Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placentalmammals

Jonathan Romiguier, Vincent Ranwez, Frédéric Delsuc, Nicolas Galtier, Emmanuel J.P. Douzery

Less Is More in Mammalian Phylogenomics: AT-Rich Genes Minimize Tree Conflicts and Unravel the Root of Placental Mammals

Jonathan Romiguier,*[,1] Vincent Ranwez,[2] Frédéric Delsuc,[1] Nicolas Galtier,[1] and Emmanuel J.P. Douzery[1]

[1]CNRS, Université Montpellier, Institut des Sciences de l'Evolution, Montpellier, France

[2]Montpellier SupAgro, AGAP, Montpellier, France

***Corresponding author:** E-mail: jonathan.romiguier@gmail.com.

Associate Editor: Emma Teeling

Abstract

Despite the rapid increase of size in phylogenomic data sets, a number of important nodes on animal phylogeny are still unresolved. Among these, the rooting of the placental mammal tree is still a controversial issue. One difficulty lies in the pervasive phylogenetic conflicts among genes, with each one telling its own story, which may be reliable or not. Here, we identified a simple criterion, that is, the GC content, which substantially helps in determining which gene trees best reflect the species tree. We assessed the ability of 13,111 coding sequence alignments to correctly reconstruct the placental phylogeny. We found that GC-rich genes induced a higher amount of conflict among gene trees and performed worse than AT-rich genes in retrieving well-supported, consensual nodes on the placental tree. We interpret this GC effect mainly as a consequence of genome-wide variations in recombination rate. Indeed, recombination is known to drive GC-content evolution through GC-biased gene conversion and might be problematic for phylogenetic reconstruction, for instance, in an incomplete lineage sorting context. When we focused on the AT-richest fraction of the data set, the resolution level of the placental phylogeny was greatly increased, and a strong support was obtained in favor of an Afrotheria rooting, that is, Afrotheria as the sister group of all other placentals. We show that in mammals most conflicts among gene trees, which have so far hampered the resolution of the placental tree, are concentrated in the GC-rich regions of the genome. We argue that the GC content—because it is a reliable indicator of the long-term recombination rate—is an informative criterion that could help in identifying the most reliable molecular markers for species tree inference.

Key words: **phylogenomics, placental mammal, GC-content, incomplete lineage sorting, biased gene conversion, Afrotheria.**

Article

Introduction

Most evolutionary biology studies rely on a well-resolved phylogenetic tree. Over the last decade, this requirement has been mainly achieved through the use of molecular data. Initially, limited to a few genes, molecular data sets are now growing considerably thanks to high-throughput sequencing. With whole-genome data sets, resolving a species phylogeny is no longer a matter of quantity of unambiguously aligned sites. However, despite the rise of phylogenomics (Eisen and Fraser 2003; Delsuc et al. 2005), important nodes remain unresolved.

One of the most iconic examples is the phylogeny of placental mammals, which was among the first to be studied in the light of molecular data. Challenging the classical morphological classification, multigene studies suggested three major clades of placental mammals (Madsen et al. 2001; Murphy et al. 2001): Afrotheria (e.g., elephants and tenrecs), Xenarthra (e.g., armadillos and sloths), and Boreoeutheria, further divided in two groups, that is, Euarchontoglires (e.g., primates and rodents) and Laurasiatheria (e.g., ruminants, cetaceans, bats, and carnivores). These clades are now well established and have been confirmed by several subsequent studies (Delsuc et al. 2002; Scally et al. 2002; Prasad et al. 2008; Meredith et al. 2011). Nevertheless, other nodes are trickier to unravel.

The root of placental mammals has been one of the most controversial and difficult nodes to resolve. Several contradictory studies have led to competing hypotheses, that is, 1) the Xenarthra rooting (Afrotheria + Boreoeutheria) (Waddell et al. 2001; Kriegs et al. 2006; O'Leary et al. 2013), 2) the Afrotheria rooting (Xenarthra + Boreoeutheria) (Murphy et al. 2001; Waddell and Shelley 2003; Nikolaev et al. 2007; Nishihara et al. 2007; Meredith et al. 2011; McCormack et al. 2012), and 3) the Atlantogenata rooting (Afrotheria + Xenarthra) (Huchon et al. 2002; Murphy et al. 2007; Wildman et al. 2007; Hallström et al. 2007; Kjer and Honeycutt 2007; Hallström and Janke 2008; Prasad et al. 2008; Meredith et al. 2011; Song et al. 2012). However, recent studies based on retroelement insertions suggest that the placental root might be impossible to resolve (Churakov et al. 2009; Nishihara et al. 2009). Other remaining uncertainties include relationships within Laurasiatheria, the position of Scandentia (tree shrews), and the relations among

the three major clades of rodents (Meredith et al. 2011). These discrepancies were first believed to be resolvable with the advent of whole-genome sequencing. Nevertheless, despite the availability of 39 mammalian genomes (Ensembl release 67; Birney et al. 2004), the uncertainties still persist.

Actually, a growing corpus of studies has revealed conflicting evolutionary histories among genes (Degnan and Rosenberg 2006, 2009). According to the literature, these discordances among gene trees could mainly be due to coalescent stochasticity (Pamilo and Nei 1988; Degnan and Rosenberg 2009; Hobolth et al. 2011), when incomplete sorting of the ancestral polymorphism occurs during successive speciation events, leading to gene genealogies that differ from the species phylogeny. This phenomenon called incomplete lineage sorting (ILS) has been detected in several different taxa (Jennings and Edwards 2005; Pollard et al. 2006; Matzke et al. 2012), including hominids (Patterson et al. 2006; Hobolth et al. 2011). ILS has mainly been reported to occur during rapid bursts of speciation, a phenomenon encountered in mammals for the basal divergence of placentals and the explosive radiation of Laurasiatheria orders for instance (Springer et al. 2003; Meredith et al. 2011).

For these reasons, dealing with ILS seems to be one of the keys to unravel the phylogeny of placental mammals. To this end, several coalescent-based methods have been developed (Liu et al. 2009, 2010). Recent applications of these approaches to placentals have led to conflicting results with respect to the root position. Flanking regions of ultraconserved elements strongly support the Afrotheria rooting (McCormack et al. 2012), whereas coding sequences strongly support the alternative Atlantogenata rooting (Song et al. 2012). This discrepancy suggests that, irrespective of tree-building methods, the very nature of the data is of primary concern (e.g., coding vs. noncoding). Here, we suggest an alternative approach. Instead of trying to explicitly account for ILS, we propose to minimize its impact by focusing on a subset of the data less likely to be affected by this phenomenon. Identifying indicators of the ILS probability of a gene genealogy is therefore required.

