC (or selection) is switched off or weakened (as in Chondrostei), or remain stable if the two forces are unchanged for a long enough period of time.

Biased Gene Conversion or Selection for GC?

Our results on divergence and polymorphisms in rDNA are compatible with gBGC, but they do not rule out the possibility of selection for GC content, which is formally almost indistinguishable from gBGC. It has been postulated that because genomes are organized in isochores in mammals and birds but not in amphibians and fishes, high GC content could be an adaptation to homoeothermy (Bernardi et al. 1985; Bernardi 1993, 2000; Varriale et al. 2008). Our finding of high GC content in cold-blooded Ostariophysi (not to mention commelinid monocots) is incompatible with selective hypotheses based on thermal adaptation. It could still be possible that high GC was selected because it confers higher bend ability to the DNA molecule (Vinogradov 2003) or because it stabilizes RNA (Bernardi 2007). However, few data support these hypotheses, and they cannot explain why such selection should be higher in groups as diverse as Ostariophysi, birds, mammals, and commelinids. GC content could be selected because it stabilizes helices. However, this hypothesis does not explain why GC content in loops also increases in parallel. Although we cannot completely exclude the possibility that the patterns detected in this study are produced by some selective advantage, it seems more likely to us that they have been produced by a neutral mechanism, namely gBGC.

rDNA as a Marker of gBGC in Eukaryotes

By combining information of both interspecific divergence and intraspecific polymorphisms, we explored the evolutionary processes driving rDNA base composition in different lineages of eukaryotes, especially in vertebrates and angiosperms, and identified gBGC (or selection for GC) as the dominant GC-increasing force. This suggests that the GC content in 18S rDNA can be used as a reliable marker of gBGC to scan other groups of eukaryotes.

We analyzed GC content in 18S rDNA in some of the most representative kingdoms of eukaryotes. Besides the above discussed amniotes, Ostariophysi, and commelinids, we found that *Giardia* among Diplomonadina, Ulvophyceae among green algae, Doridina nudibranch gastropod mollusks, and Asterozoa echinoderms among Metazoa, exhibit the highest GC among all eukaryotes. GC content in *Giardia* is clearly out of the range of all other eukaryotes and might be explained by very specific mechanisms. The other groups suggest that gBGC could be a mechanism that shapes genome landscapes in more clades of eukaryotes than previously suspected. The sparse distribution of these groups also suggests that either gBGC has independently evolved many times or has recurrently intensified at different phylogenetic scales. Although general trends can hardly be drawn, the phylogenetic pattern in Metazoa is worth noting. Basal Metazoan groups exhibit very low GC content, whereas medium and strong enrichments seem to occur in protostomes and deuterostomes, respectively (fig. 4B).

Our results are based on one genetic marker, the rDNA. Yet, they can extrapolate to the rest of the genome if two assumptions are verified: 1) that the mechanism of gBGC is similar during ectopic and allelic recombination and 2) that gBGC affects translated and nontranslated DNA in a similar way. Although these two assumptions are worth verifying experimentally, there is no a priori reason to suspect they are invalid. Furthermore, the fact that we find similar patterns of GC enrichment in rDNA in groups in which gBGC has been documented using genome-wide analyses (e.g., mammals, birds, and grasses) suggests that rDNA constitutes a reliable marker of this molecular process at a broad phylogenetic scale. Based on these results, we hypothesize that gBGC might be active in more eukaryotic groups than previously thought, a hypothesis that requires to be confirmed or falsified using genome-wide analyses. Accumulation of genomic data thanks to high-throughput sequencing technologies would certainly allow this in the near future.

Supplementary Material

Supplementary table S1 and figures S1–S6 are available at *Molecular Biology and Evolution*online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank J. Dutheil and S. Blanquart for discussions and two anonymous reviewers for helpful comments that improved the quality of the manuscript. This work was supported by the Agence Nationale de la Recherche (CoGeBi program, grant ANR-08-GENM-036-01) and European Research Council (PopPhyl project, ERC 23291). This publication is the contribution 2011-032 of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554—CNRS).

References

Alfaro ME, Santini F, Brock C, Alamillo H, Dornburg A, Rabosky DL, Carnevale G, Harmon LJ. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc Natl Acad Sci U S A*. 106:13410–13414.

Arndt PF, Petrov DA, Hwa T. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol Biol Evol*. 20:1887–1896.

