HAL
open science

Incomplete Lineage Sorting in Mammalian Phylogenomics

Celine Scornavacca, Nicolas Galtier

To cite this version:

Celine Scornavacca, Nicolas Galtier. Incomplete Lineage Sorting in Mammalian Phylogenomics. Systematic Biology, 2016, 66, pp.112 - 120. 10.1093/sysbio/syw082 . hal-01815360

HAL Id: hal-01815360

https://hal.science/hal-01815360v1

Submitted on 30 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Point of View

Syst. Biol. 66(1):112–120, 2017
© The Author(s) 2016. Published by Oxford University Press, on behalf of the Society of Systematic Biologists.
All rights reserved. For Permissions, please email: journals.permissions@oup.com
DOI:10.1093/sysbio/syw082
Advance Access publication September 7, 2016

Incomplete Lineage Sorting in Mammalian Phylogenomics

Celine Scornavacca and Nicolas Galtier*

UMR 5554—Institute of Evolutionary Sciences, University Montpellier, CNRS, IRD, EPHE, Place E. Bataillon—CC64, 34095 Montpellier, France
**Correspondence to be sent to: UMR 5554—Institute of Evolutionary Sciences, University Montpellier, CNRS, IRD, EPHE, Place E. Bataillon—CC64, 34095 Montpellier, France; E-mail: nicolas.galtier@univ-montp2.fr.*

Received 18 January 2016; reviews returned 25 March 2016; accepted 4 September 2016
Associate Editor: Tanja Stadler

Abstract.—The impact of incomplete lineage sorting (ILS) on phylogenetic conflicts among genes, and the related issue of whether to account for ILS in species tree reconstruction, are matters of intense controversy. Here, focusing on full-genome data in placental mammals, we empirically test two assumptions underlying current usage of tree-building methods that account for ILS. We show that in this data set (i) distinct exons from a common gene do not share a common genealogy, and (ii) ILS is only a minor determinant of the existing phylogenetic conflict. These results shed new light on the relevance and conditions of applicability of ILS-aware methods in phylogenomic analyses of protein coding sequences. [Coalescence; exon; Supertree; phylogeny; placental.]

Dealing with the incongruence between genes is a central issue in current phylogenomics (Galtier and Daubin 2008; Rannala and Yang 2008; Degnan and Rosenberg 2009; Szöllősi et al. 2015). For a number of reasons, gene trees sometimes depart from species trees and differ from each other. Explanations for such conflicts include gene tree-building errors, undetected paralogy, horizontal gene transfer, and incomplete lineage sorting (ILS), when within-species polymorphism lasts longer than the time between two successive speciations. ILS, which was identified as a potential source of phylogenetic incongruence decades ago (Doyle 1997; Pamilo and Nei 1988), has recently attracted a high level of attention after the discovery that species-tree estimation methods neglecting ILS, such as concatenation (=supermatrix) approaches, can be inconsistent in the presence of ILS (Degnan and Rosenberg 2006; Liu et al. 2015). A number of ILS-aware algorithms correcting for this bias have been developed (e.g., Rannala and Yang 2003; Kubatko et al. 2009; Liu et al. 2009, 2010; Bryant et al. 2012; Wu 2012; Mirarab et al. 2014a; Bayzid et al. 2015; Mirarab and Warnow 2015), assessed (Simmons and Gatesy 2015; Simmons et al. 2016), and increasingly used in real data analyses (e.g., Song et al. 2012; Betancur-R et al. 2013; Zhong et al. 2013; Xi et al. 2014; Mirarab et al. 2014b).

It is noteworthy that the rise of interest for ILS in the phylogenomic literature was in large part driven by theoretical, not empirical, arguments. Gatesy and Springer (2013, 2014) and Springer and Gatesy (2014, 2016) vehemently criticized a number of analyses based on ILS-aware methods, recalling that these are obviously justified only if (i) ILS is a major source of conflicts, (ii) "genes" correspond to genomic blocks each having a unique history of coalescence, and (iii) blocks are sufficiently large and informative such that reasonably reliable gene trees can be reconstructed (Xi et al. 2015). Surprisingly, despite the considerable amount of literature dedicated to ILS-aware methods, these aspects have hardly been quantified from real data so far.

Here, we empirically investigate whether the conditions of applicability of coalescence-based methods are met in mammals—a group in which the controversy has been particularly intense. Using protein-coding sequence alignments, we analyze the congruence of the phylogenetic signal between and within exons for various nodes of the mammalian tree. We show that coding sequences do not behave as homogeneous phylogenetic markers, and that most of the phylogenetic incongruence in mammals is unlikely to be explained by ILS. We identify exons as reasonable units of gene-tree-based analysis in mammals, and suggest that, as far as currently available data sets in mammals are concerned, ILS-aware methods are not obviously superior to—but typically slower and less flexible than—previously published super-tree methods.

Data Sets

Sequence alignments of 7349 distinct mammalian coding exons were retrieved from the Orthomam v9 database (Douzery et al. 2014), which is based on ENSEMBL v79. Orthomam v9 includes 43 fully sequenced mammalian species, of which 39 are placentals, 3 are marsupials, and 1 is a monotreme (platypus *Ornithorynchus anatinus*). Platypus was here disregarded due to the risk of misalignment and saturation. For each exon alignment, a phylogenetic tree was reconstructed using RAxML (Stamatakis 2006) under the GTR+Gamma model and 100 bootstrap replicates were performed.