An in-depth analysis of hominid genomes reported that ILS is associated with an increased recombination rate (Hobolth et al. 2011). An explanation for this effect is that recombination tends to increase local levels of polymorphism by alleviating the effect of background selection (Charlesworth et al. 1993), thus increasing the probability of ILS. Beyond ILS, recombination is a common issue in phylogeny reconstruction as it can mix genes in neighboring segments with different histories (Posada and Crandall 2002; Ruths and Nakhleh 2005; Degnan and Rosenberg 2006). For all these reasons, genes with low recombination rates are predicted to be better markers of the species phylogeny (Hobolth et al. 2011; Escobar et al. 2011). However, local recombination rates are only available in a handful of species and are known to have changed frequently throughout the evolutionary history of placentals (Ptak et al. 2005). Fortunately, the recombination rate is positively associated with the GC content in mammals (Duret and Arndt 2008). This effect is due to a neutral mechanism, that is, biased gene conversion, a DNA repair bias toward GC acting on heteroduplexes induced by meiotic recombination (Galtier et al. 2001; Marais 2003; Duret 2009; Romiguier et al. 2010). Consequently, GC content correlates even more strongly with ILS than the recombination rate estimates in hominids (Hobolth et al. 2011). Indeed, GC content is probably a better predictor of long-term recombination than current recombination rate estimates (Lartillot 2013).

Orthologous coding sequences are the largest and most commonly used resource available for mammalian phylogenomics (Ranwez et al. 2007). Their GC content is readily accessible and could be a proxy for the probability of a gene to experiment ILS. To test this hypothesis, we analyzed the accuracy of the phylogeny of a gene with regard to its GC content in mammals. We found that AT-rich genes are more congruent markers of the species phylogeny than GC-rich genes, on average, and we focused on AT-rich genes to further investigate the difficult nodes of the placental tree.

Results

Higher Probability of Topology Incongruence in GC-Rich Gene Trees

We compared 13,111 mammal gene trees with an assumed true species tree backbone (controversial nodes in literature kept unresolved, see topology in fig. 1 and Materials and Methods). In most of the following results, 13,111 genes were ranked according to their GC3 (G + C content of third codon position) and then divided in 131 subsets of 100 genes. Figure 2*A* shows the main evidence that the GC-richer gene markers are more likely to be inaccurate for inferring mammalian phylogeny. Indeed, GC3 was positively correlated with the average tree error. Regardless of the metrics used to measure the distance between topologies, this correlation remained highly significant: Robinson–Foulds distance (P value < 0.0001, $r^2 = 0.70$), quartet similarity (P value < 0.0001, $r^2 = 0.77$), or triplet similarity (shown in fig. 2*A*, gray circles, P value < 0.0001, $r^2 = 0.85$). The error rate for GC-rich genes was therefore up to 5-fold higher than that of most AT-rich genes.

GC and AT cumulative subsets (red and blue triangles) illustrated the same trend. Adding an increasing number of GC-rich gene trees nearly always increased the level of discrepancy among them (red triangles). In contrast, adding more AT-rich gene trees always improved the global topological agreement (blue triangles). The whole data set (right side of the plot) could thus be curated by the removal of GC-rich genes, whereas removing AT-rich genes increases the amount of gene tree heterogeneity.

This fact alone suggests that using GC-rich markers should be avoided in phylogenomic studies. Nevertheless, the reason underlying this strong effect is unclear. Obviously, alignments with fewer characters (fewer sites and/or species) generally produce a less accurate phylogeny, as they are more prone to both stochastic and systematic errors (Phillips et al. 2004). In agreement with this hypothesis, we found a negative correlation between the GC content of an alignment and its number of sites (P value < 0.0001, $r^2 = 0.03$) or species

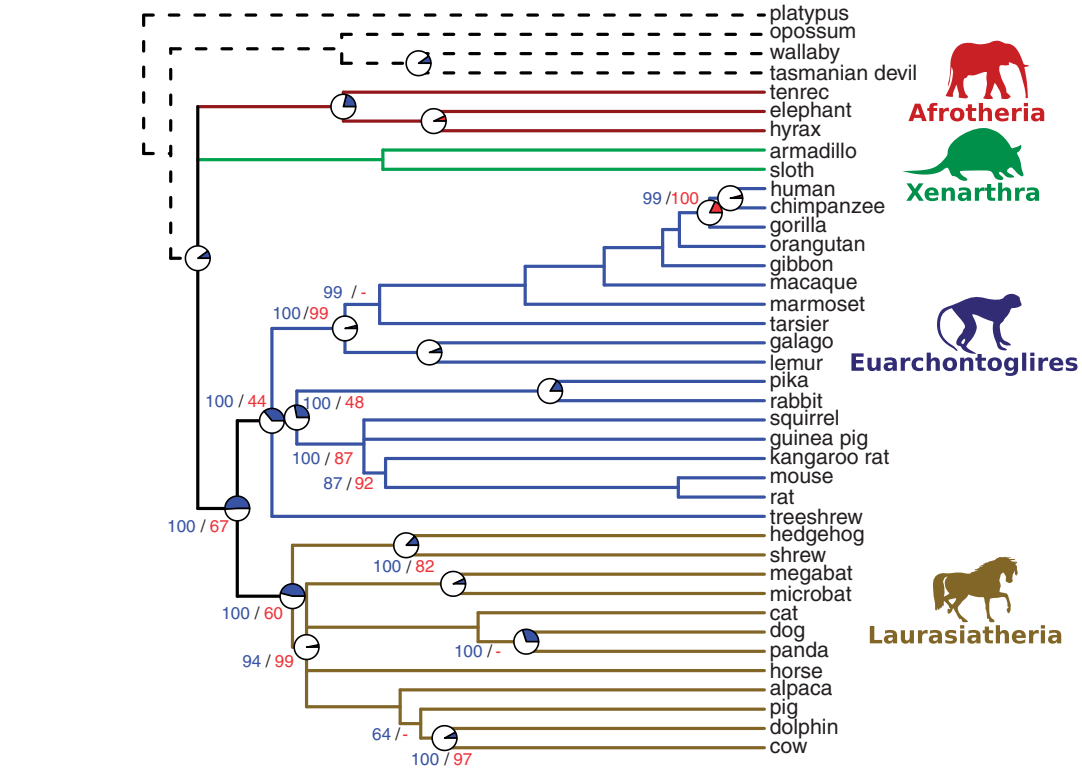

Fig. 1. GC effect on the reference tree topology. This reference tree topology leaves the most debated nodes of the mammalian phylogenetic tree unresolved (see Reference Tree section in the Materials and Methods). Branch lengths are proportional to the divergence dates (except for distant nonplacental species with dotted branches) as found in the TimeTree database (Hedges et al. 2006). Pie charts refer to the "Supertree approach" section and show the explained variance in node support by GC3 (blue for a negative effect of GC3, red for a positive one). We show only significant effects for the gap-homogenized data set (the same alignment gap quantity among GC-rich and AT-rich genes). Blue and red numbers refer to the "Supermatrix approach" analysis and show bootstrap supports for concatenation of the top 100 AT-rich genes (in blue) versus the concatenation of the top 100 GC-rich genes (in red). Equal support values are not shown. A dash indicates that the node was not retrieved.

