Supplementary material for: Genome-scale coestimation of species and gene trees

Bastien Boussau[1,2*], Gergely J. Szöllősi[1], Laurent Duret[1], Manolo Gouy[1], Eric Tannier[1], Vincent Daubin[1]

[1]Université de Lyon ; Université Lyon 1 ;
CNRS ; INRIA ; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive,
43 boulevard du 11 novembre 1918,
Villeurbanne F-69622, France.
[2]Department of Integrative Biology, UC Berkeley
4163A Valley Life Sciences Bldg
Berkeley, CA 94720-3140

[*]To whom correspondence should be addressed; E-mail: bastien.boussau@univ-lyon1.fr

October 7, 2012

*Corresponding author

Contents

Supplementary figures

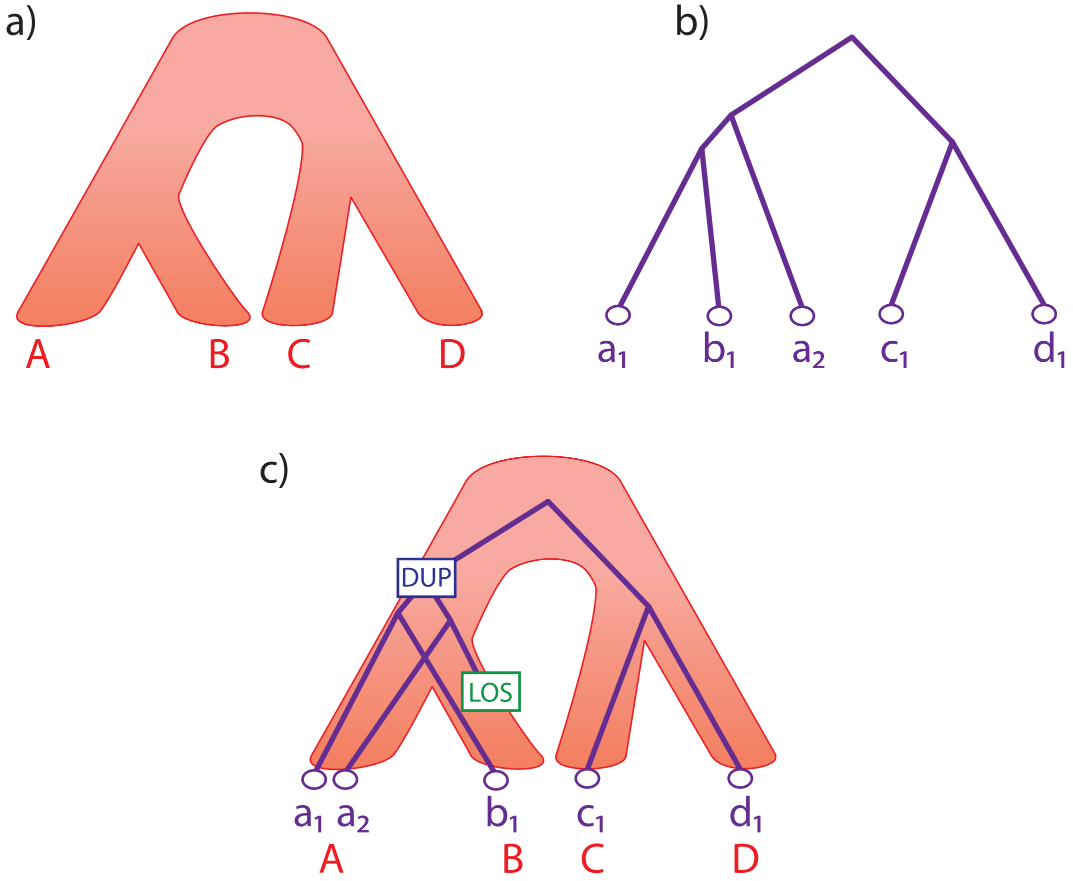


Figure 1: Species tree and gene tree. a) A species tree with four species A to D. b) A gene tree with 5 genes from the genomes of species A to D. Genes are named according to the species they are found in, and a number. Species A therefore has two genes in this gene family, a_1 and a_2. Here the gene tree is represented as rooted; usually, gene trees inferred using classical models of sequence evolution are not rooted, and the root has to be searched for. c) Reconciled gene tree. Events of gene duplication and loss have been placed to explain the gene tree shown in b given the species tree shown in a. In this scenario, here the most parsimonious one, a duplication (DUP) and a loss (LOS) are necessary.

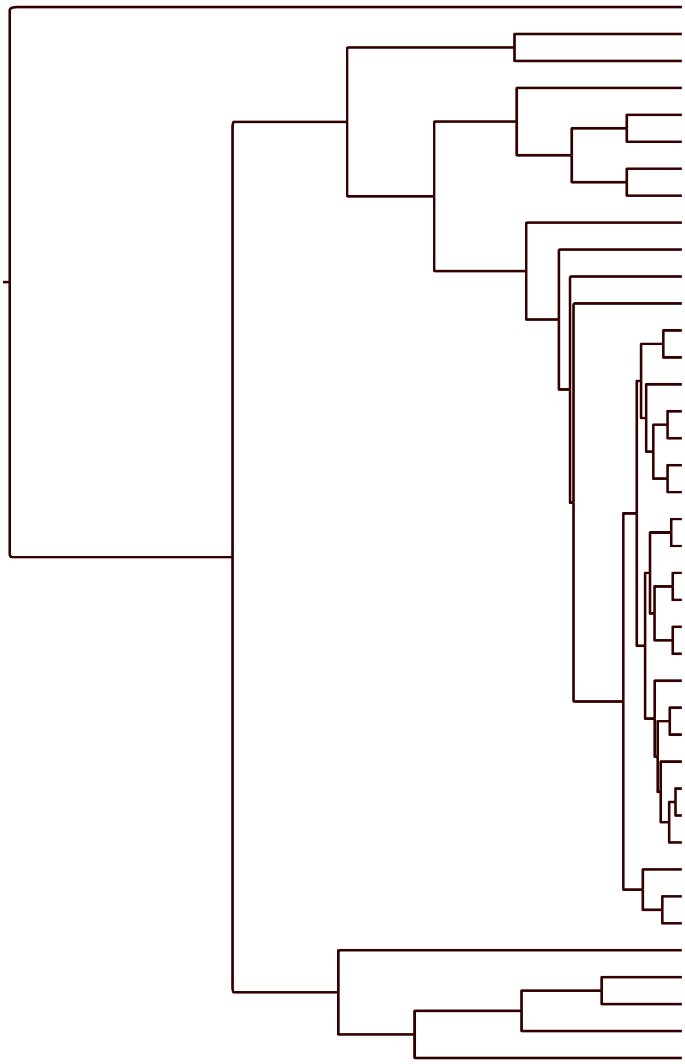


Figure 2: Species tree topology used in the simulations.