Downloaded from https://academic.oup.com/sysbio/article-abstract/66/1/112/2449707
by BIU Montpellier user
on 14 June 2018

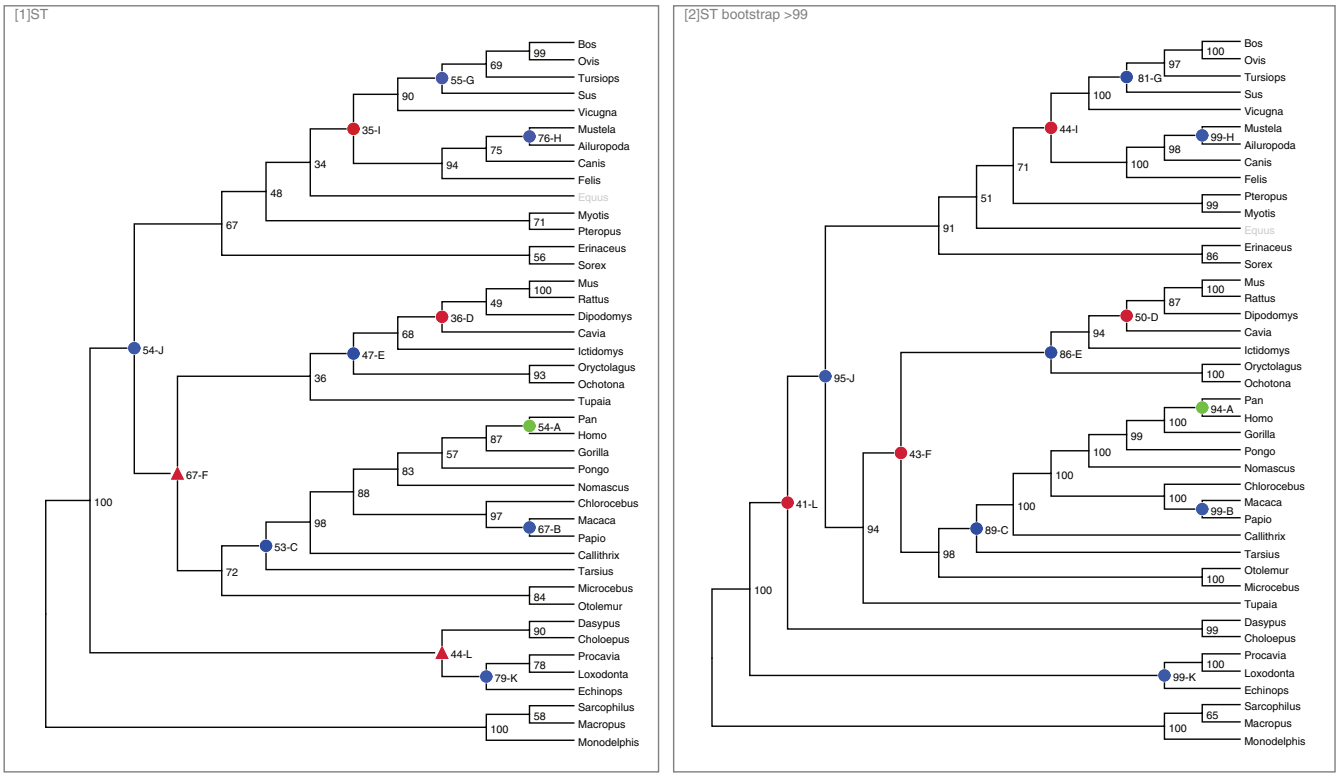

FIGURE 1. Exon-based mammalian supertree analysis. Left, SuperTriplets analysis of 5299 exon trees; Right, same analysis after nodes from source trees with bootstrap support lower than 99 were collapsed. Highlighted nodes define the (Y,Z) subtrees in the (O,(X,(Y,Z))) four-clade analyses. Nodes D, F, I, and L are controversial; they appear in red in the color version of the figure. ILS is documented at node A (green in the color version of the figure). Triangles: nodes that differ between the two trees. Node labels: letters refer to node ID in Table 1, numbers correspond to statistical support, that is, the proportion of consistent rooted triplets in source trees. The position of the horse *Equus caballus* was not taken into account in the definition of node I. Trees were drawn and annotated using Dendroscope 3 (Huson and Scornavacca 2012).

We focused on 12 non-trivial nodes of the placental tree, to which letter codes were assigned (Fig. 1 and Table 1). Four of these nodes—position of the placental root (L), early divergence within non-Eulipotyphla Laurasiatheria (I), early divergence within Euarchontoglires (F), and early divergence within Rodentia (D)—are yet unresolved (Meredith et al. 2011; Romiguier et al. 2013a), and were called "controversial". Regarding the other eight nodes, which were called "uncontroversial", a consensus has been reached in the literature. ILS has been previously demonstrated to affect gene trees for the uncontroversial *Homo/Pan/Gorilla* divergence (Hobolth et al. 2007; Dutheil et al. 2009, node A in Fig. 1). Estimates of the age of nodes were obtained from the TimeTree database (www.timetree.org).

RESULTS

Protein-Coding Sequences as Phylogenomic Entities

A number of data analyses based on ILS-aware methods have used protein-coding sequences as "loci", or "genes", or "c-genes" *sensu* Doyle (1997)—that is, genomic entities from which gene trees are estimated—implicitly assuming that the various exons of a given gene share a common, unique history of coalescence (e.g., Song et al. 2012). We questioned this assumption by asking whether exons carried by the same gene, hereafter called *co-genic exons*, yield trees more similar than exons from distinct genes. To this aim, we created five distinct mammalian data sets by sub-sampling species (Table 1). Species sub-sampling was designed in such a way that the five data sets cover a wide range of phylogenetic depth and difficulty. For each data set, we selected coding sequence alignments in which no species was missing, and at least two exons were available—applying these criteria to the full data set would yield a very small number of genes, hence the sub-sampling. Then, we calculated the Robinson and Foulds (1981) distance between exon trees and contrasted co-genic versus non co-genic pairs of exons. We found that, across the five data sets, the distributions of Robinson–Foulds distances were indistinguishable between the two categories of exon pairs (Fig. 2): two co-genic exons did not yield more similar trees than two exons carried by distinct genes, contradicting an implicit assumption behind the gene tree-based analysis of protein coding sequences. Qualitatively similar results were obtained when we generated 100 bootstrap trees per exon and compared the distributions of Robinson–Foulds distance between

Downloaded from https://academic.oup.com/sysbio/article-abstract/66/1/112/2449707
by BIU Montpellier user
on 14 June 2018