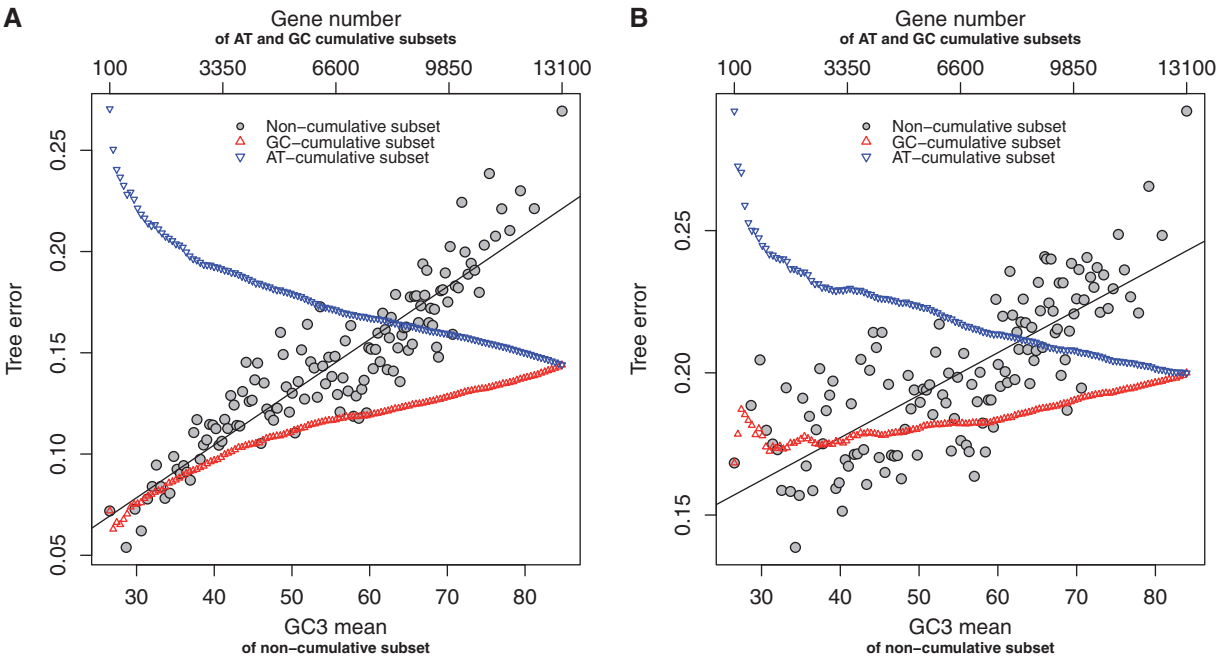

Fig. 2. GC effect on gene tree error. Each gray circle represents 100 alignments (for a total of 131), positioned according to their GC3 mean (*X* axis) and their average tree errors (*Y* axis). A gene tree error is here the proportion of false species triplets of a tree compared with the true species genealogy (reference tree topology of fig. 1). Triangles stand for cumulative subsets, where *X* axis is the number of genes of the subset. GC cumulative subsets contain an increasing number of GC-rich genes, AT cumulative subsets, an increasing number of AT-rich genes. (*A*) For the raw data set without any control. (*B*) For a modified data set where alignment gappiness (site and taxon number) is homogenized along the GC-gradient.

(P value < 0.0001, $r^2 = 0.17$). This recalls GC biases reported in studies using polymerase chain reaction (PCR)-based genome sequencing methods (Aird et al. 2011; Benjamini and Speed 2012; Dabney and Meyer 2012). It therefore appears that GC-rich genes yield less informative alignments in currently available databases. To check that the elevated error rate in GC-rich gene trees might be due to missing characters, we repeated our analyses after correcting for the number of sites and species along the GC gradient of our 13,111 alignments (see Materials and Methods). The results confirm the effect of the GC3 without any putative bias due to the character quantity (fig. 2*B*). The correlation is weaker but remains highly significant for all metrics: Robinson–Foulds distance (P value < 0.0001, $r^2 = 0.33$), quartet similarity (P value < 0.0001, $r^2 = 0.49$), and triplet similarity (fig. 2*B*, gray circles, P value < 0.0001, $r^2 = 0.59$). Independently of missing character and taxa, GC-rich genes clearly produce an excess of anomalous gene trees.

GC-Rich Genes Tend to Miss the Monophyly of the Deepest Placental Nodes

Supertree Approach

To characterize errors induced by the GC content, we computed one supertree per noncumulative subset of 100 gene trees, extracted the support values for each well-established clade of the reference tree (fig. 1), and correlated them to the average gene GC3.

The pie charts on nodes in figure 1 show the percentage of their variance in support values explained by GC3 (i.e., r^2). Blue pies represent a negative effect of GC3 on support values, and red pies indicate a positive effect. To be as conservative as possible, we only considered significant correlations in the gap-homogenized data set (see Materials and Methods).

We observed that the support value of most nodes was negatively correlated with GC3. This result is in agreement with figure 2, where GC-rich genes tended to produce topological errors, thus decreasing the support values of consensual nodes. Expectedly, this effect was more striking with raw data, which also encompassed the effect of missing data on topological inference (not shown).

Interestingly, the three nodes most affected by the GC effect were deep in the placental tree. Their support values were all negatively correlated with GC3 (P value < 0.0001): Euarchontoglires ($r^2 = 0.77$ for raw data, $r^2 = 0.36$ after gap homogenization), Laurasiatheria ($r^2 = 0.79$ for raw data, $r^2 = 0.46$ after gap homogenization), and Boreoeutheria ($r^2 = 0.82$ for raw data, $r^2 = 0.51$ after gap homogenization). Other nodes negatively affected by GC content were Caniformia (dog + panda) and Afrotheria.

Supermatrix Approach

We concatenated the top 100 GC-richest and the top 100 AT-richest genes in two supermatrices. For a fair comparison, we used alignments obtained after gap homogenization. The two supermatrices thus contained 81,924 sites, with the same number of species and the same proportion of missing data. We used each of these two supermatrices to infer a phylogeny of mammals using maximum likelihood as implemented in RAxML (Stamatakis 2006b) and compared the support values in figure 1 (blue numbers for the AT-rich supermatrix, red numbers for the GC-rich supermatrix).

These results are in close agreement with the supertree approach. Once again, the deepest nodes receive stronger support from AT-rich genes. In the GC-rich supermatrix tree, Boreoeutheria, Euarchontoglires, and Laurasiatheria support values, respectively, dropped from 100 to 67, 100 to 44, and 100 to 60. Some other support values also decreased (e.g., Glires dropped from 100 to 48), and some well-established clades were not even recovered: dog + panda, pig + cow + dolphin, and tarsier + anthropoids.