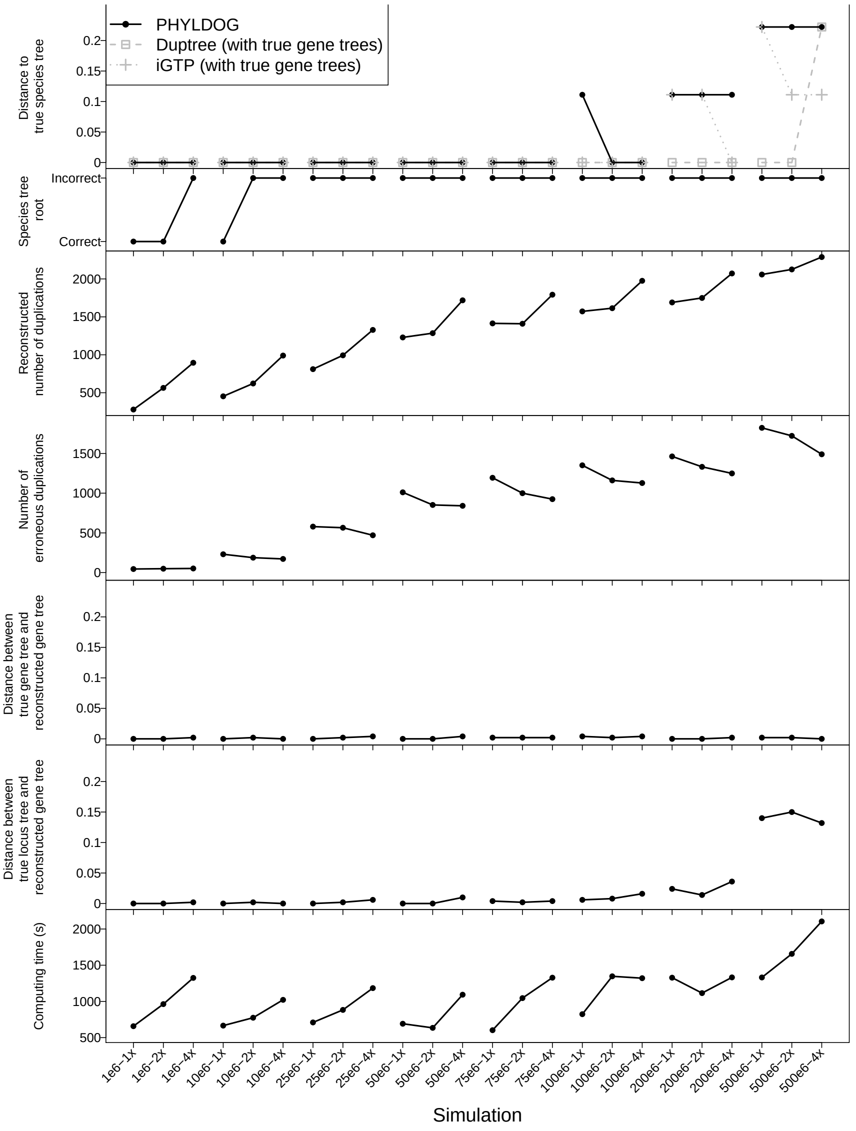


Figure 3: Accuracy of PHYLDOG on simulations involving Incomplete Lineage Sorting. All 24 data sets of simulated gene trees for 12 fly genomes were obtained from Rasmussen and Kellis (2012). These data sets were simulated with varying rates of duplication (1, 2, or 4 times the rate found in real data), and varying effective population sizes (from 1 million individuals, about the effective population size for *Drosophila melanogaster* (Rasmussen and Kellis 2012) to 500 million). All data sets contain 500 gene families. The figure shows, from top to bottom: i) the distance between the reconstructed and the true species tree, for PHYLDOG with joint reconstruction of species and gene trees, and for duptree (Wehe et al. 2008) starting from the true gene trees and iGTP (Chaudhary et al. 2010) starting from the true gene trees. ii) PHYLDOG recovery of the true root position for the species tree. iii) Reconstructed total number of duplications for the 500 gene families. iv) Difference between the reconstructed and the true (simulated) numbers of duplications. v) Robinson and Foulds normalized distance between the true (simulated) gene trees and the reconstructed gene trees by PHYLDOG. vi) Robinson and Foulds normalized distance between the true (simulated) locus trees and the reconstructed gene trees by PHYLDOG. vii) Computing time for PHYLDOG.

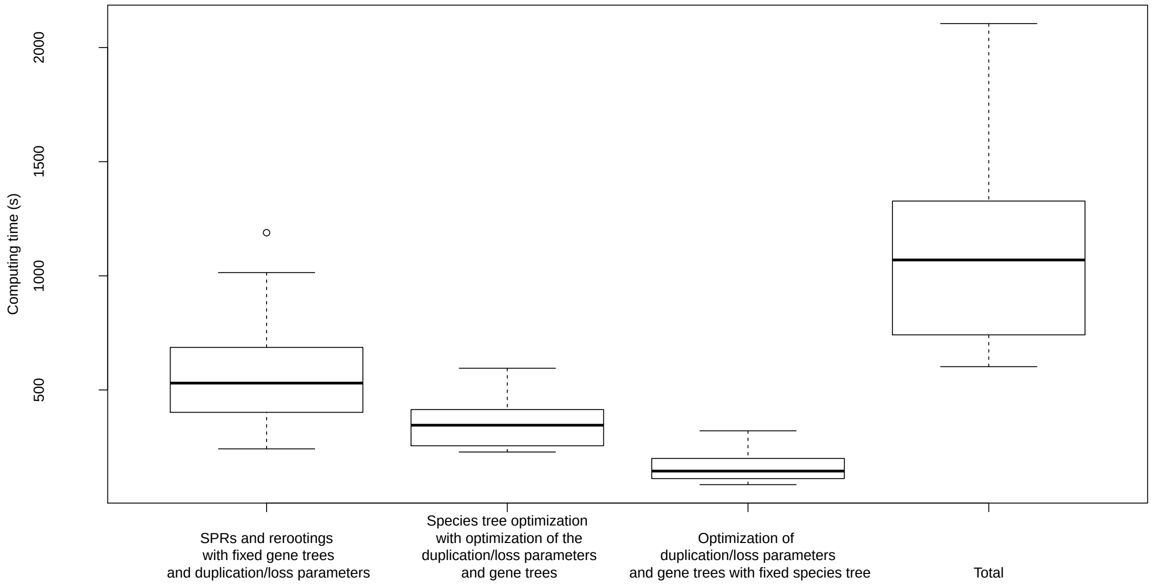


Figure 4: Total computing time and computing time for three steps of PHYLDOG's algorithm on the 24 simulations with 12 species and 500 gene trees. Computations were done on the supercomputer Jade, containing 1536 Bi-Xeon E5472 (Harpertown, 3GHz) and 1344 Bi-Xeon X5560 (Nehalem, 2,8GHz).

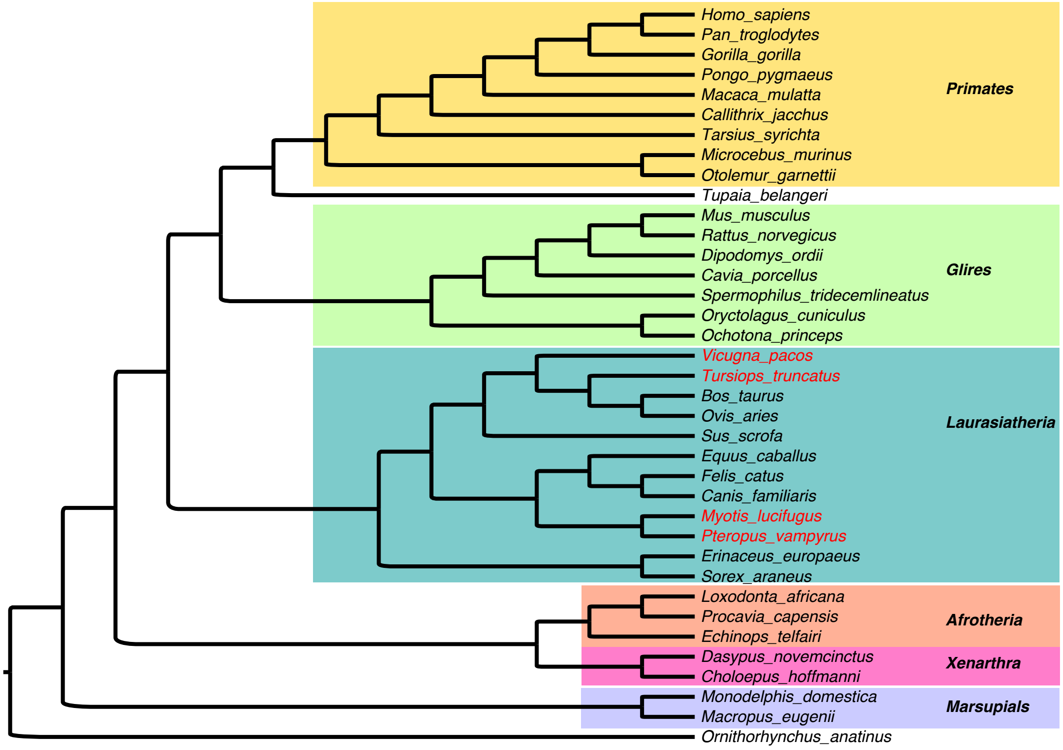


Figure 5: Species tree topology used by the Ensembl data base. Only mammalian species are shown.