Arnheim N, Krystal M, Schmickel R, Wilson G, Ryder O, Zimmer E. 1980. Molecular evidence for genetic exchange among ribosomal genes on non-homologous chromosomes in man and apes. *Proc Natl Acad Sci U S A*. 77:7323–7327.

Arnheim N, Treco D, Taylor B, Eicher EM. 1982. Distribution of ribosomal gene length variants among mouse chromosomes. *Proc Natl Acad Sci U S A*. 79:4677–4680.

Averbeck KT, Eickbush TH. 2005. Monitoring the mode and tempo of concerted evolution in the *Drosophila melanogaster* rDNA locus. *Genetics* 171:1837–1846.

Bazin E, Duret L, Penel S, Galtier N. 2005. Polymorphix, a sequence polymorphism database. *Nucleic Acids Res*. 33:D481–D484.

Benevolenskaya EV, Kogan GL, Tulin AV, Philipp D, Gvozdev VA. 1997. Segmented gene conversion as a mechanism of correction of 18S rRNA pseudogene located outside of rDNA cluster in *D. melanogaster*. *J Mol Evol*. 44:646–651.

Bernardi G. 1993. The vertebrate genome: isochores and evolution. *Mol Biol Evol*. 10:186–204.

Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* 241:3–17.

Bernardi G. 2007. The neoselectionist theory of genome evolution. *Proc Natl Acad Sci U S A*. 104:8385–8390.

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J. 1985. The mosaic genome of warmblooded vertebrates. *Science* 228:953–958.

Bird AP. 1980. DNA methylation and the frequency of Cpg in animal DNA. *Nucleic Acids Res*. 8:1499–1504.

Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol*. 19:1181–1197.

Brinkmann H, Venkatesh B, Brenner S, Meyer A. 2004. Nuclear protein-coding genes support lungfish and not the coelacanth as the closest living relatives of land vertebrates. *Proc Natl Acad Sci U S A*. 101:4900–4905.

Brown DD, Wensink PC, Jordan E. 1972. A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. *J Mol Biol*. 63:57–73.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 17:540–552.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*. 6:361–375.

Douzery EJP, Huchon D. 2004. Rabbits, if anything, are likely Glires. *Mol Phylogenet Evol*. 33:922–935.

Dover G. 1994. Concerted evolution, molecular drive and natural selection. *Curr Biol*. 4:1165–1166.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*. 4:e1000071.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*. 10:285–311.

Duret L, Semon M, Piganeau G, Mouchiroud D, Galtier N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* 162:1837–1847.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol*. 8:255.

Dutheil JY, Jossinet F, Westhof E. 2010. Base pairing constraints drive structural epistasis in ribosomal RNA sequences. *Mol Biol Evol*. 27:1868–1876.

Eickbush TH, Eickbush DG. 2007. Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* 175:477–485.

Elder JF, Turner BJ. 1995. Concerted evolution of repetitive DNA sequences in Eukaryotes. *Q Rev Biol*. 70:297–320.

Endow SA, Komma DJ. 1986. One-step and stepwise magnification of a bobbed lethal chromosome in *Drosophila melanogaster*. *Genetics* 114:511–523.

Escobar JS, Cenci A, Bolognini J, Haudry A, Laurent S, David J, Glémin S. 2010. An integrative tests of the dead-end hypothesis of selfing evolution in Triticeae (Poaceae). *Evolution* 64:2855–2872.

Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat Rev Genet*. 2:549–555.

Flavell RB, O'Dell M. 1976. Ribosomal RNA genes in homeologous chromosomes of groups 5 and 6 in hexaploid wheat. *Heredity* 37:377–385.

Franzke A, Mummenhoff K. 1999. Recent hybrid speciation in *Cardamine* (Brassicacea)-conversion of nuclear ribosomal ITS sequences in statu nascendi. *Theor Appl Genet*. 98:831–834.

Fuertes Aguilar J, Roselló JA, Nieto Feliner G. 1999. Nuclear ribosomal DNA (nrDNA) concerted evolution in natural and artificial hybrids of *Armeria* (Plumbaginaceae). *Mol Ecol*. 8:1341–1346.

Galtier N. 2003. Gene conversion drives GC content evolution in mammalian histones. *Trends Genet*. 19:65–68.

Galtier N, Bazin E, Bierne N. 2006. GC-biased segregation of noncoding polymorphisms in *Drosophila*. *Genetics*. 172:221–228.

Galtier N, Duret L, Glemin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet*. 25:1–5.

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.