AT-Rich Genes Support the Afrotheria Rooting Hypothesis of the Placental Tree

In the previous section, AT content and increased support for deep nodes were clearly linked. However, would this apply for the most controversial one, that is, the root of Placentalia? To answer this question, we concatenated the alignments used in figure 2*A* into 131,100-gene supermatrices and performed a maximum likelihood analysis on each of them. The corresponding bootstrap supports for the three alternative hypotheses on the placental root are displayed in figure 3, according to the average GC3 and the average tree quality of the 100 genes. The resulting trees differed markedly, most of them supporting either the Atlantogenata or the Afrotheria rooting hypotheses. Interestingly, the support values were nonrandomly distributed. Producing better trees, AT-rich genes (left part of fig. 3, 77% of the first quarter) mainly supported an Afrotheria root, whereas GC-rich genes (right part of fig. 3) were in favor of Atlantogenata. To gain insight into this trend, we computed the root-to-tip branch lengths for each Afrotheria and Xenarthra species. The resulting mean for each ML tree is reported on the top of figure 3. Notably, GC-rich genes produced trees with longer branches, not only for these taxa (P value < 0.0001, $r^2 = 0.59$) but also for the whole tree (P value < 0.0001, $r^2 = 0.66$). Associated with the increased topological error, this result is strongly suggestive of a long-branch attraction artifact (Brinkmann et al. 2005). Indeed, Afrotheria and Xenarthra are poorly sampled clades, leading to long ancestral branches that cannot be further divided. These long branches could thus be prone to attract each other, a risk probably increased in fast-evolving GC-rich genes. These results suggest that Afrotheria might be the true rooting of the placental tree, whereas the Atlantogenata rooting would be a phylogenetic reconstruction artifact due to long-branch attraction of two fast-evolving ingroup lineages.

Exclusion of Unreliable GC-Rich Genes

Avoiding GC-rich genes thus seems to be highly relevant for resolving the placental root. Would a curated data set without such recombining genes be able to resolve the whole tree? As a first curating attempt, we excluded alignments with a GC3% above the average whole-genome rate. In mammals, this average rate is roughly equal to 40% (e.g., 40.91 for human), with extremes ranging from 37.82% (opossum) to 45.49% (platypus) (Kryukov et al. 2012). Moreover, because

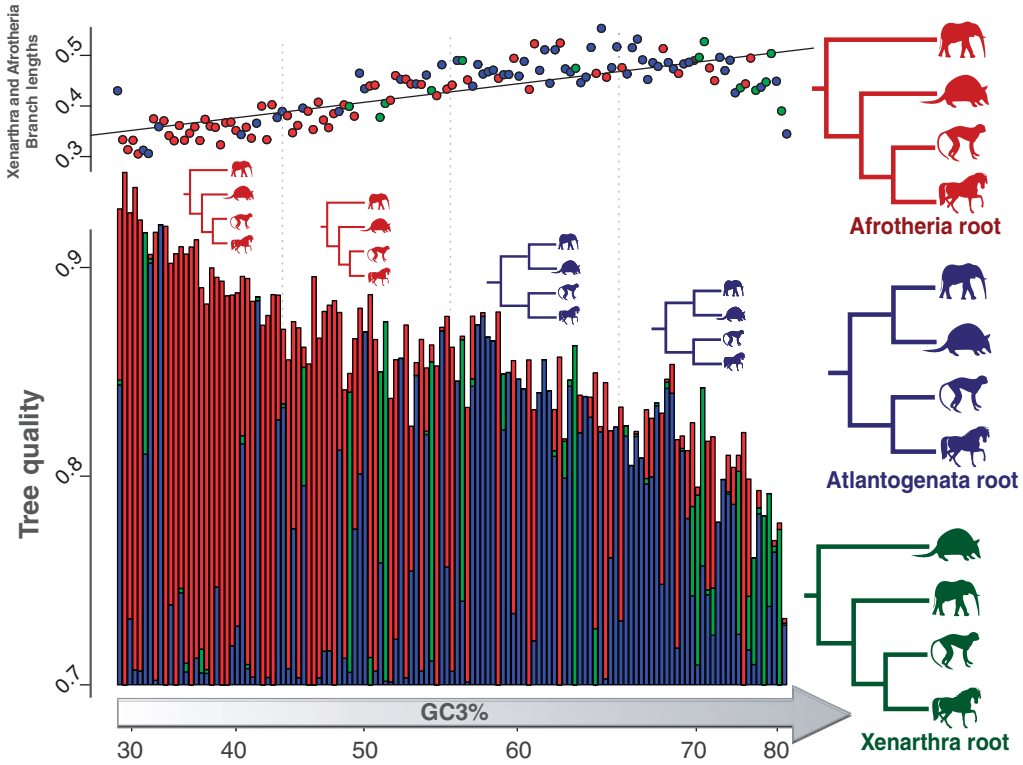

Fig. 3. Support values for the root of Placentalia according to GC content and tree quality. Each bar represents a concatenated set of 100 genes grouped according to their GC3-content (as for fig. 2*A*). The heights of the bars are proportional to the average quality value of the 100 gene trees. Tree quality represents the proportion of congruent triplets between a gene tree and the reference species tree (fig. 1) (same value as 1—gene tree error from fig. 2*A*). Red, blue, and green areas, respectively, represent the bootstrap proportion in support of Xenarthra + Boreoeutheria, Atlantogenata (Afrotheria + Xenarthra), and Afrotheria + Boreoeutheria. On the top of the figure, the mean branch length distances from the root to elephant, hyrax, tenrec, armadillo, and sloth tips are displayed for the 131 subsets. Red, blue, and green dot colors represent the color that dominates in the corresponding bar.

these AT-rich genes can still give conflicting histories, we further excluded those that produced a gene tree with a triplet error of over 10% relative to the reference tree in figure 1. This produced a large data set of 1,640 genes, which we concatenated to perform maximum likelihood inference. To our knowledge, this alignment of 4,417,485 sites is the largest data set ever analyzed in mammalian phylogeny. As such a large concatenated data set could produce overestimated bootstrap values (Nishihara et al. 2007), we analyzed a smaller data set that contains the AT-richest alignments (GC3 below 40%) for the full species set (39 taxa, 172 genes). Bootstrap values of the larger and smaller data sets are presented in figure 4. As expected from the results of figure 3, these AT-rich alignments supported the Afrotheria rooting hypothesis. The bootstrap values of the large data set were all equal to 100, with the exception of the Cetartiodactyla + Chiroptera node. However, these values dropped significantly with the small data set: 90 to 63 for the Cetartiodactyla + Chiroptera node and 100 to 78 for the Perissodactyla + Carnivora node. The topology changed for the position of the tree shrew, which was related to either Glires or Primates depending on the data set. Afrotheria as the sister group of all other placentals and the squirrel as the sister group of all other rodents were very well supported in the two analyses. Additional analyses reinforced these results: tree inferences at the amino acid level or with more stringent quality filters on afrotherian and xenarthran sequences support the same placental and rodent roots (supplementary materials, Supplementary Material online).

Discussion

The GC Syndrome

GC content is an important feature of mammalian genes and genomes, which has been linked to various issues of functional and comparative genomics such as replication timing (Pink and Hurst 2011), methylation level (Varriale and Bernardi 2010), species life-history traits (Romiguier et al. 2010), ancestral reconstructions (Romiguier et al. 2013), and of course recombination (discussed earlier). Our results highlight that GC-rich genes are generally the less reliable to reconstruct the species phylogeny of placental mammals. Two distinct factors may explain this result: 1) the number of available characters (i.e., fewer sites and species in GC-rich alignments) and 2) the GC content itself, both potentially leading to tree reconstruction artifacts.