TABLE 1. Analyzed species, data sets, and nodes

Genus	Species	Clade	D1[a]	D2	D3	D4	D5	A[b]	B	C	D	E	F	G	H	I	J	K	L
Monodelphis	*domestica*	Marsupial	*		*		*											O	O
Macropus	*eugenii*	Marsupial	*		*													O	O
Sarcophilus	*harrisii*	Marsupial																O	O
Echinops	*telfairi*	Afrotheria															O	Z	X
Loxodonta	*africana*	Afrotheria	*		*		*										O	Y	X
Procavia	*capensis*	Afrotheria	*		*												O	Y	X
Choloepus	*hoffmanni*	Xenarthra			*												X		Y
Dasypus	*novemcinctus*	Xenarthra			*		*										X		Y
Otolemur	*garnettii*	Primates		*						X		X	Z				Y	X	Z
Microcebus	*murinus*	Primates		*						X		X	Z				Y	X	Z
Tarsius	*syrichta*	Primates		*						Y		X	Z				Y	X	Z
Callithrix	*jacchus*	Primates		*						Z		X	Z				Y	X	Z
Papio	*anubis*	Primates							Z	Z		X	Z				Y	X	Z
Macaca	*mulatta*	Primates		*					Y	Z		X	Z				Y	X	Z
Chlorocebus	*sabaeus*	Primates							X	Z		X	Z				Y	X	Z
Nomascus	*leucogenys*	Primates		*					O	Z		X	Z				Y	X	Z
Pongo	*abelii*	Primates		*				O	O	Z		X	Z				Y	X	Z
Gorilla	*gorilla*	Primates		*				X	O	Z		X	Z				Y	X	Z
Homo	*sapiens*	Primates	*	*	*	*	*	Y	O	Z		X	Z				Y	X	Z
Pan	*troglodytes*	Primates	*	*				Z	O	Z		X	Z				Y	X	Z
Tupaia	*belangeri*	Scandentia		*			*						X				Y	X	Z
Ochotona	*princeps*	Lagomorpha								O	O	Y	Y				Y	X	Z
Oryctolagus	*cuniculus*	Lagomorpha			*		*			O	O	Y	Y				Y	X	Z
Ictidomys	*tridecemlineatus*	Rodentia								O	X	Z	Y				Y	X	Z
Cavia	*porcellus*	Rodentia								O	Z	Z	Y				Y	X	Z
Dipodomys	*ordii*	Rodentia								O	Y	Z	Y				Y	X	Z
Rattus	*norvegicus*	Rodentia	*							O	Y	Z	Y				Y	X	Z
Mus	*musculus*	Rodentia	*	*	*		*			O	Y	Z	Y				Y	X	Z
Sorex	*araneus*	Eulipotyphla				*							O			O	Z	X	Z
Erinaceus	*europaeus*	Eulipotyphla				*	*						O			O	Z	X	Z
Pteropus	*vampyrus*	Chiroptera				*							O			X	Z	X	Z
Myotis	*lucifugus*	Chiroptera				*	*						O			X	Z	X	Z
Equus	*caballus*	Perissodactyla			*	*	*						O				Z	X	Z
Felis	*catus*	Carnivora				*							O	O	O	Y	Z	X	Z
Canis	*lupus*	Carnivora	*		*	*							O	O	X	Y	Z	X	Z
Ailuropoda	*melanoleuca*	Carnivora	*			*							O	O	Y	Y	Z	X	Z
Mustela	*putorius*	Carnivora				*	*						O	O	Z	Y	Z	X	Z
Vicugna	*pacos*	Cetartiodactyla				*							O	X		Z	Z	X	Z
Sus	*scrofa*	Cetartiodactyla				*							O	Y		Z	Z	X	Z
Tursiops	*truncatus*	Cetartiodactyla	*			*							O	Z		Z	Z	X	Z
Ovis	*aries*	Cetartiodactyla											O	Z		Z	Z	X	Z
Bos	*taurus*	Cetartiodactyla	*		*	*	*						O	Z		Z	Z	X	Z

[a]Asterisks indicate the species that belong to data sets D1–D5 shown in Fig. 2.
[b]Symbols O, X, Y, and Z define the four clades used in analyses of nodes A–L shown in Figs. 1, 3, and 4.

bootstrap trees from co-genic versus non-co-genic pairs of exons (Supplementary Fig. S1, available on Dryad at http://dx.doi.org/10.5061/dryad.1m3s2).

Spatial Autocorrelation of the Phylogenetic Signal

To investigate in a more detailed way the spatial autocorrelation of the phylogenetic signal, we focused on 12 nodes of the mammalian tree (Fig. 1). Each of these nodes defines a rooted subtree (O,(X,(Y,Z))), where O, X, Y, and Z are four mammalian clades (Table 1). For each node, we selected exon alignments of length greater than 700 bp in which at least one species of the four clades was available. These were reduced to four-species alignments by randomly sampling one species per clade—10 replicates per node were achieved this way. Then, we counted the number of informative sites in agreement with the true phylogeny (O = X, Y = Z), hereafter called *trustworthy sites*, and the number of informative sites contradicting the true phylogeny (O = Y, X = Z, or O = Z, X = Y), called *misleading sites*. The phylogenetic quality of exon e for node m was measured as

$$Q(e,m)=(n_{\mathrm{T}}-n_{\mathrm{M}})/(n_{\mathrm{T}}+n_{\mathrm{M}}),$$

where n_{T} is the number of trustworthy sites and n_{M} the number of misleading sites in exon e regarding node m. As far as controversial nodes are concerned, Q cannot be taken as a measure of quality—the truth being unknown—but can still be used to examine the congruence between exons.