Galtier N, Lobry JR. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol*. 44:632–636.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159:907–911.

Gangloff S, Zou H, Rothstein R. 1996. Gene conversion plays the major role in controlling the stability of large tandem repeats in yeast. *EMBO J*. 15:1715–1725.

Gavrilets S. 2004. Fitness landscapes and the origin of species. Princeton (NJ): Princeton University Press.

Glémin S, Bazin E, Charlesworth D. 2006. Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proc R Soc B Biol Sci*. 273:3011–3019.

González F, Betancur J, Maurin O, Freudenstein JV, Chase MW. 2007. Metteniusaceae, an early-diverging family in the lamiid clade. *Taxon*. 56:795–800.

Gonzalez IL, Sylvester JE. 2001. Human rDNA: evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes. *Genomics* 73:255–263.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.

Gutell RR, Cannone JJ, Shang Z, Du Y, Serra MJ. 2000. A story: unpaired adenosine bases in ribosomal RNAs. *J Mol Biol.* 304: 335–354.

Haddrill PR, Charlesworth B. 2008. Non-neutral processes drive the nucleotide composition of non-coding sequences in *Drosophila*. *Biol Lett.* 4:438–441.

Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. 2007. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biol.* 8:R18.

Halanych KM. 2004. The new view of animal phylogeny. *Annu Rev Ecol Evol Syst.* 35:229–256.

Hasegawa M, Hashimoto T. 1993. Ribosomal RNA trees misleading? *Nature* 362:795.

Haudry A, Cenci A, Guilhaumon C, Paux E, Poirier S, Santoni S, David J, Glémin S. 2008. Mating system and recombination affect molecular evolution in four Triticeae species. *Genet Res.* 90:97–109.

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 6:e1001115.

Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* 6:e1001107.

Hillis DM, Moritz C, Porter CA, Baker RJ. 1991. Evidence for biased gene conversion in concerted evolution of ribosomal DNA. *Science* 251:308–310.

Hudelot C, Gowri-Shankar V, Jow H, Rattray M, Higgs PG. 2003. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. *Mol Phylogenet Evol.* 28:241–252.

Hugall AF, Foster R, Lee MSY. 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. *Syst Biol.* 56:543–563.

Inoue JG, Miya M, Tsukamoto K, Nishida M. 2004. Mitogenomic evidence for the monophyly of elopomorph fishes (Teleostei) and the evolutionary origin of the leptocephalus larva. *Mol Phylogenet Evol.* 32:274–286.

Karehed J. 2001. Multiple origin of the tropical forest tree family Icacinaceae. *Am J Bot.* 88:2259–2274.

Koch MA, Dobes C, Mitchell-Olds T. 2003. Multiple hybrid formation in natural populations: concerted evolution of the internal transcribed spacer of nuclear ribosomal DNA (ITS) in North American *Arabis divricarpa* (Brassicaceae). *Mol Biol Evol.* 20:338–350.

Kovarik A, Matyasek R, Leitch A, Gazdova B, Fulnecek J, Bezdek M. 1997. Variability in CpNpG methylation in higher plant genomes. *Gene* 204:25–33.

Kovarik A, Matyasek R, Lim KY, Skalicka K, Koukalova B, Knapp S, Chase M, Leitch AR. 2004. Concerted evolution of 18-5.8-26S rDNA repeats in *Nicotiana* allotetraploids. *Biol J Linn Soc.* 82:615–625.

Kovarik A, Pires JC, Leitch AR, Lim KY, Sherwood AM, Matyasek R, Rocca J, Soltis DE, Soltis PS. 2005. Rapid concerted evolution of nuclear ribosomal DNA in two tragopogon allopolyploids of recent and recurrent origin. *Genetics* 169:931–944.

Kudla G, Helwak A, Lipinski L. 2004. Gene conversion and GC-content evolution in mammalian Hsp70. *Mol Biol Evol.* 21:1438–1444.

Kupriyanova NS. 2000. Conservation and variation of ribosomal DNA in eukaryotes. *Mol Biol.* 34:637–647.

Lavoué S, Miya M, Inoue JG, Saitoh K, Ishiguro NB, Nishida M. 2005. Molecular systematics of the gonorynchiform fishes (Teleostei) based on whole mitogenome sequences: implications for higher-level relationships within the Otocephala. *Mol Phylogenet Evol.* 37:165–177.

Liao DQ. 1999. Concerted evolution: molecular mechanism and biological implications. *Am J Hum Genet.* 64:24–30.