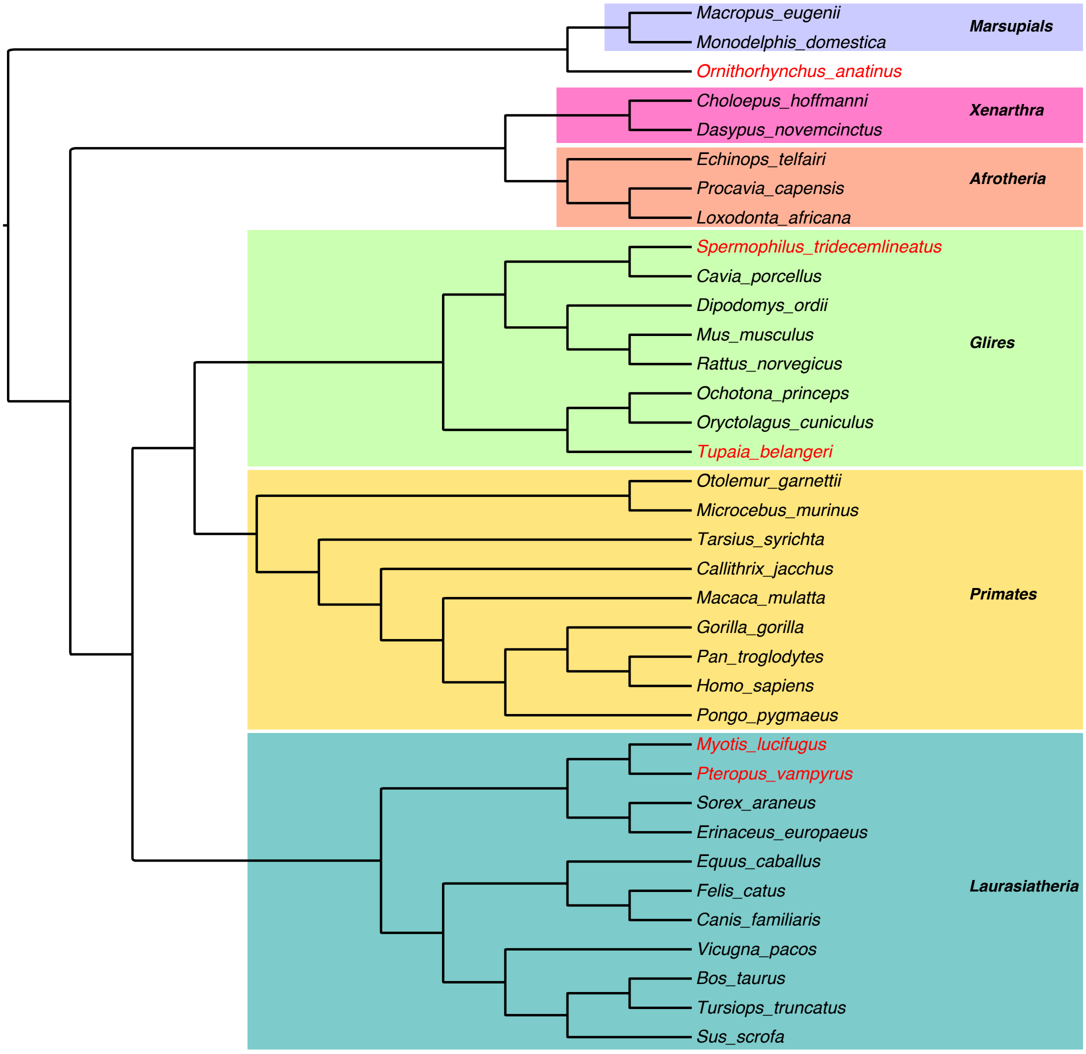


Figure 6: Species tree obtained by iGTP (Chaudhary et al. 2010) with the duplication-loss method based on PhyML input gene trees. Species names in red show incongruences with the tree inferred by PHYLDOG.

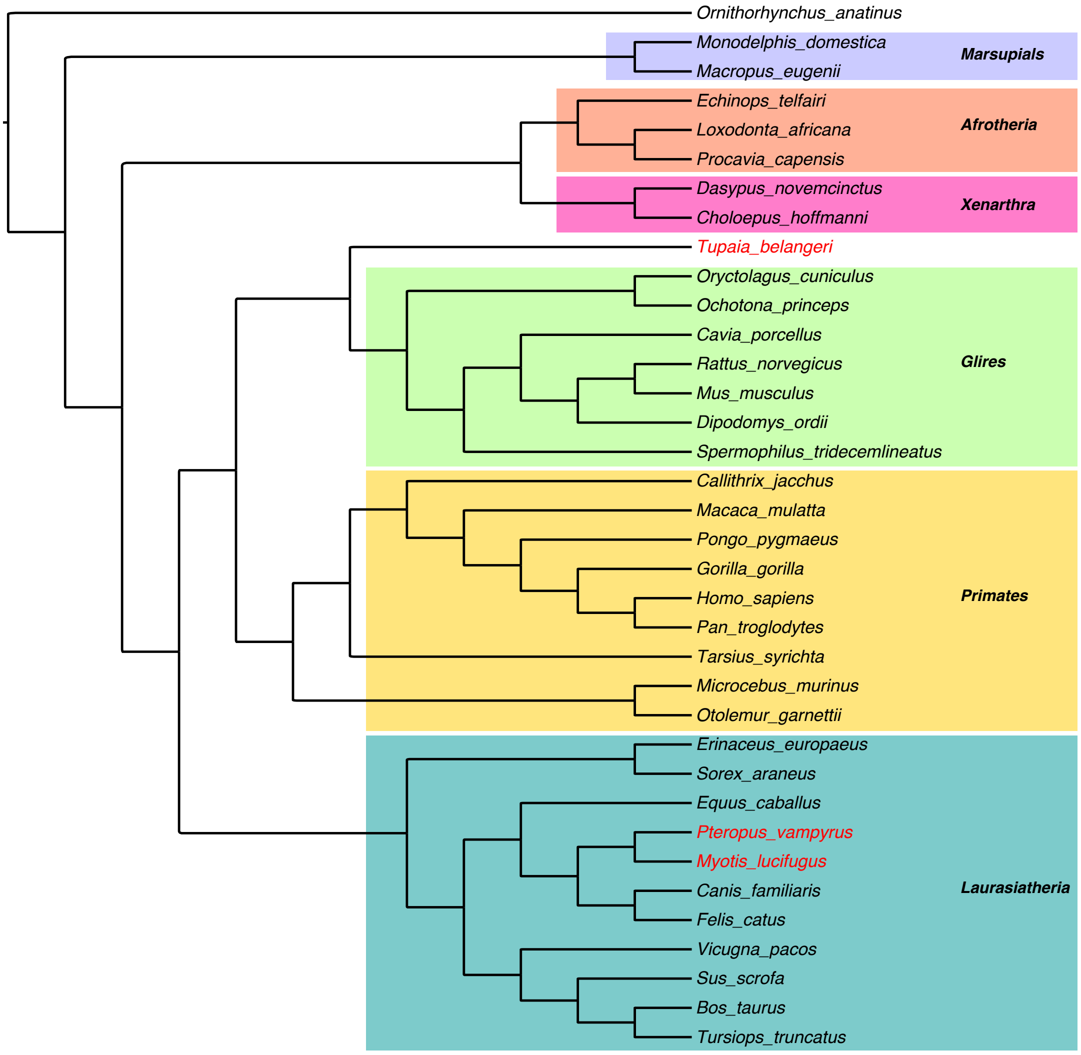


Figure 7: Species tree obtained by duptree (Wehe et al. 2008) with the gene tree parsimony method based on PhyML input gene trees. Species names in red show incongruences with the tree inferred by PHYLDOG.

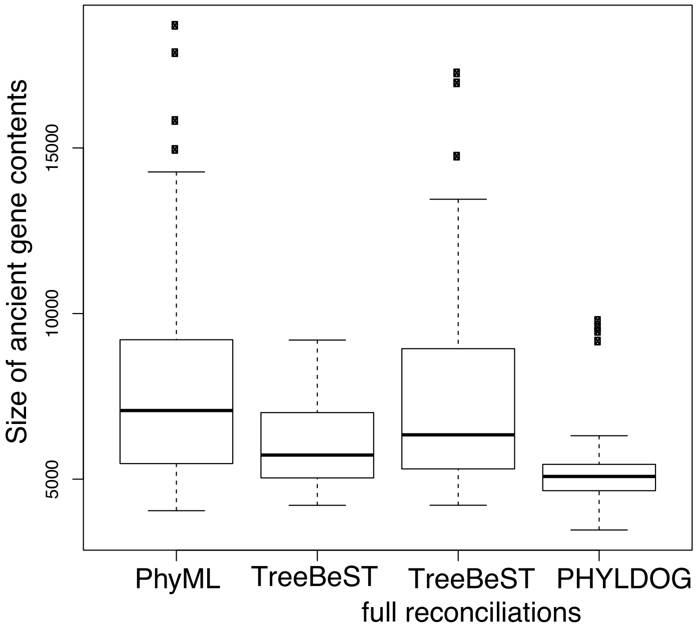


Figure 8: Sizes of ancestral gene contents (the lower the better), inferred from PhyML, TreeBeST with full parsimonious reconciliations, TreeBeST with TreeBeST partial reconciliations or PHYLDOG trees. Comparison between PHYLDOG trees and TreeBeST trees: Paired Student t-test p-value= 0.002196, Paired Wilcoxon test p-value = 0.00036. Comparison between PHYLDOG trees and TreeBeST trees with partial reconciliations: Paired Student t-test p-value= 6.615e-06, Paired Wilcoxon test p-value = 2.91e-11.

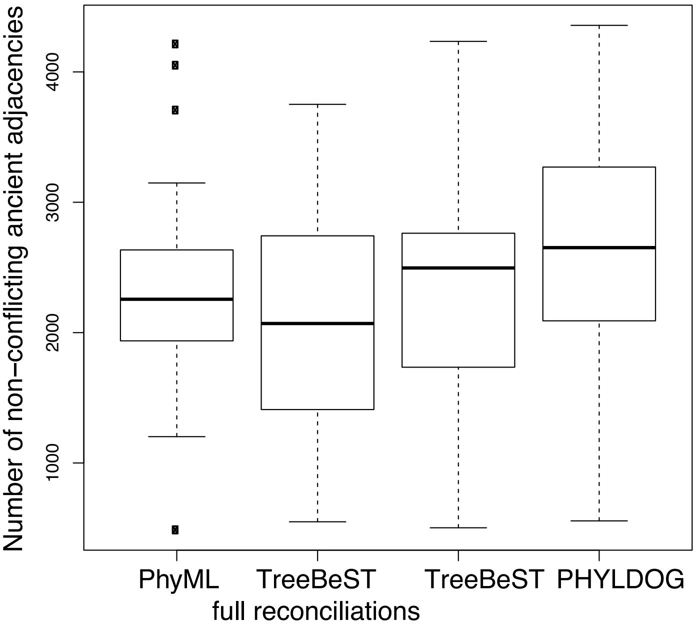


Figure 9: Number of non-conflicting adjacencies in ancestral genomes (the higher the better), inferred from PhyML, TreeBeST with full parsimonious reconciliations, TreeBeST with TreeBeST partial reconciliations or PHYLDOG trees. Comparison between PHYLDOG trees and TreeBeST trees: Paired Student t-test p-value= 0.00026, Paired Wilcoxon test p-value = 1.91e-07. Comparison between PHYLDOG trees and TreeBeST trees with partial reconciliations: Paired Student t-test p-value= 0.03, Paired Wilcoxon test p-value = 0.01.