The first factor, that is, the smaller number of characters in GC-rich regions, could be explained by biological factors, but there is no clear consensus on this subject. On the one hand, Oliver and Marín (1996) suggested that GC-rich coding sequence regions are longer, presumably because of the AT bias in the stop-codon composition. On the other hand, Duret

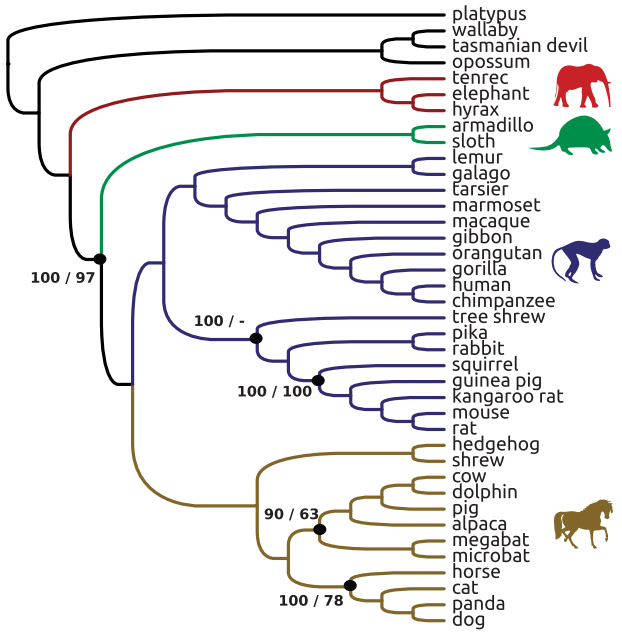

Fig. 4. Maximum likelihood support values of AT-rich data sets for the debated nodes of placental phylogeny. The first bootstrap value is for the 1,640-locus data set (4,417,485 sites), the second for the 172-gene (398,418 sites) data set. The bootstrap values of all other nodes are equal to 100.

et al. (1995) reported that AT-rich genes code for longer proteins and suggested biological links related to the isochore structure of mammals (Bernardi 1985). However, a database bias cannot be excluded, and GC-rich genes could be underrepresented because of methodological issues related to the sequencing of GC-rich genomic regions. Indeed, GC-content biases of standard and high-throughput sequencing are well documented in the literature, even though extreme AT-rich and GC-rich genes both seem to be affected by the phenomenon (Aird et al. 2011; Benjamini and Speed 2012; Dabney and Meyer 2012). This effect—biological or methodological—has never been taken into account in phylogenetic studies. AT-rich genes are currently represented by less-gapped alignments in Ensembl (Birney et al. 2004) and/or OrthoMam (Ranwez et al. 2007), and probably in other genomic databases. From a practical standpoint, this report is relevant for most phylogeny projects. Particularly in mammals, this GC syndrome should be kept in mind when selecting a subset of phylogenetic markers already present in databases or when sequencing a specific marker in several nonmodel species.

In addition to the effect of alignment gaps, we report an impact of the GC content itself on phylogenetic reconstruction. GC-rich genes increased the discrepancies between the gene genealogies and the species phylogeny. This was in agreement with our expectations, that is, that the GC content is a good long-term recombination marker, which in turn is known to increase ILS (Hobolth et al. 2011).

Biased Gene Conversion Leads to Homoplasy and Model Misspecifications

ILS does not seem to be sufficient to explain the extent of discrepancies observed among GC-rich gene trees. Here, we propose an extra hypothesis. Many authors have hypothesized that the GC-content distribution is due to a neutral mechanism, the so-called biased gene conversion (Galtier et al. 2001; Galtier and Duret 2007; Duret 2009). According to this model, a bias in the DNA repair machinery would result in a meiotic distortion favoring GC over AT alleles in high recombination regions (Eyre-Walker 1993). By increasing the GC content of recombination hotspots, this mechanism would give rise to around 100 kb regions of high average GC content. These so-called GC-rich isochore regions (Bernardi 1985) have higher gene densities and methylation rates (Eyre-Walker and Hurst 2001; Kudla et al. 2006). These regions, although important from a functional viewpoint, could be penalized by biased gene conversion. Indeed, it has been reported that the neutral nature of this mechanism could counteract natural selection and promote the fixation of deleterious mutations from A or T toward G or C (Montoya-Burgos et al. 2003; Galtier et al. 2009; Necsulea et al. 2011). However, recombination hotspots arise and disappear at a relatively fast rate in these regions (e.g., Ptak et al. [2005] report that the location of recombination hotspots was not conserved during the 6 Ma divergence between human and chimp). This prompts GC-rich genes to undergo short biased gene conversion events followed by long periods during which natural selection takes over again, thus likely promoting compensatory mutations toward AT (partly repairing deleterious substitutions fixed after biased gene-conversion episodes). Under this dynamic substitution pattern, sites tend to alternate GC and AT states, depending on whether biased gene conversion switches on or off. Such turnover leads to multiple substitutions, which are known to blur the phylogenetic signal. In our opinion, this homoplasy phenomenon could also explain the higher topological error of GC-rich genes. Ephemeral hotspots of biased gene conversion could also strongly increase model misspecification problems due to heterotachy (Delsuc et al. 2006; Nishihara et al. 2007) or base-compositional heterogeneity (Nabholz et al. 2011). Indeed, it is worth noting that the average GC3 correlates with the variance in GC3 among taxa (mean values of our 131 groups of 100 genes, $r^2 = 0.63$).

These hypotheses are in agreement with most of our results. Indeed, the GC-content dynamics of mammals has been recently described (Romiguier et al. 2010; Lartillot 2013). In these studies, nonmodel species with a very fast GC-content evolutionary pattern were identified, including the tenrec, lagomorphs (rabbit and pika), and the shrew. Interestingly, nodes involving these species are among the few recent nodes significantly affected by the GC content in our analysis (fig. 2). This is particularly striking for the shrew, which was reported to be the placental mammal most influenced by biased gene conversion (Romiguier et al. 2010). It is worth noting that ILS, homoplasy, or model misspecification induced by biased gene conversion are not exclusive hypotheses. Whatever the reason, recombination is the main culprit.

In this regard, our real ability to estimate long-term recombination is crucial. As stated in the introduction, GC content is the best-known predictor of the long-term average recombination rate, due to the effects of biased gene conversion.

However, biased gene conversion is supposed to require time to properly imprint the signature of a recombination hotspot. Consequently, GC content best reflects the local recombination rate in case of high hotspot stability. Interestingly, such a high hotspot stability has been reported in Canidae, which was one of the nodes here most affected by the GC content (fig. 2, dog + panda node). Indeed, dogs and their wild relatives exhibit a nonfunctional PRDM9, a protein involved in the short life cycle of recombination hotspots in other mammals (Muñoz Fuentes et al. 2011). Because of this loss of function, local GC content in the dog quite accurately reflects the recombination rates (Axelsson et al. 2012). This could likely explain the strong GC effect found for the Caniformia node (dog + panda).