Liao DQ. 2000. Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. *J Mol Evol.* 51:305–317.

Lim KY, Kovarik A, Matyasek R, Bezdek M, Lichtenstein CP, Leitch AR. 2000. Gene conversion of ribosomal DNA in *Nicotiana tabacum* is associated with undermethylated, decondensed and probably active gene units. *Chromosoma* 109:161–172.

Long EO, Dawid IB. 1980. Repeated genes in eukaryotes. *Annu Rev Biochem.* 49:727–764.

Loytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A.* 102:10557–10562.

Ludwig W, Strunk O, Westram R, et al. (32 co-authors). 2004. ARB: a software environment for sequence data. *Nucleic Acids Res.* 32:1363–1371.

Lundberg J. 2001. Phylogenetic studies in the Euasterids II with particular reference to Asterales and Escalloniaceae. Uppsala (Sweden): Acta Universitatis Upsaliensis.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates Inc.

Mallatt J, Winchell CJ. 2007. Ribosomal RNA genes and deuterostome phylogeny revisited: more cyclostomes, elasmobranchs, reptiles, and a brittle star. *Mol Phylogenet Evol.* 43:1005–1022.

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454:479–485.

Mano S, Innan H. 2008. The evolutionary rate of duplicated genes under concerted evolution. *Genetics* 180:493–505.

Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* 19:330–338.

Marmur J, Doty P. 1959. Heterogeneity in deoxyribonucleic acids. I. Dependence on composition of the configurational stability of deoxyribonucleic acids. *Nature* 183:1427–1429.

Meer MV, Kondrashov AS, Artzy-Randrup Y, Kondrashov FA. 2010. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature* 464:279–282.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol.* 21:984–990.

Miya M, Takeshima H, Endo H, et al. (12 co-authors). 2003. Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. *Mol Phylogenet Evol.* 26:121–138.

Montoya-Burgos JI, Boursot P, Galtier N. 2003. Recombination explains isochores in mammalian genomes. *Trends Genet.* 19:128–130.

Murata Y, Nikaido M, Sasaki T, Cao Y, Fukumoto Y, Hasegawa M, Okada N. 2003. Afrotherian phylogeny as inferred from complete mitochondrial genomes. *Mol Phylogenet Evol.* 28:253–260.

Nagylaki T. 1983. Evolution of a finite population under gene conversion. *Proc Natl Acad Sci U S A.* 80:6278–6281.

Nickrent DL, Der JP, Anderson FE. 2005. Discovery of the photosynthetic relatives of the "Maltese mushroom" *Cynomorium*. *BMC Evol Biol.* 5:38.

Olmstead RG, Kim KJ, Jansen RK, Wagstaff SJ. 2000. The phylogeny of the Asteridae sensu lato based on chloroplast *ndhF* gene sequences. *Mol Phylogenet Evol.* 16:96–112.

Petes TD. 1980. Unequal meiotic recombination within tandem arrays of yeast ribosomal DNA genes. *Cell* 19:765–774.

Pruesse E, Quast C, Knittel K, Fuchs B, Ludwig W, Peplies J, Glöckner FO. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35:7188–7196.

R Development Core Team. 2009. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Rauscher JT, Doyle JJ, Brown AH. 2004. Multiple origins and nrDNA internal transcribed spacer homeologue evolution in the *Glycine tomentella* (Leguminosae) allopolyploid complex. *Genetics* 166:987–998.

Rogers SO, Bendich AJ. 1987. Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. *Plant Mol Biol.* 9:509–520.

Romiguier J, Ranwez V, Douzery EJP, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res.* 20:1001–1009.

Schlotterer C, Tautz D. 1994. Chromosomal homogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. *Curr Biol.* 4:777–783.

Smit S, Widmann J, Knight R. 2007. Evolutionary rates vary among rRNA structural elements. *Nucleic Acids Res.* 35:3339–3354.

Smit S, Yarus MY, Knight R. 2006. Natural selection is not required to explain universal compositional patterns in rRNA secondary structure categories. *RNA—a Publication of the RNA Society.* 12:1–14.

Soltis DE, Soltis PS, Chase MW, et al. (16 co-authors). 2000. Angiosperm. phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. *Bot J Linn Soc.* 133:381–461.

Soltis DE, Soltis PS, Nickrent DL, et al. (16 co-authors). 1997. Angiosperm. phylogeny inferred from 18S ribosomal DNA sequences. *Ann Mo Bot Gard.* 84:1–49.