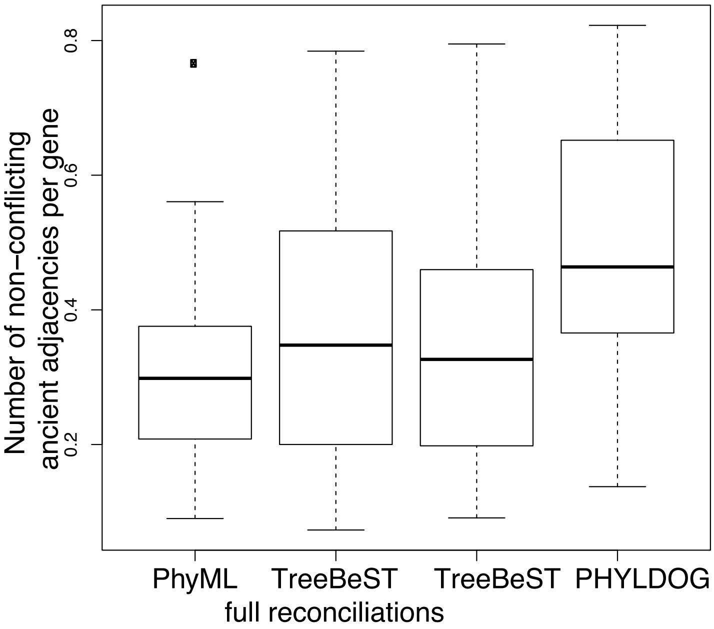
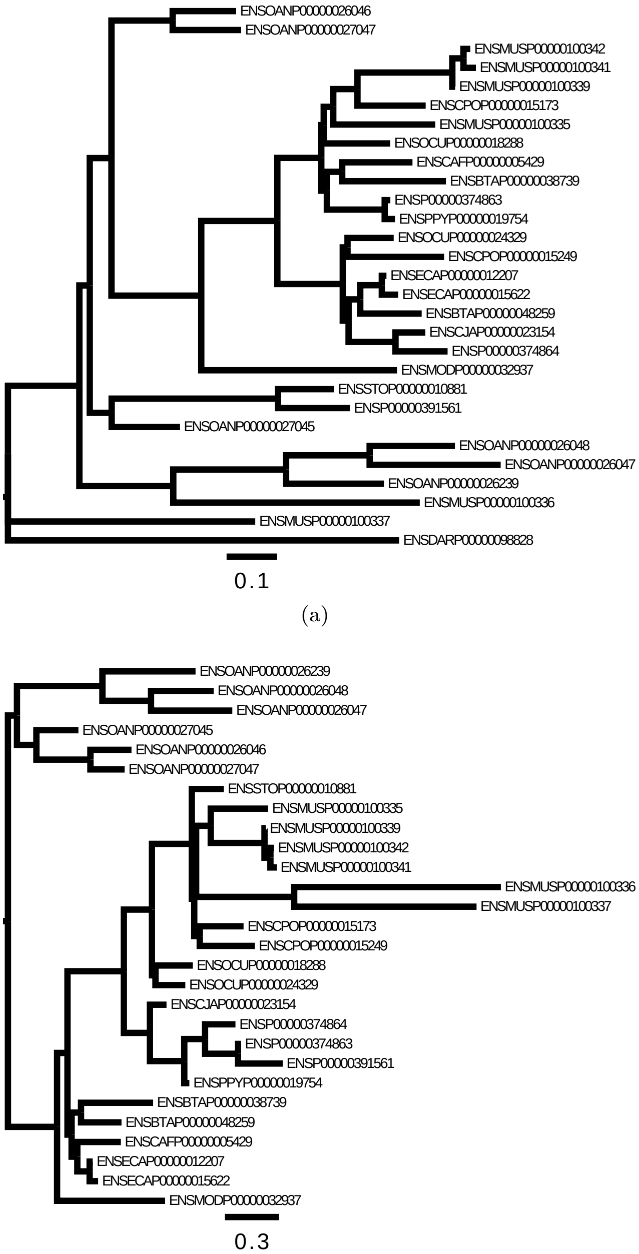


Figure 10: Normalized number of non-conflicting adjacencies in ancestral genomes (the higher the better), inferred from PhyML, TreeBeST with full parsimonious reconciliations, TreeBeST with TreeBeST partial reconciliations or PHYLDOG trees. Comparison between PHYLDOG trees and TreeBeST trees: Paired Student t-test p-value= 8.328e-12, Paired Wilcoxon test p-value = 2.91e-11. Comparison between PHYLDOG trees and TreeBeST trees with partial reconciliations: Paired Student t-test p-value= 1.82e-09, Paired Wilcoxon test p-value = 4.075e-10.



(a)

(b)

Figure 11: Example gene trees as in Fig. 5 but with sequence names instead of species names. (a) Gene tree as reconstructed by TreeBeST. This gene tree also includes a sequence from *Danio rerio*, ENSDARP00000098828, removed in Fig. 5 for clarity. (b) Gene tree as reconstructed by PHYLDOG.

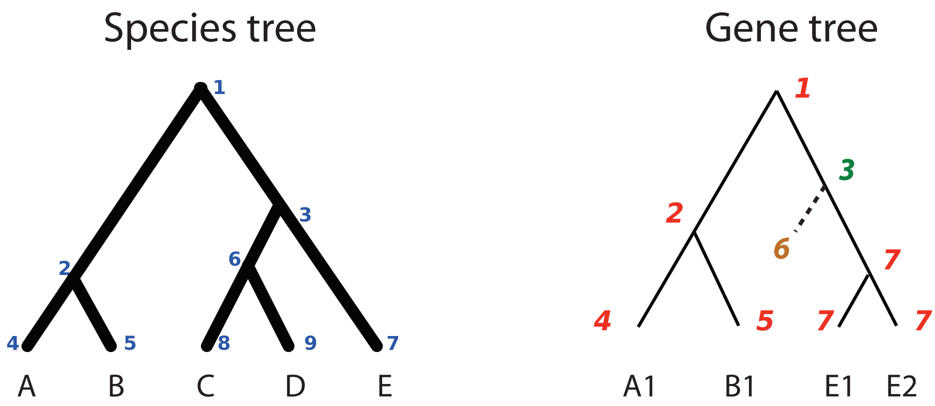


Figure 12: Mapping between the species tree and a gene tree. Left: Species tree, with its nodes numbered, from 1 at the root to higher numbers as nodes are further from the root. Right: Gene tree whose nodes are numbered in reference to the species tree. A hidden node is shown in green, and a lost node in orange.

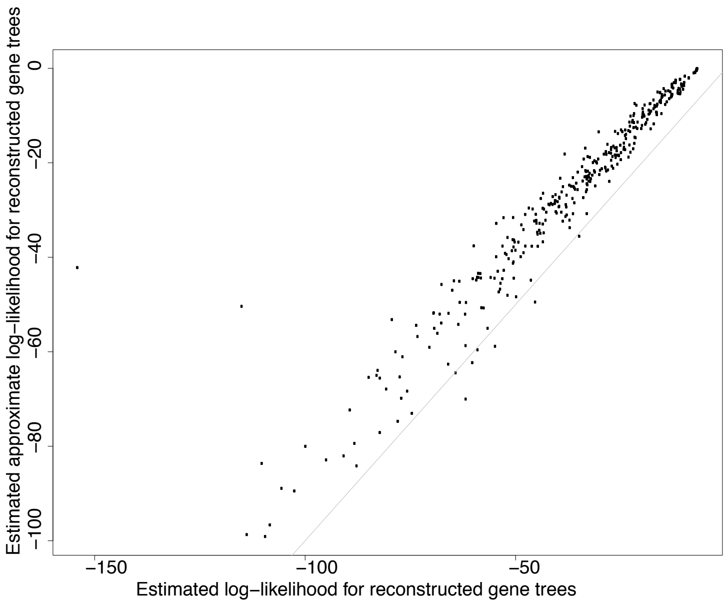


Figure 13: Correlation between the "true" likelihood as implemented in Akerborg et al. (2009) (x) and our approximate likelihood (y), computed for 400 simulated gene trees ($r^2 = 0.96$). The line $y = x$ is shown in grey. Two outlier points are caused by optimization problems in the computation of the "true" likelihood.