The Root of Placentalia: Afrotheria?

As shown in figure 3, the Afrotheria and the Atlatogenata rooting hypotheses dominate our data set. This is consistent with recent molecular studies, which excluded a Xenarthra root of Placentalia and supported an Atlantogenata (Song et al. 2012) or an Afrotheria root (McCormack et al. 2012). Nevertheless, a recent study including a large morphological character data set supports a Xenarthra rooting (O'Leary et al. 2013). However, as illustrated in its supplementary material, the morphological characters analyzed without the addition of molecular data did not retrieve three of the four placental super-orders (Afrotheria, Laurasiatheria, and Euarchontoglires). This suggests a pervasive conflict between morphological and molecular data regarding the deepest branches of placentals—presumably because of the predominant morphological homoplasies documented in this clade (Springer et al. 2007).

As illustrated on the top of the figure 3, GC-rich genes have a faster evolutionary trend. Indeed, trees with high average root-to-tip branch lengths (above 0.4) for Afrotheria and Xenarthra tend to support an Atlantogenata root. In phylogeny, it is well known that high evolutionary rates can lead to homoplasy and long-branch attraction (Brinkmann et al. 2005; Nishihara et al. 2007). Here, we suggest that these artifacts are by-products of biased gene conversion. Indeed, in addition to the theoretical insights introduced above, concrete examples of recombination-associated episodes of accelerated evolution have been reported. While inducing a sudden burst of GC increase (a striking example is provided by the Fxy gene in mouse [Montoya-Burgos et al. 2003]), biased gene conversion is responsible for the fastest evolving regions in the human genome, thus misleading natural selection scans (Galtier and Duret 2007; Kostka et al. 2011). Close to the placental root, such episodes would increase the length of the already long Afrotheria and Xenarthra branches, potentially prone to long-branch attraction (Nishihara et al. 2007). Biased gene conversion increases this risk and could lead to overestimating the confidence in their sister-group relationship, the so-called Atlantogenata clade. Our results show that, actually, this node is supported in the first place by the less-reliable phylogenetic markers of the genome: fast-evolving GC-rich genes, which are prone to recombination, ILS, homoplasy, and topological errors on widely accepted nodes. On a genomic scale, equally averaging the signal of all genes leads to mixing support for the root of Placentalia. Taking the better reliability of AT-rich genes into account allowed us to uncover their agreement in favor of the Afrotheria rooting hypothesis. Furthermore, we note that sequences of Afrotherian and Xenarthran species come from low-coverage genomes (1.6–2x), which could result in a much higher error rate, larger amount of missing data, less accurate alignments or improper orthology assessment. These issues are particularly true in GC-rich regions (Aird et al. 2011; Benjamini and Speed 2012; Dabney and Meyer 2012), which seem definitely less reliable to unravel the placental root.

Interestingly, a recent study based on noncoding ultraconserved elements similarly supported the Afrotheria rooting (McCormack et al. 2012). This might be due to the fact that these noncoding regions are most commonly located in AT-rich isochores, that is, the recombination coldspots of mammalian genomes (Fullerton et al. 2001; Duret and Arndt 2008). In confirmation of this hypothesis, the GC content of the whole data set of McCormack et al. (2012) is as low as 38% (917 loci), when compared with the 55% of our 13,111 coding sequences, which are known to often be located near recombination hotspots (Fullerton et al. 2001; Duret and Arndt 2008). This difference could explain why similar coalescent-based methods (Liu et al. 2009) support Atlantogenata with coding sequences (Song et al. 2012), whereas an Afrotheria rooting is supported when applied to noncoding sequences (McCormack et al. 2012).

To unravel the trickiest nodes of the placental tree, we suggest focusing on markers located in AT-rich isochores (coding/noncoding sequences or rare genomic changes) and use a broader taxon sampling. Indeed, trying to cure the existing data set did not allow resolving all nodes (see the tree shrew placement and the laurasiatherian orders), and relies on arbitrary thresholds. In addition to increase taxon sampling, this issue could be addressed by improving current inference methods. It is well known that gene trees could be regarded as contradicting testimonies of the past. Current coalescent methods (Liu et al. 2009, 2010) try to resolve the conflicts, even when the stories differ markedly, whereas curated data sets ignore the claims of the least reliable witnesses. We suggest that the ideal solution is somewhere in-between—taking all the available information into account but putting different weight on it. Based on several criteria such as the GC content, coalescent methods should give more importance to more reliable gene trees. Reconstruction of ancestral substitutions along each branch of a tree (Romiguier et al. 2012; Dutheil et al. 2012) can pinpoint local increases in GC content and could be used to more precisely infer which genes are unreliable with respect to specific nodes. This combination of increased taxon sampling (Meredith et al. 2011), improved coalescent methods (Liu et al. 2009, 2010), and accurate sequence reliability measures is probably the key to ultimately producing a robust, fully resolved placental tree.

Conclusion

Through a genomic scale analysis, this study provides the first evidence of a strong base composition effect on the accuracy of a gene phylogeny with respect to the species phylogeny. Footsteps of recombination hotspots, GC-rich genes tend to give rise to erroneous species phylogenies, which is in agreement with well-studied phenomena such as ILS and biased gene conversion. We believe that the GC content is one out of several potential criteria required to distinguish the most reliable phylogenetic markers. On the basis of a genome-wide trend, we suggest a possible resolution of the long-debated position of the placental root. Although providing an explanation for recent disagreeing results (Song et al. 2012; McCormack et al. 2012), this study suggests that recombination hotspots, and thus GC-rich sequences, prevented us from identifying Afrotherians as the first offshoot of the placental tree.

Materials and Methods

Data Set

Sequence alignments were extracted from the OrthoMaM v7 database (Ranwez et al. 2007), which contains curated 1-to-1 orthologous markers from the 39 mammalian genomes available in Ensembl, release 67 (Birney et al. 2004). We used all available coding sequences (13,111), representing more than 24,000,000 nucleotide sites. We also used the maximum likelihood tree provided with each alignment in OrthoMaM (Ranwez et al. 2007). These trees have been inferred under a GTR + GAMMA model using RAxML (Stamatakis 2006b) on nucleotide alignments obtained with MAFFT (Katoh et al. 2002), refined by MACSE (Ranwez et al. 2011), and curated using trimAl (Capella-Gutiérrez et al. 2009).

GC3 Content

The GC content is related to recombination because of biased gene conversion, which is a neutral mechanism (Galtier et al. 2001; Galtier and Duret 2007; Duret 2009). To avoid the confounding effect of natural selection, we computed the GC percentage at the third position of codons (GC3). Mostly synonymous, substitutions on these positions are supposed to more accurately measure recombination. One GC3 was computed per alignment using the Biopython library (Cock et al. 2009). The GC3 of an alignment is defined as the mean of the G + C percent on third codon position of each taxon.