Soltis PS, Soltis DE. 1998. Molecular evolution of 18S rDNA in Angiosperms: implications for character weighting in phylogenetic analysis. In: Soltis DE, Soltis PS, Doyle JJ, editors. Molecular systematics of plants II. Norwell (MA): Kluwer Academic Publishers. p. 188–210.

Spencer CCA. 2006. Human polymorphism around recombination hotspots. *Biochem Soc Trans.* 34:535–536.

Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci U S A.* 48:582–592.

Sullivan JP, Lundberg JG, Hardman M. 2006. A phylogenetic analysis of the major groups of catfishes (Teleostei: Siluriformes) using *rag1* and *rag2* nuclear gene sequences. *Mol Phylogenet Evol.* 41:636–662.

Swalla BJ, Smith AB. 2008. Deciphering deuterostome phylogeny: molecular, morphological and palaeontological perspectives. *Philos Trans R Soc B Biol Sci.* 363:1557–1568.

Szostak JW, Wu R. 1980. Unequal crossing over in the ribosomal DNA of *Saccharomyces cerevisiae*. *Nature* 284:426–430.

Tamura MN, Yamashita J, Fuse S, Haraguchi M. 2004. Molecular phylogeny of monocotyledons inferred from combined analysis of plastid *matK* and *rbcL* gene sequences. *J Plant Res.* 117:109–120.

The Angiosperm Phylogeny Group. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc.* 161:105–121.

Townsend TM, Larson A, Louis E, Macey JR. 2004. Molecular phylogenetics of Squamata: the position of snakes, amphisbaenians, and dibamids, and the root of the squamate tree. *Syst Biol.* 53:735–757.

van Tuinen M, Sibley CG, Hedges SB. 2000. The early history of modern birds inferred from DNA sequences of nuclear and mitochondrial ribosomal genes. *Mol Biol Evol.* 17:451–457.

Varriale A, Torelli G, Bernardi G. 2008. Compositional properties and thermal adaptation of 18S rRNA in vertebrates. *RNA—a Publication of the Rna Society.* 14:1492–1500.

Venkatesh B, Erdmann MV, Brenner S. 2001. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. *Proc Natl Acad Sci U S A.* 98:11382–11387.

Vinogradov AE. 2003. DNA helix: the importance of being GC-rich. *Nucleic Acids Res.* 31:1838–1844.

Wada A, Suyama A. 1986. Local stability of DNA and RNA secondary structure and its relation to biological functions. *Prog Biophys Mol Biol.* 47:113–157.

Walsh JB. 1985. Interaction of selection and biased gene conversion in a multigene family. *Proc Natl Acad Sci U S A.* 82:153–157.

Wang HC, Hickey DA. 2002. Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes. *Nucleic Acids Res.* 30:2501–2507.

Wang HC, Singer GAC, Hickey DA. 2004. Mutational bias affects protein evolution in flowering plants. *Mol Biol Evol.* 21:90–96.

Wang HC, Xia XH, Hickey D. 2006. Thermal adaptation of the small subunit ribosomal RNA gene: a comparative study. *J Mol Evol.* 63:120–126.

Wanga H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, Davis CC, Latvis M, Manchester SR, Soltis DE. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc Natl Acad Sci U S A.* 106:3853–3858.

Webster MT, Axelsson E, Ellegren H. 2006. Strong regional biases in nucleotide substitution in the chicken genome. *Mol Biol Evol.* 23:1203–1216.

Webster MT, Smith NG. 2004. Fixation biases affecting human SNPs. *Trends Genet.* 20:122–126.

Wendel JF, Schnabel A, Seelanan T. 1995. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proc Natl Acad Sci U S A.* 92:280–284.

Winchell CJ, Sullivan J, Cameron CB, Swalla BJ, Mallatt J. 2002. Evaluating hypotheses of deuterostome phylogeny and chordate evolution with new LSU and SSU ribosomal DNA data. *Mol Biol Evol.* 19:762–776.

Worberg A, Alford MH, Quandt D, Borsch T. 2009. Huerteales sister to Brassicales plus Malvales, and newly circumscribed to include *Dipentodon*, *Gerrardina*, *Huertea*, *Perrottetia*, and *Tapiscia*. *Taxon.* 58:468–478.

Xia X, Xie Z, Kjer KM. 2003. 18S ribosomal RNA and tetrapod phylogeny. *Syst Biol.* 52:283–295.