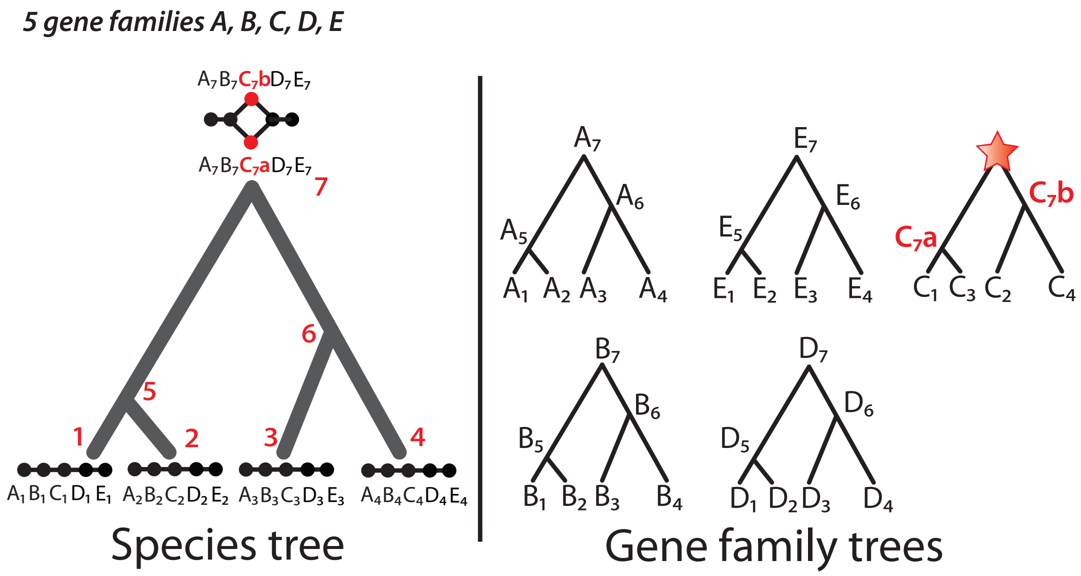


Figure 14: Use of parsimoniously inferred ancestral syntenies to uncover errors in reconstructed gene trees. See section S9 for details.

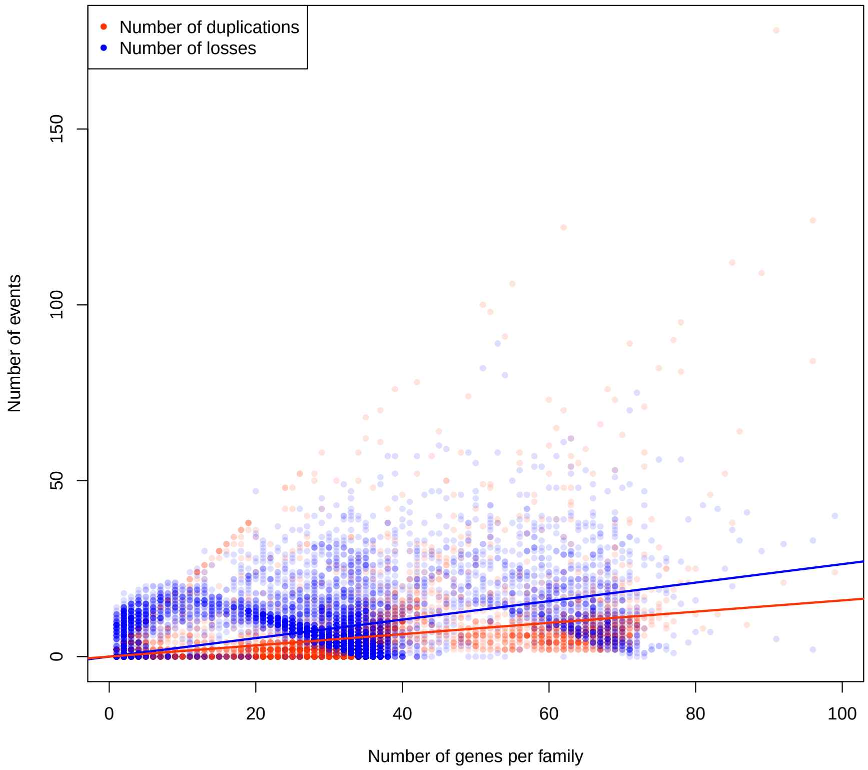


Figure 15: Number of duplications and losses per gene family as inferred by PHYLDOG, with respect to the number of genes in gene families. Two linear regressions passing through (0,0) have been drawn for duplications (red) and losses (blue).

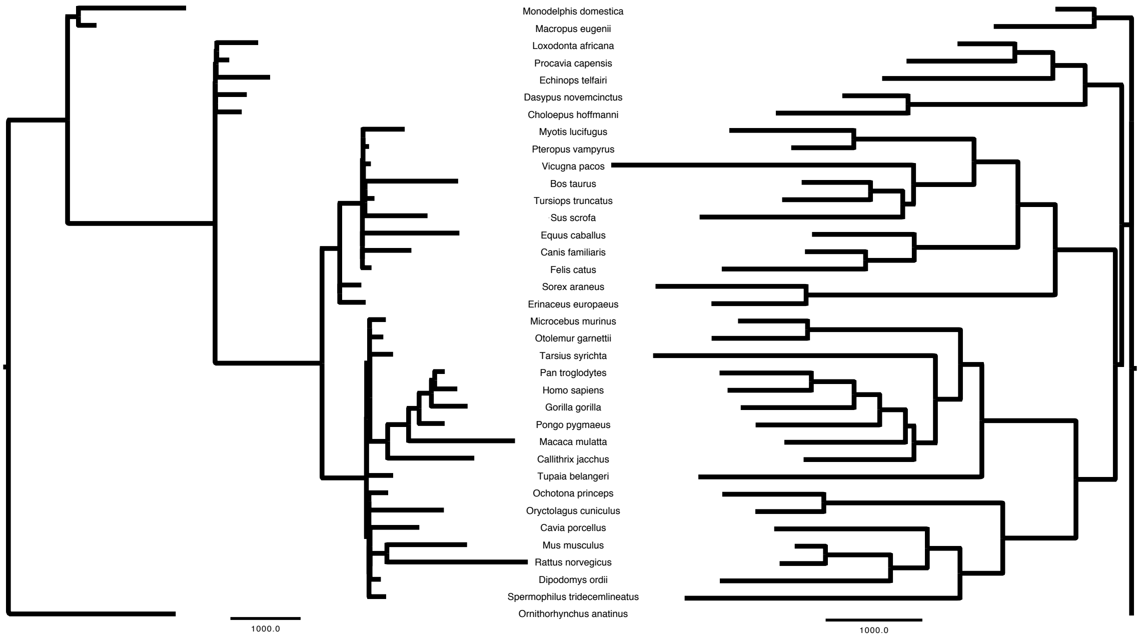


Figure 16: Mammalian tree reconstructed by PHYLDOG, with branch lengths representing the number of duplications (left) or losses (right) over all gene families. Number of losses cannot be recovered with our model for the two branches descending from the root and are therefore set to 0. We find evidence of discrepancies in sequencing quality, where the alpaca (*Vicugna pacos*), tarsier (*Tarsius syrichta*), tree shrew (*Tupaia belangeri*) and squirrel (*Spermophilus tridecemlineatus*) show very high numbers of losses, which can be linked to their low sequencing coverage (respectively 2.51x, 1.82x, 2x, 1.9x) and relatively distant evolutionary neighbors. The same cause may explain the large numbers of losses found on the branches leading to Insectivora (*Sorex araneus* and *Erinaceus europaeus*, 1.9x, 1.86x) and to Xenarthra (*Dasypus novemcinctus*, *Choloepus hoffmanni*, 2x, 2.05x).

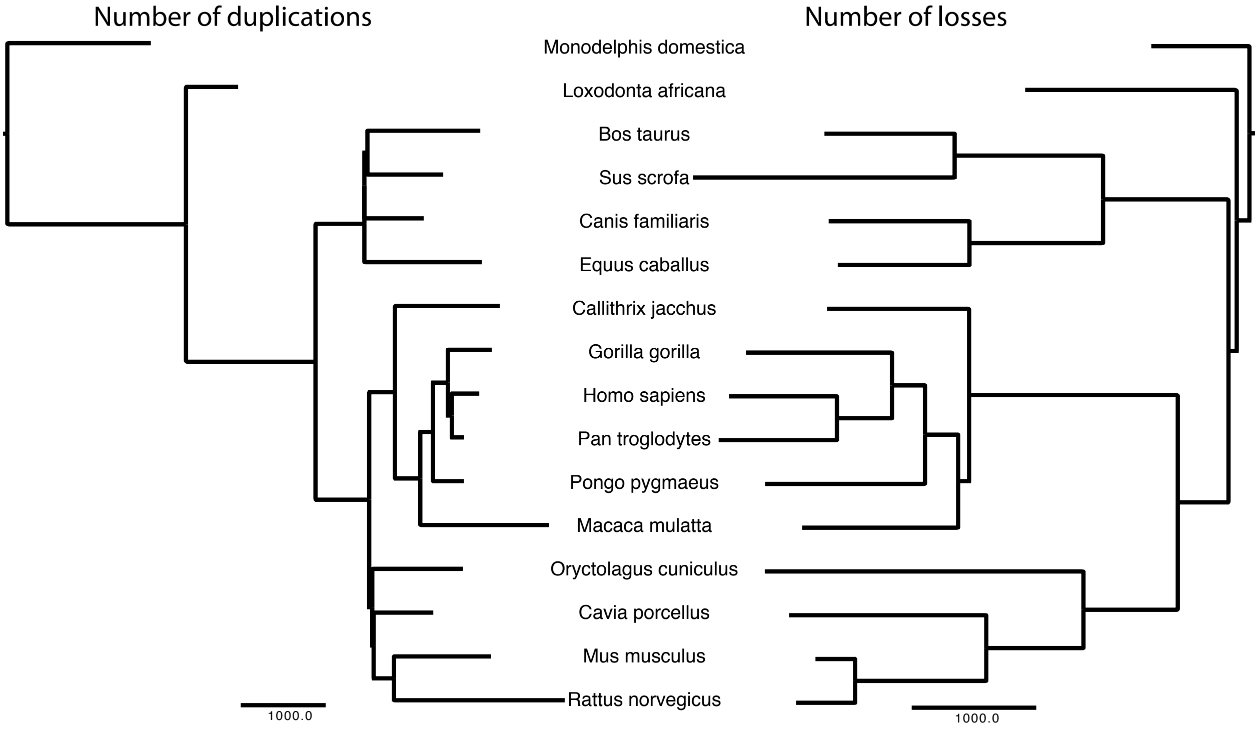


Figure 17: Same tree as in Fig. S16, but in which only therian mammals whose genome sequences have a coverage above 5x are shown. Branch lengths represent the total numbers of duplications (left) or losses (right).