Control of Gene Length and Number of Species

The number of sites and species available per gene could influence the accuracy of its inferred phylogeny. Analyses were performed on two data sets to distinguish the effect of GC content/recombination and that of the alignment gappiness: the original one consisting of alignments from OrthoMaM, and the gap-homogenized one. The latter data set attempted to homogenize the number of species and the number of sites along the GC3 gradient of our 13,111 genes. First, we sorted all genes according to their GC3 content to pair each gene with another: the least GC-rich gene was paired with the most GC-rich gene, the second least GC-rich gene was paired with the second most GC-rich gene, and so on. For each pair, when a site contained a missing character state for a given species in one alignment, we replaced the corresponding nucleotide in the other alignment by the same missing character state. Thus, alignments of a pair had the same structure, with the same species, gaps and missing data. With this gap-homogenized data set, we controlled for the fact that a contrast between the AT- and GC-richest genes would be only due to a different amount of characters.

A maximum likelihood phylogenetic tree was recomputed for each gap-homogenized alignment. Those trees have been inferred according to the procedure described earlier for OrthoMaM (Ranwez et al. 2007). Note that in 0.6% cases, alignments had too many gaps to be analyzed.

Gene Tree Subsets

We divided our 13,111 gene trees into 131 subsets according to three different strategies. In the first strategy, clusters of 100 gene trees were built according to the GC3 content of their alignments: the first cluster contained the 100 trees inferred from the 100 most AT-rich alignments, the last one contained the 100 trees inferred from the 100 most GC-rich alignments. We called this the noncumulative method, in contrast to strategies 2 and 3. In the second strategy, the first subset contained the 100 most AT-rich genes, the second subset the 200 most AT-rich genes, and so on, until the 131st that contained all the gene trees. Hence, those subsets contained not only an increasing amount of information but also an increasing number of trees from GC-rich genes. We called this the GC-cumulative method. Similarly, in the AT-cumulative method (third approach), subsets contained an increasing amount of information, but also an increasing number of trees from AT-rich genes.

Reference Tree

We compared all our gene trees with a reference species tree. This reference tree, based on the literature (Springer et al. 2004; Meredith et al. 2011), only contains nodes that are well accepted by the scientific community (see topology of fig. 1). We left unresolved the debated nodes: the root of Placentalia (relationships among Afrotheria, Xenarthra and Boreoeutheria), relationships among Cetartiodactyla, Perissodactyla, Chiroptera, and Carnivora, position of Scandentia (tree shrew) within Euarchontoglires, and the position of the squirrel relative to murids and caviomorph rodents.

Measuring Error in Gene Tree Reconstructions

We used several topological distances to measure the topological error of a gene tree: their percentage of absent bipartitions (Robinson and Foulds 1981), triplets, and quartets (Bansal et al. 2009) compared with resolved nodes of the reference species tree. Absent bipartitions were computed via part of the Robinson–Foulds metrics (number of partitions implied by the reference tree but not in the gene tree) using a homemade program based on the Bio++ library

(Dutheil et al. 2006). Triplet and quartet distances were computed with the DQUAD program (Ranwez et al. 2010). As pinpointed by Bansal et al. (2009), tree measures do not react similarly to topological errors, we hence test those three different distances to verify that the better results obtained with AT-rich genes were independent of the considered measure.

Supertree Approach

We used a supertree approach to determine which node was more affected by the GC content. Thanks to the SuperTriplet program (Ranwez et al. 2010), we computed 131 branch supports for each well-established node according to the 131 subsets of 100 trees.

Supermatrix Method

We compared topology and support values obtained from a supermatrix containing the 100 most AT-rich genes versus those obtained from a supermatrix containing the 100 most GC-rich genes. Concatenated alignments were used to infer the maximum likelihood tree using RAxML v7.2.8, GTRCAT model (for nucleotide analyses) or PROTCATLG model (for amino-acid analyses) (Stamatakis 2006a, 2006b) and 100 bootstrap replicates. We used the same procedure for the 131 subsets of 100 genes (results of fig. 3), the 1,640-gene data set (all alignments below 40% GC3 content with a triplet topology error proportion below 0.1), and the 172-gene data set (all alignments below 40% of GC3 content containing the 39 species of this study).

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank three anonymous referees for constructive comments. This work was supported by Agence Nationale de la Recherche grants ANR-10-BINF-01-02 "Ancestrome" and by European Research Council Advanced Grant ERC 232971 "PopPhyl." This publication is contribution number 2013-069 of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554-CNRS-UM2-IRD).

References

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 12:R18.

Axelsson E, Webster MT, Ratnakumar A, Ponting CP, Lindblad-Toh K. 2012. Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome Res.* 22:51–63.

Bansal MS, Dong J, Fernández-Baca D. 2009. Comparing and aggregating partially resolved trees. *Science* 412:6634–6652.

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 40:e72.

Bernardi G. 1985. The mosaic genome of warm-blooded vertebrates. *Science* 228:953–958.

Birney E, Andrews TD, Bevan P, et al. (48 co-authors). 2004. An overview of Ensembl. *Genome Res.* 14:925–928.

Brinkmann H, Van Der Giezen M, Zhou Y, Poncelin De Raucourt G, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol.* 54:743–757.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134: 1289–1303.

Churakov G, Kriegs JO, Baertsch R, Zemann A, Brosius J, Schmitz J. 2009. Mosaic retroposon insertion patterns in placental mammals. *Genome Res.* 19:868–875.

Cock PJA, Antao T, Chang JT, et al. (11 co-authors). 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423.

Dabney J, Meyer M. 2012. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques* 52:87–94.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2:e68.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol.* 24: 332–340.

Delsuc F, Brinkmann H, Chourrout D, Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* 439:965–968.

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

Delsuc F, Scally M, Madsen O, Stanhope MJ, De Jong WW, Catzeflis FM, Springer MS, Douzery EJP. 2002. Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. *Mol Biol Evol.* 19:1656–1671.

Duret L. 2009. Mutation patterns in the human genome: more variable than expected. *PLoS Biol.* 7:e1000028.

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

Duret L, Mouchiroud D, Gautier C. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J Mol Evol.* 40:308–317.

Dutheil J, Gaillard S, Bazin E, Glémin S, Ranwez V, Galtier N, Belkhir K. 2006. Bio ++ : a set of C ++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics* 7:188.

Dutheil JY, Galtier N, Romiguier J, Douzery EJ, Ranwez V, Boussau B. 2012. Efficient selection of branch-specific models of sequence evolution. *Mol Biol Evol.* 29:1861–1874.

Eisen JA, Fraser CM. 2003. Phylogenomics: intersection of evolution and genomics. *Science* 300:1706–1707.

Escobar JS, Scornavacca C, Cenci A, Guilhaumon C, Santoni S, Douzery EJP, Ranwez V, Glémin S, David J. 2011. Multigenic phylogeny and analysis of tree incongruences in Triticeae (Poaceae). *BMC Evol Biol.* 11:181.

Eyre-Walker A. 1993. Recombination and mammalian genome evolution. *Proc Biol Sci.* 252:237–243.

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

Fullerton SM, Carvalho AB, Clark AG. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol Biol Evol.* 18:1139–1142.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* 23:273–277.