For each node m and each replicate, we focused on pairs of co-genic exons, arbitrarily labeled the members of a pair as e_1 and e_2, and calculated the across

Downloaded from https://academic.oup.com/sysbio/article-abstract/66/1/112/2449707
by BIU Montpellier user
on 14 June 2018

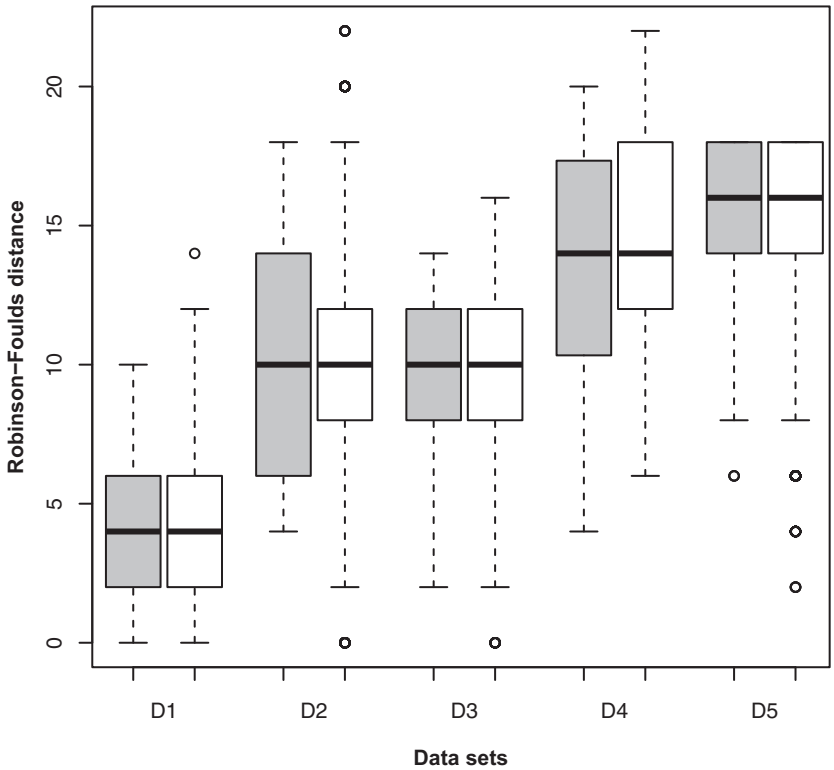

FIGURE 2. Phylogenetic incongruence between co-genic and non-co-genic exons. Each pair of boxes is for a data set (12–15 species, see Table 1). Gray: pairs of co-genic exons. White: pairs of non-co-genic exons. *Y*-axis: Robinson–Foulds distance between exon trees.

genes correlation coefficient of Q(e_1,m) and Q(e_2,m). The distribution of correlation coefficients among nodes (median across replicates) was roughly centered on zero (Fig. 3a), confirming the lack of a congruent phylogenetic signal between co-genic exons, and the generality of this effect across the whole mammalian tree. Only one node (H, i.e., monophyly of ferret and giant panda) resulted in a significantly positive correlation.

Then, we conducted the exact same analysis using half-exons rather than exons. Exons alignments were split in two parts containing the same number of informative nucleotides (excluding gaps and missing data). Then n_T, n_M, and Q were calculated separately for the 5′-half and the 3′-half of each alignment. For 10 of the 12 nodes considered here, a significant, positive correlation was found between the phylogenetic signals carried by the two halves of an exon (Fig. 3b). This analysis does not demonstrate that the two halves of an exon share a common genealogy (Springer and Gatesy 2016), but suggests that large exons might plausibly be used as genomic units for gene tree-based phylogenomic analysis in mammals.

Plausibility of ILS as the Main Source of Phylogenomic Incongruence

ILS is a source of phylogenetic conflict when genetic polymorphism is retained across two (or more) successive events of speciations. This implies that, for a given internal node, the total number of ILS-induced misleading sites cannot be higher than the number of polymorphic sites in the ancestral population. More precisely, consider a node of the species tree of the form (O,(X,(Y,Z))) and call X*, Y*, and Z* the ancestors to X, Y, and Z, respectively, at the time of the split between Y and Z. The number of ILS-induced misleading sites for node (O,X,(Y,Z))) equals the number of sites at which either X* and Y*, or X* and Z*, shared a derived allelic state. In the worst case of two simultaneous events of speciation, and assuming a panmictic, Wright–Fisher ancestral population and neutral mutations, the expectation for this number is $\theta/3$, where $\theta = 4N_e\mu$ is four times the product of ancestral effective population size, N_e, and mutation rate, μ (Wright 1938; Supplementary Fig. S2, available on Dryad).

Ancestral θ cannot be directly measured, but genome-wide estimates of heterozygosity, π, from wild-caught individuals are available in a number of extant mammalian species, including primates (Perry et al. 2012; Prado-Martinez et al. 2013), rodents (Halligan et al. 2013; Romiguier et al. 2014; Deinum et al. 2015), lagomorphs (Carneiro et al. 2012; Romiguier et al. 2014), perissodactyls (Jónsson et al. 2014) and carnivorans (Liu et al. 2014). In none of these species did the average π exceed 1% in non-coding regions, and 0.35% in coding regions. Knowing that in a panmictic population π is an unbiased estimate of θ (Tajima 1983), this provides an upper bound of ~0.12% for the expected proportion of ILS-induced misleading sites in

Downloaded from https://academic.oup.com/sysbio/article-abstract/66/1/112/2449707
by BIU Montpellier user
on 14 June 2018