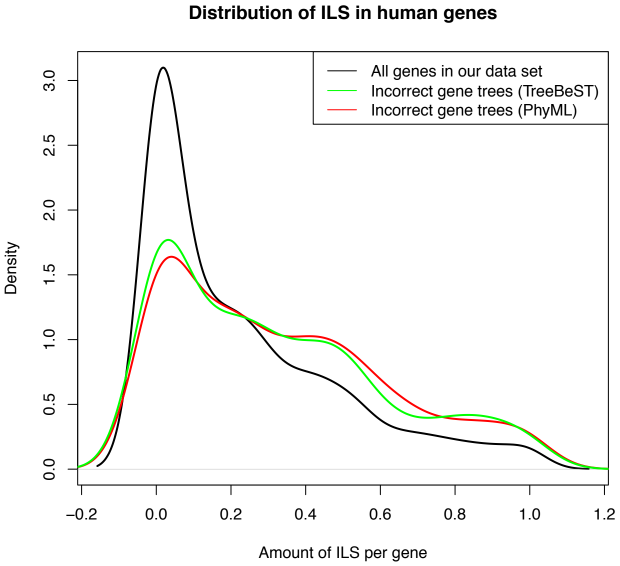


Figure 18: Distribution of the amount of Incomplete Lineage Sorting (ILS) among gene families in primates according to Scally et al. (2012). Black line: all genes in our data set. Green and red lines: distribution of ILS in gene families whose tree reconstructed by TreeBeST and PhyML respectively differ from ((*Homo sapiens*, *Pan troglodytes*), *Gorilla gorilla*)).

Supplementary table

Table 1: Ensembl V. 57 genome coverages per species

Bos taurus	7x
Callithrix jacchus	6x
Canis familiaris	7.5x
Cavia porcellus	6.79x
Choloepus hoffmanni	2.05x
Dasypus novemcinctus	2x
Dipodomys ordii	1.85x
Echinops telfairi	2x
Equus caballus	6.79x
Erinaceus europaeus	1.86x
Felis catus	1.87x
Gorilla gorilla	35x
Homo sapiens	8x
Loxodonta africana	7x
Macaca mulatta	5.2x
Macropus eugenii	2x
Monodelphis domestica	7.33x
Microcebus murinus	1.93x
Mus musculus	7x
Myotis lucifugus	1.7x
Ochotona princeps	1.93x
Ornithorhynchus anatinus	6x
Oryctolagus cuniculus	7x
Otolemur garnettii	1.5x
Pan troglodytes	6x
Pongo pygmaeus	6x
Procavia capensis	2.19x
Pteropus vampyrus	2.63x
Rattus norvegicus	7x
Sorex araneus	1.9x
Spermophilus tridecemlineatus	1.90x
Sus scrofa	"high coverage"
Tarsius syrichta	1.82x
Tupaia belangeri	2x
Tursiops truncatus	2.59x
Vicugna pacos	2.51x

1 Derivation of the analytical formula for computing $L(T, S, N|A)$

We defined $L(T, S, N|A)$ as

$$L(T,S,N|A) = \prod_i L(T_i|A_i) \times L(S,N|T_i)$$

The likelihood of the gene tree given the alignment $L(T_i|A_i)$ is proportional to the probability of the alignment given the tree:

$$L(T_i|A_i) = K_1 \times P(A_i|T_i)$$

with K_1 a constant. As in our model, when the gene tree is fixed, the probability of the alignment is independent from S and N, we have

$$L(T_i|A_i) = K_1 \times P(A_i|T_i,S,N)$$

Similarly, the likelihood $L(S,N|T_i)$ of the species tree S and parameters of duplication and loss N given the gene tree T_i is:

$$L(S,N|T_i) = K_2 \times P(T_i|S,N)$$

In consequence

$$\begin{aligned}
L(T,S,N|A) &= \prod_i L(T_i|A_i) \times L(S,N|T_i) \\
L(T,S,N|A) &= K_3 \times \prod_i P(A_i|T_i,S,N) \times P(T_i|S,N) \\
L(T,S,N|A) &= K_3 \times \prod_i P(A_i|T_i,S,N) \times P(T_i,S,N)/P(S,N) \\
L(T,S,N|A) &= K_3 \times \prod_i P(A_i,T_i,S,N)/P(S,N) \\
L(T,S,N|A) &= K_3 \times \prod_i P(T_i,S,N|A_i) \times P(A_i)/P(S,N)
\end{aligned}$$

Since A_i is constant and if we assume a uniform prior on S,N, we get

$$\begin{aligned}
L(T,S,N|A) &= K_4 \prod_i \times P(S,N,T_i|A_i) \\
L(T,S,N|A) &= K_4 \times P(S,N,T|A)
\end{aligned}$$

with K_4 a constant, and since we assume that all gene families evolve independently.

Consequently, one can view $L(T,S,N|A)$ as the likelihood of our hierarchical model, or as an unnormalized posterior probability with uniform probabilities on the species tree, on the parameters of duplications and losses, and on the parameters of the models of sequence evolution.

2 Algorithms to approximate $L(S, N|T_i)$

We use a reconciliation algorithm derived from Zmasek and Eddy (2001) (Fig. S12), in which nodes of the gene tree are mapped onto nodes of the species tree according to the following principle. For a node n of the gene tree, $L(n)$ denotes the set of species that have a gene that is a descendant of n, and for a set of species S, $lca(S)$ denotes the node that is the last common ancestor of all species in S. Then a node n in the gene tree is mapped to $\Lambda(n) = lca(L(n))$, and we associate to each node n of the gene tree $\Lambda(n)$. Moreover, we augment the gene tree graph with "hidden nodes", *i.e.* nodes that are not visible in the gene tree because one of their two descendants has been lost, and we also map them to nodes of the species tree. We also map these lost descendants, and augment the gene tree with special "lost nodes".

We use this mapping to aid us in the computation of an approximate likelihood of the species tree and parameters of duplication and loss given the gene tree, $L(S, N|T_i)$.

To each node u of the species tree, let $\mathcal{T}(u)$ be the subforest of the gene tree which is the graph induced by all nodes mapped to u by the function Λ. For a component T of $\mathcal{T}(u)$, a vertex is an *exemplar* if it is a leaf, or an internal node of degree 2, or the root if it has degree 1. The number of exemplars k_T of a component T of $\mathcal{T}(u)$ gives the number of paralogous genes obtained by the duplication/loss process along branch i leading to node u of the species tree. If T is a single node, $k_T = 1$, unless it corresponds to a "lost node", in which case $k_T = 0$.

We then take for granted these inferred numbers k_T of paralogous genes after a duplication/loss process witnessed by all the exemplars of each component of $\mathcal{T}(u)$. We use our branch-wise birth-death model to compute the probability $P_{u,1\to k_T}$ that a component T of $\mathcal{T}(u)$ has k_T exemplars. The adequate formula can be found in Bartholomay (1958), equation 69:

$$P_{u,1\to k_T} = \sum_{n=0}^{min(1,k_T)} (-1)^n \binom{1}{n} \lambda^{k_T-n}\mu^{1-n}(e^{(\lambda-\mu)}-1)^{1+k_T-2n}(\lambda e^{(\lambda-\mu)}-\mu)^{-1-k_T+n}(\mu e^{(\lambda-\mu)}-\lambda)^n \tag{1}$$

Using these branch probabilities, and assuming the evolution of a gene along a branch is independent of its evolution along another branch, one can compute an approximate likelihood of the species tree and parameters of duplication and loss given the gene tree, $L(S, N|T_i)$, as follows (Equation 2):

$$L(S, N|T_i) \approx \prod_{u\in\mathcal{N}(\mathrm{S})} \prod_{T\in\mathcal{T}(u)} P_{u,1\to k_T} \text{ with } k_T \in [0; +\infty[\tag{2}$$

where $\mathcal{N}$(S) represents the set of all nodes u of the species tree S. This product is computed over all nodes u of the species tree, and for each node u, over all components T of $\mathcal{T}(u)$ inferred by the algorithm described above. Then, k_T is the number of exemplars in a component T of $\mathcal{T}(u)$.