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

Hallström BM, Janke A. 2008. Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations. *BMC Evol Biol.* 8:162.

Hallström BM, Kullberg M, Nilsson MA, Janke A. 2007. Phylogenomic data analyses provide evidence that Xenarthra and Afrotheria are sister groups. *Mol Biol Evol.* 24:2059–2068.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22: 2971–2972.

Hobolth A, Dutheil J, Hawks J, Schierup M. 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res.* 21:349–356.

Huchon D, Madsen O, Sibbald MJJB, Ament K, Stanhope MJ, Catzeflis F, De Jong WW, Douzery EJP. 2002. Rodent phylogeny and a timescale for the evolution of Glires: evidence from an extensive taxon sampling using three nuclear genes. *Mol Biol Evol.* 19:1053–1065.

Jennings WB, Edwards SV. 2005. Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees. *Evolution* 59: 2033–2047.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.

Kjer KM, Honeycutt RL. 2007. Site specific rates of mitochondrial genomes and the phylogeny of eutheria. *BMC Evol Biol.* 7:8.

Kostka D, Hubisz MJ, Siepel A, Pollard KS. 2011. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol Biol Evol.* 29:1047–1057.

Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, Schmitz J. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. *PLoS Biol.* 4:e91.

Kryukov K, Sumiyama K, Ikeo K, Gojobori T, Saitou N. 2012. A new database (GCD) on genome composition for eukaryote and prokaryote genome sequences and their initial analyses. *Genome Biol Evol.* 4:501–512.

Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.* 4:e180.

Lartillot N. 2013. Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Mol Biol Evol.* 30:489–502.

Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol Biol.* 10:302.

Liu L, Yu L, Pearl DK, Edwards SV. 2009. Estimating species phylogenies using coalescence times among sequences. *Syst Biol.* 58:468–477.

Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, Amrine HM, Stanhope MJ, DeJong WW, Springer MS. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* 409: 610–614.

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

Matzke A, Churakov G, Berkes P, Arms EM, Kelsey D, Brosius J, Kriegs JO, Schmitz J. 2012. Retroposon insertion patterns of neoavian birds: strong evidence for an extensive incomplete lineage sorting era. *Mol Biol Evol.* 29:1497–1501.

McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res.* 22:746–754.

Meredith RW, Janecka JE, Gatesy J, et al. (22 co-authors). 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* 334:521–524.

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

Muñoz Fuentes V, Di Rienzo A, Vilà C. 2011. Prdm9, a major determinant of meiotic recombination hotspots, is not functional in dogs and their wild relatives, wolves and coyotes. *PLoS One* 6: e25498.

Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* 409:614–618.

Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res.* 17:413–421.

Nabholz B, Künstner A, Wang R, Jarvis E, Ellegren H. 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. *Mol Biol Evol.* 28:2197–2210.

Necsulea A, Popa A, Cooper DN, Stenson PD, Mouchiroud D, Gautier C, Duret L. 2011. Meiotic recombination favors the spreading of deleterious mutations in human populations. *Hum Mutat.* 32: 198–206.

Nikolaev S, Montoya-Burgos JI, Margulies EH, Program NCS, Rougemont J, Nyffeler B, Antonarakis SE. 2007. Early history of mammals is elucidated with the ENCODE multiple species sequencing data. *PLoS Genet.* 3:e2.

Nishihara H, Maruyama S, Okada N. 2009. Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. *Proc Natl Acad Sci U S A.* 106: 5235–5240.

Nishihara H, Okada N, Hasegawa M. 2007. Rooting the eutherian tree: the power and pitfalls of phylogenomics. *Genome Biol.* 8:R199.

O'Leary MA, Bloch JI, Flynn JJ, et al. (23 co-authors). 2013. The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science* 339:662–667.

Oliver JL, Marín A. 1996. A relationship between GC content and coding-sequence length. *J Mol Evol.* 43:216–223.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol Biol Evol.* 5:568–583.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441:1103–1108.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol.* 21:1455–1458.

Pink CJ, Hurst LD. 2011. Late replicating domains are highly recombining in females but have low male recombination rates: implications for isochore evolution. *PLoS One* 6:e24480.

Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet.* 2:e173.

Posada D, Crandall KA. 2002. The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol.* 54:396–402.

Prasad AB, Allard MW, Green ED. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol Biol Evol.* 25:1795–1808.

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Pääbo S. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet.* 37:429–434.

Ranwez V, Criscuolo A, Douzery EJP. 2010. SuperTriplets: a triplet-based supertree approach to phylogenomics. *Bioinformatics* 26:115–123.

Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak MK, Douzery EJ. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol Biol.* 7:241.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS One* 6:e22594.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math Biosci.* 53:131–147.

Romiguier J, Figuet E, Galtier N, Douzery EJP, Boussau B, Dutheil JY, Ranwez V. 2012. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. *PLoS One* 7:e33852.

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(8):1001–1009.

Romiguier J, Ranwez V, Douzery EJP, Galtier N. 2013. Genomic evidence for large, long-lived ancestors to placental mammals. *Mol Biol Evol.* 30:5–13.

Ruths D, Nakhleh L. 2005. Recombination and phylogeny: effects and detection. *Int J Bioinformatics Res Appl.* 1:202–212.

Scally M, Madsen O, Douady CJ, Jong WWD, Stanhope MJ, Springer MS. 2002. Molecular evidence for the major clades of placental mammals. *J Mammal Evol.* 8:239–277.

Song S, Liu L, Edwards SV, Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc Natl Acad Sci U S A.* 109:14942–14947.

Springer MS, Burk-Herrick A, Meredith R, Eizirik E, Teeling E, O'Brien SJ, Murphy WJ. 2007. The adequacy of morphology for reconstructing the early history of placental mammals. *Syst Biol.* 56:673–684.

Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc Natl Acad Sci U S A.* 100:1056–1061.

Springer MS, Stanhope MJ, Madsen O, de Jong WW. 2004. Molecules consolidate the placental mammal tree. *Trends Ecol Evol.* 19: 430–438.

Stamatakis A. 2006a. Phylogenetic models of rate heterogeneity: a high performance computing perspective In: Proceedings of the 20th IEEE International Parallel Distributed Processing Symposium (IPDPS 2006); 2006 April 25–29; Rhodes Island, Greece. New York: ACM.

Stamatakis A. 2006b. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Varriale A, Bernardi G. 2010. Distribution of DNA methylation, CpGs, and CpG islands in human isochores. *Genomics* 95:25–28.

Waddell PJ, Kishino H, Ota R. 2001. A phylogenetic foundation for comparative mammalian genomics. *Genome Inform.* 12: 141–154.

Waddell PJ, Shelley S. 2003. Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, gamma-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. *Mol Phylogenet Evol.* 28:197–224.

Wildman DE, Uddin M, Opazo JC, Liu G, Lefort V, Guindon S, Gascuel O, Grossman LI, Romero R, Goodman M. 2007. Genomics, biogeography, and the diversification of placental mammals. *Proc Natl Acad Sci U S A.* 104:14395–14400.