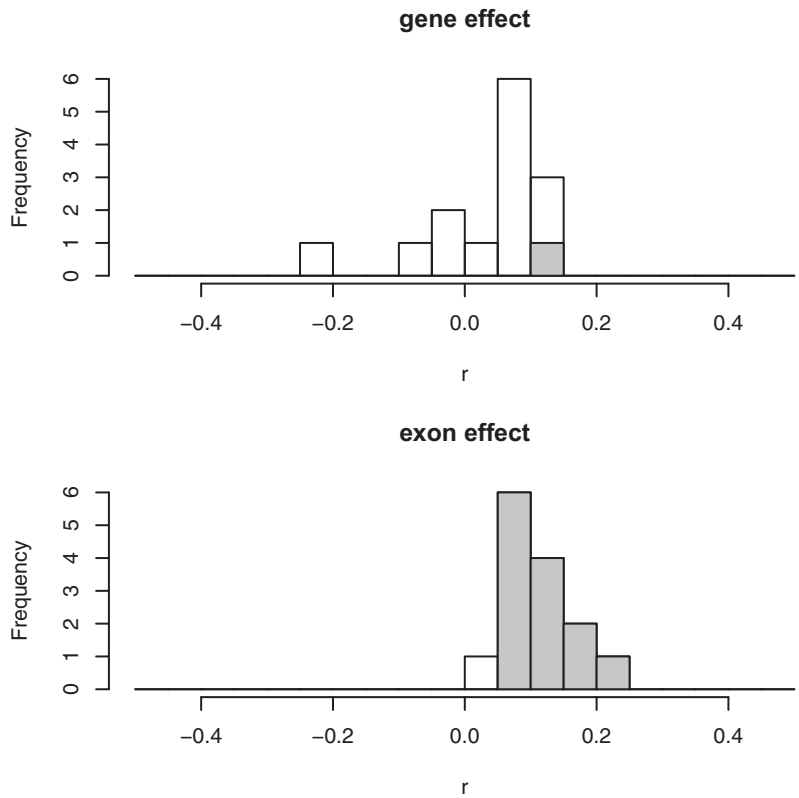

FIGURE 3. Spatial autocorrelation of the phylogenetic signal: gene effect versus exon effect. Top: across-nodes distribution of correlation coefficient of phylogenetic quality between pairs of co-genic exons. Bottom: across-nodes distribution of correlation coefficient of phylogenetic quality between the two halves of an exon. Gray: significantly positive.

protein coding sequences for any particular node of our reference mammalian phylogeny—assuming that the level of genetic polymorphism in ancestral species was similar to that of extant ones.

Figure 4 shows the per-base proportion of misleading sites, n_M/L (where L is the total alignment length), as a function of the age of the node, for 12 mammalian nodes (closed circles). In only 2 of the 12 analyzed nodes was the observed amount of misleading sites compatible with the effect of ILS only. These correspond to recent divergences within primates, including the well-documented (gorilla (human,chimpanzee)). For all the other nodes, and particularly the most ancient ones, ILS can only explain a small fraction of the observed phylogenetic incongruence, unless one assumes that the genetic diversity of ancestral mammalian species was 10 times as high as than that of mice, voles, and rabbits—a highly implausible hypothesis (Romiguier et al. 2013b, 2014).

The proportion of misleading sites is positively correlated with node age in our 4-taxa analysis (Fig. 4), reflecting the accumulation of homoplasy as sequences diverge and multiple substitutions accumulate. A number of these multiple substitutions, however, can be identified by sampling more species. We performed a similar analysis using as source data the reconstructed ancestral sequences of clades X, Y, Z, and O, rather than randomly drawn representatives. For each clade, the sequence of the most recent common ancestor of the sampled species was estimated under the HKY model (Hasegawa et al. 1985) by selecting for each site the marginally most probable state, as described by Yang et al. (1995), assuming a discrete gamma distribution of rates across sites. The proportion of misleading sites in this analysis was still generally higher than expected under ILS only (Fig. 4, open circles). For three nodes—divergence of Laurasiatheria, of Boreoeutheria, and of Glires—ILS might non-negligibly contribute to the phylogenetic conflict generated before the ancestors of clades X, Y, and Z.

GC-Content Effect

For each uncontroversial node, we calculated the GC-content at third codon positions (GC3) of each exon alignment by averaging across sequences. We found that Q, the phylogenetic quality of an exon, was negatively correlated to GC3 in seven out of eight uncontroversial nodes, the only exception being the *Homo/Pan/Gorilla* divergence. In four cases, the negative correlation was significant (*E*, *HK*, *J*, Fig. 1, Table 1). This result corroborates the report by Romiguier et al. (2013a) of a higher reliability of AT-rich than GC-rich alignments in mammalian phylogenomics.

Downloaded from https://academic.oup.com/sysbio/article-abstract/66/1/112/2449707
by BIU Montpellier user
on 14 June 2018

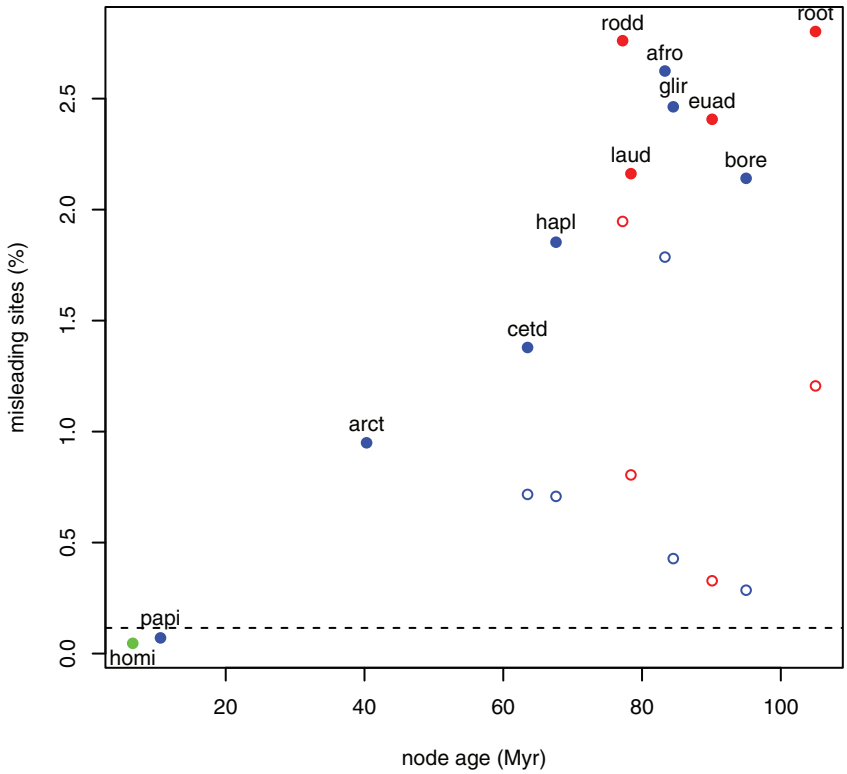

FIGURE 4. Amount of phylogenetic conflict at 12 mammalian nodes. Each dot is for a node. *X*-axis: node age. *Y*-axis: percentage of misleading sites among all sites. Dotted horizontal line: maximal expected percentage of ILS-induced misleading sites assuming panmixia and ancestral effective population sizes similar to extant ones.