As an example, for the gene tree and species tree shown in Fig. S12, $L(S, N|T_i)$ is computed as follows:

$$L(S, N|T_i) = P_{1,1\to1} \times P_{2,1\to1} \times P_{5,1\to1} \times P_{4,1\to1} \times P_{3,1\to1} \times P_{6,1\to0} \times P_{7,1\to2} \tag{3}$$

This formula approximates $L(S, N|T_i)$ because we assume that gene family evolution along one branch of the species tree is independent from gene family evolution along another branch of the species tree. This assumption leads us to neglect some of the least parsimonious scenarios, and to view cases where orthologous genes have been independently lost in two sister lineages as a single loss in the parent lineage.

3 Estimating the branch-wise expected numbers of duplications and losses

Approximating the likelihood $L(S, N|T_i)$ requires estimating values for the branch-wise expected numbers of gene duplications and losses. This estimation benefits from the assumption that gene family evolution along a branch of the species tree is independent from other branches. Finding the most likely λ and μ for a branch leading to node u of the species tree therefore only requires considering the counts of exemplars associated to this branch in the most parsimonious reconciliation through an analytical solution, which has the merit of being much faster than numerical optimization. We consider all counts of times where there were k exemplars at the end of branch leading to node u, with $k \in [0; \infty[$; more precisely, we sum these counts in three bins: we use the numbers of times 0, 1 and ≥ 2 lineages have been found at the end of the branch leading to node u. Using these counts only, maximum likelihood values of λ and μ can be computed as shown below.

Equation 1 provides formulas for computing the probability $P_{u,1\rightarrow k}($ of having k exemplars through a birth-death process with birth expected number λ and death expected number μ.

For simplicity, we set:

$$\begin{aligned} A &= (e^{(\lambda-\mu)} - 1)^{1+k-2n} \\ B &= (\lambda e^{(\lambda-\mu)} - \mu)^{-1-k+n} \\ C &= (\mu e^{(\lambda-\mu)} - \lambda)^{n} \end{aligned}$$

Then we have the probability $P_{1,0}(1)$ of having 0 lineages at the end of a branch:

$$P_{1,0}(1) = \mu \frac{A}{B} = X$$

The probability $P_{1,1}(1)$ of having 1 lineages at the end of a branch:

$$P_{1,1}(1) = \lambda\mu \frac{A}{B}^2 - \frac{C}{B} = \frac{\lambda A}{B} X - \frac{C}{B} = Y$$

The probability $P_{1,\geq 2}(1)$ of having 2 or more lineages at the end of a branch:

$$P_{1,\geq 2}(1) = 1 - (P_{1,0}(1) + P_{1,1}(1)) = \frac{Y}{B/(\lambda A) - 1} = Z$$

We therefore have a set of three equations from which one can compute the values of λ and μ. Solving these equations yields:

$$\mu = \frac{ln(X-1) - ln(\frac{Z}{Y+Z} - 1)}{\frac{Z}{X(Y+Z)} - 1} \quad (4)$$

$$\lambda = \frac{Z}{X(Y+Z)}\mu \quad (5)$$

We thus use these two equations to obtain maximum likelihood values for the expected numbers of duplications λ and of losses μ for a given branch leading to node u of the species tree, using the frequencies of cases where 0, 1, or 2 or more exemplars were associated to node u in the gene trees.

4 Algorithms to simultaneously infer species and gene trees

4.1 Maximizing $L(S, N|T_i)$: rooting the gene tree

For a given rooted binary gene tree and a given rooted binary species tree, section 2 shows how to compute $L(S, N|T_i)$. However, gene trees are usually not rooted, unless they are inferred through molecular clock models Kumar (2005) or through non-reversible models of evolution Yang and Roberts (1995), Galtier and Gouy (1998), Huelsenbeck et al. (2002), Yap and Speed (2005), Boussau and Gouy (2006). And even in such cases, there may be little signal for the position of the root. Our procedure therefore evaluates $L(S, N|T_i)$ for all possible roots of the gene trees, and thus provides the most probable root position, according to the birth-death model of gene duplications and losses.

To compute the likelihood for all possible roots of the gene trees, instead of iterating the gene tree traversal algorithm described in section 2 for all n possible positions of the root, which would involve n gene tree traversals in total, we rely on a double-recursive tree traversal algorithm, as in Chen et al. (2000), and thus only require 2 tree traversals. However, contrary to Chen et al. (2000) who used this algorithm to find the most parsimonious reconciliation between an unrooted gene tree and a rooted species tree, we use the algorithm to find the most probable reconciliation between an unrooted gene tree and a rooted species tree (see section 2).

4.2 Finding the best gene tree given a species tree: optimizing $L(G_i)$

The preceding section explained how a particular unrooted gene tree and a species tree can be used to maximize $L(S, N|T_i)$. If gene trees could be known *a priori*, this species tree likelihood would be enough to find the species tree that maximizes the product of all species tree likelihoods for each family. However, gene trees are not given but are usually estimated based on a sequence alignment, through maximization of Felsenstein (1981) likelihood. We therefore search for the gene tree that maximizes $L(G_i)$ according to equation 2 in the main manuscript, which accounts for both sequence and gene family evolution. This search is done through commonly used tree search heuristics; for the sake of computational efficiency, we first use a simple Nearest Neighbor Interchange (NNI) Guindon and Gascuel (2003) strategy, as follows:

1. For each branch of the current gene tree topology t where a duplication has been inferred, the topologies t_1 and t_2 are obtained through NNIs. We thus obtain $L(S, N|t_1)$ and $L(S, N|t_2)$.
2. If $L(S, N|t_1) > L(S, N|t)$ or $L(S, N|t_2) > L(S, N|t)$, we compute $L(t_x|A_i)$, where t_x corresponds to t_1 or t_2.
3. If $L(G_i)$ obtained with t_x is better than the current $L(G_i)$, the NNI is accepted and the algorithm resumes with t_x as the new current gene tree topology.

Step 1 in this algorithm implicitly assumes that the starting gene tree is the most likely according to $L(t_x|A_i)$, as it only computes $L(S, N|t_x)$, $x \in (1, 2)$ which amounts to considering that only $L(S, N|t_x)$ can increase. In practice, starting trees can be obtained by PhyML Guindon and Gascuel (2003) for instance, or are built by PHYLDOG with a BIONJ (Gascuel 1997) or a PhyML-like NNI-based algorithm if not provided.

In addition to this NNI-based algorithm, we have also implemented a SPR-based algorithm, which is used only for fixed species trees. This SPR-based algorithm works as follows, using a user-defined threshold d (in practice, on the Ensembl data set, we used $d = 8$):

1. For each node of the current gene tree topology t, generate all topologies t_x that can be reached by pruning the node and regrafting it at a distance inferior to a pre-defined threshold d. Only nodes that may decrease the number of losses or duplications are tried as regrafting points.
2. For each t_x, compute $L(S, N|t_x)$
3. Let $t*$ be the t_x gene tree topology that maximizes $L(S, N|t_x)$
4. If $L(S, N|t*) > L(S, N|t)$, use $t*$ as the new current gene tree topology.
5. Go back to step 1 until no further improvement can be found. Let's call t the tree found at the end of this algorithm

In practice, it was found that this latter SPR-based algorithm provides more accurate gene trees, and was used to generate the gene trees for the mammalian data set.

4.3 Finding the best species tree given several gene family trees: optimizing $L(T, S, N|A)$

The likelihood defined in eq. 1 in the main manuscript can be used to compute the likelihood of a species tree, parameters of duplication and loss and gene trees given sequence alignments. As the preceding sections have shown how one could find the gene tree that maximizes $L(G_i)$ for a given species tree, what remains to be explained is how to search for the most likely species tree topology. To search for the most likely species tree topology, classical tree exploration algorithms can be used, with the simplification that the species tree does not have branch lengths, and the particularity that it needs to be rooted. We use an algorithm that can be divided in 5 steps.

1. INITIALIZATION
   Start with:

   - a random species tree as current species tree (alternatively a user-input tree can be used or PHYLDOG can reconstruct an MRP tree from families with at most 1 gene per species)
   - gene trees maximizing $L(T_i|A_i)$ as reconstructed by PHYLDOG (alternatively BIONJ gene trees, or user-input gene trees)
   - default parameters of duplication and loss (alternatively user-input parameters)
   - various user-input options: j the maximum distance for Subtree Prunings And Regraftings (SPRs) in the species tree, parameters of gene-specific models of sequence evolution, etc...