Exon Tree Analysis

If ILS was indeed a minor determinant of phylogenetic conflict in mammals, we would expect ILS-aware methods to behave more or less similarly to ILS-naive ones. To test this prediction, we conducted a gene tree phylogenomic analysis in mammals based on ORTHOMAM v9 using exons as our source data set. Marsupials (*Monodelphis domestica*, *Macropus eugenii*, and *Sarcophilus harrisii*) were used to root placental trees. We kept only those trees containing at least one of the three outgroup species and we checked whether the outgroup species formed a monophyletic clade; if yes, we rooted the tree on this clade using BppReroot of the BppSuite (Guéguen et al. 2013), otherwise we discarded the tree. This procedure generated 5299 rooted exon trees, a forest we here denote by $\boldsymbol{F}$.

We applied three methods to infer a species tree from $\boldsymbol{F}$: SuperTriplets (Ranwez et al. 2010), a supertree method for rooted trees that does not explicitly account fo ILS, ASTRAL (Mirarab et al. 2014a, 2015) and MP-EST (Liu et al. 2010), two gene-tree-based methods that were specifically developed to cope for ILS. SuperTriplets and ASTRAL accept polytomic trees, and were thus applied both to $\boldsymbol{F}$ and to the forest of trees, denoted by $\boldsymbol{F}^{99}$, obtained from $\boldsymbol{F}$ by collapsing internal branches with a bootstrap value lower than 99. The running time for SuperTriplets was of 6 s, while ASTRAL took around 32 min to compute a species tree for the unrooted version of $\boldsymbol{F}$, and more than 22 h for the unrooted version of $\boldsymbol{F}^{99}$. MPEST_1.5 took around 46 min (10 independent rounds were performed).

The results of the SuperTriplets analysis are shown in Fig. 1, where on the left we have the species tree obtained from $\boldsymbol{F}$ and on the right the one obtained from $\boldsymbol{F}^{99}$. The two trees differ regarding the position of the placental root, of horse *Equus caballus* and of tree shrew *Tupaia belangeri*—three controversial branching orders. Moreover, they also differ in the support that SuperTriplets provided for each node of the estimated species tree. Importantly, the uncontroversial nodes, which were reconstructed in both the $\boldsymbol{F}$ and $\boldsymbol{F}^{99}$ analyses, each obtained a stronger statistical support in the latter analysis. The trees reconstructed by ASTRAL on $\boldsymbol{F}$ and $\boldsymbol{F}^{99}$ (Supplementary Fig. S3, available on Dryad) were identical to the ones reconstructed by SuperTriplets, respectively, on $\boldsymbol{F}$ and $\boldsymbol{F}^{99}$. The MP-EST species tree, obtained from the analysis of $\boldsymbol{F}$, differed from the one obtained by SuperTriplets and ASTRAL on $\boldsymbol{F}$ regarding the position of *Tupaia* and the position of *Cavia*, which is unresolved as well (Supplementary Fig. S3, available on Dryad). The fact that ASTRAL and MP-EST behaved very similarly to SuperTriplets is consistent with the suggestion that ILS does not substantially impact this mammalian data set.

Downloaded from https://academic.oup.com/sysbio/article-abstract/66/1/112/2449707
by BIU Montpellier user
on 14 June 2018

Discussion

Our analysis of the distribution of the phylogenetic signal across nodes, genes, and exons in mammals revealed that, for this data set, (i) exons are more plausible phylogenomic units than coding sequences, and (ii) ILS is only a minor determinant of the existing phylogenetic conflict between sites.

It has previously been argued that typical mammalian genes occupy genomic windows substantially wider than recombinational units, and are therefore unlikely to share a common genealogy (Gatesy and Springer 2013, 2014). We here empirically confirm this claim by detecting essentially no correlation in phylogenetic signal between co-genic exons. This argues against the usage of protein coding sequences as phylogenomic units in gene-tree-based phylogenomic analyses, especially if methods explicitly modeling the processes that generate conflict between gene trees, such as ILS-aware methods, are to be used (e.g., Song et al. 2012). The expected impact of ILS on phylogenetic conflicts, and therefore the length of c-genes, is dependent on species sampling and recombination rate (Springer and Gatesy 2016), and the option we took here (c-genes = exons) will presumably not fit every data set.

The relative impact on phylogenetic conflict of ILS versus undetected multiple substitutions has been investigated based on simulations (e.g., Huang et al. 2010; Lanier and Knowles 2015). Here, we show that, with the exception of very recent nodes, the density of misleading sites in ORTHOMAM v9 coding sequence alignments exceeds by large the typical within-species heterozygosity. This implies that the majority of misleading sites must have appeared posterior to speciation events by a process different from ILS—presumably multiple substitutions. Whether we will once be in position to resolve the most difficult nodes of the mammalian tree, therefore, seems to critically depend on our ability to correctly model the substitution process, not only the coalescence process, as suggested by, for example, the existence of a GC3 effect on gene tree accuracy (Romiguier et al. 2013a, this study). The work of De Maio et al. (2015) has a place in this context, offering a natural avenue for considering ILS and complex models of sequence evolution concurrently.

The exon trees we analyzed are noisy, as demonstrated by the relatively high Robinson–Foulds distances reported in Fig. 2. To reduce the amount of noise in gene trees, it is tempting to use larger phylogenomic units, such as coding sequences (Song et al. 2012) or bins of alignments (Mirarab et al. 2014b). But by doing so, one departs an implicit assumption underlying ILS-aware methods, which is that the history of a phylogenomic unit is represented by a single gene tree (Gatesy and Springer 2013, 2014; Liu and Edwards 2015). If this assumption is to be relaxed, then ILS-aware methods appear to loose their justification, compared with other super-tree methods. Using the fast and flexible SuperTriplets algorithm (Ranwez et al. 2010), we show that a pre-processing of gene trees that only selects highly supported nodes is an alternative way to reduce the phylogenetic noise (Fig. 1), which in this case yielded a species tree in good agreement with supermatrix estimates (Meredith et al. 2011) with fairly high support for most uncontroversial nodes.

Our analysis suggests that ILS is not the main cause of phylogenetic conflict in this mammalian data set, and rather points to undetected multiple substitutions, alignment errors, or hidden paralogy as the dominant conflict-generating processes. Besides, recent comparative genomic analyses suggest that in mammals hybridization and horizontal gene transfer can occur between closely related but differentiated taxa (e.g., Sankararaman et al. 2014; Li et al. 2016). This is another potential source of phylogenetic conflict, which somewhat mimics ILS if restricted to recently diverged lineages, and has been largely neglected in the mammalian phylogenomic literature so far.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.1m3s2.

Funding

This work was supported by Agence Nationale de la Recherche grant ANR-10-BINF-01-01 (Ancestrome).

Acknowledgments

We thank the Montpellier Bioinformatics and Biodiversity platform for the computing facilities and Bastien Boussau and three anonymous reviewers for helpful comments and discussions.

References

Bayzid M.S., Mirarab S., Boussau B., Warnow T. 2015. Weighted statistical binning: Enabling statistically consistent genome-scale phylogenetic analyses. PLoS One 10:e0129183.

Betancur-R R., Li C., Munroe T.A., Ballesteros J.A., Ortí G. 2013. Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). Syst. Biol. 62:763–785.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.

Carneiro M., Albert F.W., Melo-Ferreira J., Galtier N., Gayral P., Blanco-Aguiar J.A., Villafuerte R., Nachman M.W., Ferrand N. 2012. Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. Mol. Biol. Evol. 29:1837–1849.

De Maio N., Schrempf D., Kosiol C. 2015. PoMo: an allele frequency-based approach for species tree estimation. Syst. Biol. 64:1018–1031.

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Deinum E.E., Halligan D.L., Ness R.W., Zhang Y.H., Cong L., Zhang J.X., Keightley P.D. 2015. Recent evolution in *Rattus norvegicus* is

Downloaded from https://academic.oup.com/sysbio/article-abstract/66/1/112/2449707
by BIU Montpellier user
on 14 June 2018

shaped by declining effective population size. Mol. Biol. Evol. 32:2547–2558.
Douzery E.J., Scornavacca C., Romiguier J., Belkhir K., Galtier N., Delsuc F., Ranwez V. 2014. OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. Mol. Biol. Evol. 31:1923–1928.
Doyle J.J. 1997. Trees within trees: genes and species, molecules and morphology. Syst. Biol. 46:537–553.
Dutheil J.Y., Ganapathy G., Hobolth A., Mailund T., Uyenoyama M.K., Schierup M.H. 2009. Ancestral population genomics: the coalescent hidden Markov model approach. Genetics 183: 259–274.
Galtier N., Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Phil. Trans. R. Soc. Lond. B Biol. Sci. 363: 4023–4029.
Gatesy J., Springer M.S. 2013. Concatenation versus coalescence versus "concatalescence". Proc. Natl. Acad. Sci. USA. 110:E1179.
Gatesy J., Springer M.S. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol. Phylogenet. Evol. 80:231–266.
Guéguen L., Gaillard S., Boussau B., Gouy M., Groussin M., Rochette N.C., Bigot T., Fournier D., Pouyet F., Cahais V., Bernard A., Scornavacca C., Nabholz B., Haudry A., Dachary L., Galtier N., Belkhir K., Dutheil J.Y. 2013. Bio++: efficient extensible libraries and tools for computational molecular evolution. Mol. Biol. Evol. 30:1745–1750.
Halligan D.L., Kousathanas A., Ness R.W., Harr B., Eöry L., Keane T.M., Adams D.J., Keightley P.D. 2013. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PLoS Genet. 9:e1003995.
Hasegawa M., Kishino H., Yano T. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J; Mol. Evol. 22:160–174.
Hobolth A., Christensen O.F., Mailund T., Schierup M.H. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. 3:e7.
Huang H., He Q., Kubatko L.S., Knowles L.L. 2010. Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Syst. Biol. 59:573–583.
Huson D.H., Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. 61: 1061–1067.
Jónsson H., Schubert M., Seguin-Orlando A., Ginolhac A., Petersen L., Fumagalli M., Albrechtsen A., Petersen B., Korneliussen T.S., Vilstrup J.T., Lear T., Myka J.L., Lundquist J., Miller D.C., Alfarhan A.H., Alquraishi S.A., Al-Rasheid K.A., Stagegaard J., Strauss G., Bertelsen M.F., Sicheritz-Ponten T., Antczak D.F., Bailey E., Nielsen R., Willerslev E., Orlando L. 2014. Speciation with gene flow in equids despite extensive chromosomal plasticity. Proc. Natl. Acad. Sci. USA. 111:18655–18660.
Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973.
Lanier H.C., Knowles L.L. 2015. Applying species-tree analyses to deep phylogenetic histories: challenges and potential suggested from a survey of empirical phylogenetic studies. Mol. Phylogenet. Evol. 83:191–199.
Li G., Davis B.W., Eizirik E., Murphy W.J. 2016. Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). Genome Res. 26:1–11.
Liu L., Yu L., Kubatko L., Pearl D.K., Edwards S.V. 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320–328.
Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol.10:302.
Liu L., Edwards S.V. 2015. Comment on "Statistical binning enables an accurate coalescent-based estimation of the avian tree". Science 350:171.
Liu L., Xi Z., Wu S., Davis C.C., Edwards S.V. 2015. Estimating phylogenetic trees from genome-scale data. Ann. N.Y. Acad. Sci. 1360:36–53.
Liu S., Lorenzen E.D., Fumagalli M., Li B., Harris K., Xiong Z., Zhou L., Korneliussen T.S., Somel M., Babbitt C., Wray G., Li J., He W., Wang Z., Fu W., Xiang X., Morgan C.C., Doherty A., O'Connell M.J., McInerney J.O., Born E.W., Dalén L., Dietz R., Orlando L., Sonne C., Zhang G., Nielsen R., Willerslev E., Wang J. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell 157:785–794.
Meredith R.W., Janečka J.E., Gatesy J., Ryder O.A., Fisher C.A., Teeling E.C., Goodbla A., Eizirik E., Simão T.L., Stadler T., Rabosky D.L., Honeycutt R.L., Flynn J.J., Ingram C.M., Steiner C., Williams T.L., Robinson T.J., Burk-Herrick A., Westerman M., Ayoub N.A., Springer M.S., Murphy W.J. 2011. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334:521–524.
Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014a. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548.
Mirarab S., Bayzid M.S., Boussau B., Warnow T. 2014b. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346:1250463.
Mirarab S., Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52.
Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.
Perry G.H., Melsted P., Marioni J.C., Wang Y., Bainer R., Pickrell J.K., Michelini K., Zehr S., Yoder A.D., Stephens M., Pritchard J.K., Gilad Y. 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. Genome Res. 22:602–610.
Prado-Martinez J., Sudmant P.H., Kidd J.M., Li H., Kelley J.L., Lorente-Galdos B., Veeramah K.R., Woerner A.E., O'Connor T.D., Santpere G., Cagan A., Theunert C., Casals F., Laayouni H., Munch K., Hobolth A., Halager A.E., Malig M., Hernandez-Rodriguez J., Hernando-Herraez I., Prüfer K., Pybus M., Johnstone L., Lachmann M., Alkan C., Twigg D., Petit N., Baker C., Hormozdiari F., Fernandez-Callejo M., Dabad M., Wilson M.L., Stevison L., Camprubí C., Carvalho T., Ruiz-Herrera A., Vives L., Mele M., Abello T., Kondova I., Bontrop R.E., Pusey A., Lankester F., Kiyang J.A., Bergl R.A., Lonsdorf E., Myers S., Ventura M., Gagneux P., Comas D., Siegismund H., Blanc J., Agueda-Calpena L., Gut M., Fulton L., Tishkoff S.A., Mullikin J.C., Wilson R.K., Gut I.G., Gonder M.K., Ryder O.A., Hahn B.H., Navarro A., Akey J.M., Bertranpetit J., Reich D., Mailund T., Schierup M.H., Hvilsom C., Andrés A.M., Wall J.D., Bustamante C.D., Hammer M.F., Eichler E.E., Marques-Bonet T. 2013. Great ape genetic diversity and population history. Nature 499:471–475.
Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.
Rannala B., Yang Z. 2008. Phylogenetic inference using whole genomes. Annu. Rev. Genomics Hum. Genet. 9:217–231.
Ranwez V., Criscuolo A., Douzery E.J. 2010. SuperTriplets: a triplet-based supertree approach to phylogenomics. Bioinformatics 26:i115–i123.
Robinson D.R., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
Romiguier J., Ranwez V., Delsuc F., Galtier N., Douzery E.J.P. 2013a. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Mol. Biol. Evol. 30:2134–2144.
Romiguier J., Ranwez V., Douzery E.J.P., Galtier N. 2013b. Genomic evidence for large, long-lived ancestors to placental mammals. Mol. Biol. Evol. 30:5–13.
Romiguier J., Gayral P., Ballenghien M., Bernard A., Cahais V., Chenuil A., Chiari Y., Dernat R., Duret L., Faivre N., Loire E., Lourenco J.M., Nabholz B., Roux C., Tsagkogeorga G., Weber A.A., Weinert L.A., Belkhir K., Bierne N., Glémin S., Galtier N. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515:261–263.

Downloaded from https://academic.oup.com/sysbio/article-abstract/66/1/112/2449707
by BIU Montpellier user
on 14 June 2018

Sankararaman S., Mallick S., Dannemann M., Prüfer K., Kelso J., Pääbo S., Patterson N., Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507: 354–357.

Simmons M.P., Gatesy J. 2015. Coalescence vs. concatenation: sophisticated analyses vs. first principles applied to rooting the angiosperms. Mol. Phylogenet. Evol. 91:98–122.

Simmons M.P., Sloan D.B., Gatesy J. 2016. The effects of subsampling gene trees on coalescent methods applied to ancient divergences. Mol. Phylogenet. Evol. 97:76–89.

Song S., Liu L., Edwards S.V., Wu S. 2012. Resolving conflict in Eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. USA. 109:14942–14947.

Springer M.S., Gatesy J. 2014. Land plant origins and coalescence confusion. Trends Plant Sci. 19:267–269.

Springer M.S., Gatesy J. 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94:1–33.

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

Szöllősi G.J., Tannier E., Daubin V., Boussau B. 2015. The inference of gene trees with species trees. Syst. Biol. 64:e42–e62.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Wright S. 1938. The distribution of gene frequencies under irreversible mutation. Proc. Natl. Acad. Sci. USA. 24:253–259.

Wu Y. 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution 66:763–775.

Xi Z., Liu L., Rest J.S., Davis C.C. 2014. Coalescent versus concatenation methods and the placement of Amborella as sister to water lilies. Syst. Biol. 63:919–932.

Xi Z., Liu L., Davis C.C. 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. Mol. Phylogenet. Evol. 92:63–71.

Yang Z., Kumar S., Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141: 1641–1650.

Zhong B., Liu L., Yan Z., Penny D. 2013. Origin of land plants using the multispecies coalescent model. Trends Plant Sci. 18: 492–495.

Downloaded from https://academic.oup.com/sysbio/article-abstract/66/1/112/2449707
by BIU Montpellier user
on 14 June 2018