2. SPECIES TREE EXPLORATION WITH FIXED GENE TREES AND PARAMETERS OF DUPLICATION AND LOSS
   Iterate until no more improvement is made:

   - on each node n of the species tree:
     - prune the subtree rooted in n, and regraft the subtree in all possible positions less than j nodes away from the pruning node in the species tree
     - compute the associated species tree scores, with fixed expected numbers of events per branch, and keeping gene tree topologies fixed
     - if one of these SPRs increases the score, keep it as current species tree, otherwise go back to the current species tree
   - On each branch of the current species tree:
     - root the species tree
     - compute the associated species tree score, with fixed expected numbers of events per branch, and keeping gene tree topologies fixed
     - if a rooting improves the score of the species tree, keep it as current species tree, otherwise go back to the current species tree

3. SPECIES TREE EXPLORATION WITH FIXED GENE TREES BUT OPTIMIZING PARAMETERS OF DUPLICATION AND LOSS
   Iterate until no more improvement is made:

   - on each node n of the species tree:
     - prune the subtree rooted in n, and regraft the subtree in all possible positions less than j nodes away from the pruning node in the species tree
     - compute the associated species tree scores, keeping gene tree topologies fixed, and with default parameters of duplication and loss
     - update the parameters of duplication and loss for each topology obtained by one of these SPRs

  - re-compute the associated species tree scores, keeping gene tree topologies fixed, with the updated parameters of duplication and loss
  - if one of these SPRs increases the score, keep it as current species tree, otherwise go back to the current species tree
- On each branch of the current species tree:
  - root the species tree
  - compute the associated species tree score, keeping gene tree topologies fixed, and with default parameters of duplication and loss
  - update the parameters of duplication and loss
  - re-compute the associated species tree score, keeping gene tree topologies fixed, with the updated parameters of duplication and loss
  - if a rooting improves the score of the species tree, keep it as current species tree, otherwise go back to the current species tree

Repeat step 2 above, but this time estimating branch-wise expected numbers of events per branch when computing the likelihood

4. Species tree exploration optimizing both gene trees and parameters of duplication and loss
Iterate until no more improvement is made

- on each node n of the species tree:
  - make a Nearest Neighbor Interchange (NNI) around node n
  - compute the associated species tree score, keeping gene tree topologies fixed, and with default parameters of duplication and loss
  - update the parameters of duplication and loss
  - re-compute the associated species tree score, keeping gene tree topologies fixed, with the updated parameters of duplication and loss
  - if one of these NNIs increases the score, keep it as current species tree, otherwise go back to the current species tree
- On each branch of the current species tree:
  - root the species tree
  - compute the associated species tree score, keeping gene tree topologies fixed, and with default parameters of duplication and loss
  - update the parameters of duplication and loss
  - re-compute the associated species tree score, keeping gene tree topologies fixed, with the updated parameters of duplication and loss
  - if a rooting improves the score of the species tree, keep it as current species tree, otherwise go back to the current species tree

5. Optimization of the gene trees and parameters of duplication and loss with fixed species tree
   Optimize gene tree topologies using SPRs instead of NNIs, and keeping the species tree fixed.

6. Termination
   Return the species tree, the parameters of duplication and loss and the gene trees that maximize $L(T, S, N|A)$

From step 2 to 5, each evaluation of a species tree gets more and more costly, as more and more parameters are optimized. The early optimizations are therefore crude, to avoid spending computational time on species trees that are far from the optimal species tree, but the final ones are more accurate, with a step of joint species tree/gene tree estimation where both types of trees are optimized through NNIs (step 4), and a final step of SPRs on the gene trees when the species tree is fixed (step 5).

Fig. S4 shows computing time for different steps of the algorithm for the simulations of 500 gene families along a phylogeny of 12 species. The first box corresponds to step 1, in which the species tree exploration is done without optimization of the gene trees and parameters of duplication and loss, takes about half of the computing time. The second box corresponds to steps 2 to 4, and takes about a third of the time. The third box corresponds to step 5, the optimization of the gene trees using SPRs with a fixed species tree, and takes most of the remaining amount of time.

5 Parallel architecture based on a server-client framework

The complete algorithm to estimate the species tree simultaneously with gene trees is summarized in the following pseudo-code.

Algorithm 1 Optimizing the species tree as well as gene trees with a server-client architecture as implemented in PHYLDOG

```
likelihood_threshold=1e-6
$|T|$=2
if (server) {
    get initial species tree (or build a random one) and store it into $currentS$ and into $oldS$
    get the set of gene families to analyze
    create $n$ clients
    send each client the set of gene families they are in charge of
    send each client $currentS$ and arbitrary starting expected numbers of duplication and loss
    currentlk = -INFINITY
    while (iterations_without_improvement < limit) {
        receive family likelihoods $L(G_i)$ and branch-wise counts of exemplars from the clients
        compute total score ($newLk = L(currentS, S, N|A)$)
        if (newLk > currentLk)
            then {$oldS = currentS$ ; iterations_without_improvement = 0 ; }
            else {$currentS = oldS$ ; iterations_without_improvement ++ ; }
        change the topology of $currentS$ and update expected numbers of duplication and loss
        send each client $currentS$ and expected numbers of duplication and loss
    }
}
else if client {
    receive set of gene families
    receive $currentS$ and rates of duplication and loss
    read alignments and compute or read pre-computed trees that maximize $L(T_i|A_i)$ for each gene family
    while (iterations_without_improvement < limit) {
        compute family scores
        send to the server family scores and branch-wise counts of gene duplications and losses
        receive $currentS$ and rates of duplication and loss
    }
}
```

6 Simulation procedure

6.1 Simulation of 2000 gene families in 40 species

We simulated gene family alignments according to the following procedure.

First, we used a rooted clock-like species tree containing 40 species (Supplementary Fig. S2). For each of the branches of this species tree, we independently drew an average duplication rate from an exponential distribution of mean 0.7, and an average loss rate from an exponential distribution of mean 0.8. Then we simulated gene tree topologies using the species tree topology and the average duplication and loss rates above, using algorithms as in Akerborg et al. (2009), assuming half of the gene families started at the root of the tree, and the other half started uniformly over any other branch of the species tree. To achieve more realism, for each each gene tree the average duplication and loss rates were multiplied by factors drawn from a Gamma distribution with shape and rate parameters of 10, ensuring a spread between approximately 0.15 and 3. Each branch length of each gene tree topology was then randomly drawn from an exponential distribution of mean 0.04269, a value obtained from supplementary figure 7 in Rasmussen and Kellis (2007). We discarded gene family trees that contained fewer than 3 sequences, or more than 100 sequences, and obtained 2000 rooted gene family trees. We then used these gene trees to simulate 1000 bases long sequence alignments under an HKY model of sequence evolution Hasegawa et al. (1985), with parameters ($\kappa = 4$, $\theta = 0.6$, $\theta 1 = 0.6$, $\theta 2 = 0.6$), and with a discretized gamma law ($\alpha = 1$, 4 categories) and a category of invariant sites (proportion 0.1) to simulate rate heterogeneity among sites.

We used these 2000 sequence alignments alone as input to PHYLDOG, which reconstructed the species tree, gene trees, and branch-wise expected numbers of duplications and losses. Our approach emulates the situation with real data where the true evolutionary process is unknown and more complex than the models we use for reconstruction, as PHYLDOG was run with a simple JC69 model Jukes and Cantor (1969) without any model of rate heterogeneity, and does not model heterogeneities in rates of duplications and losses between gene families. Running PHYLDOG on the resulting 2000 gene alignments using 400 processes took 33 hours.

6.2 24 simulations of 500 gene families in 12 species with incomplete lineage sorting

We downloaded simulated gene trees produced in Rasmussen and Kellis (2012) with a model of gene duplication and loss as well as a model of lineage sorting in populations through the multispecies coalescent. The resulting trees have been produced from a *Drosophila* species phylogeny containing 12 species. 24 distinct simulations obtained with a range of duplication rates and a range of effective population sizes were used. Duplication rates were 1, 2, or 4 times the rate observed in the actual Drosophila genomic sequences. Effective population sizes ranged between 1 million (about the effective population size for *Drosophila melanogaster* (Rasmussen and Kellis 2012)) and 500 million. For comparison, the effective population size in humans is thought to be about 10,000 (Nei and Graur 1984), which makes these simulated data sets particularly rich with instances of incomplete lineage sorting, where gene trees differ from the species tree without duplication events.

We used these gene trees to simulate sequence alignments of codons. Branch lengths in these trees are expressed in numbers of generation. To simulate sequences based on these trees, it is necessary to convert the branch lengths into expected numbers of substitutions per codon. To this end, we multiplied the branch lengths by an estimate of the mutation rate per site per generation in *Drosophila melanogaster*, 0.0346 mutation/site/MY (Keightley et al. 2009), and divided by 3. The use of the mutation rate instead of the substitution rate likely results in an overestimate of the number of substitutions per site. We used the Yang and Nielsen model of codon evolution (Yang and Nielsen 1998), with a kappa value of 2 and an omega value of 0.05, determined as the average dN/dS over all functional categories in Supplementary Table 12 from Drosophila 12 Genomes Consortium et al. (2007). Alignments were 600 codons long, based on the average gene length of 1802.9 bases in the version 5.39 of the genome of *Drosophila melanogaster*. We also used a Gamma distribution discretized in 8 categories (Yang 1994) plus a category of invariant sites to model rate heterogeneity across sites.

We used each of these 24 sets of 500 sequence alignments as input to PHYLDOG, which reconstructed the species tree, gene trees, and branch-wise expected numbers of duplications and losses. As with the first simulation, our approach emulates the situation with real data where the true evolutionary process is unknown and more complex than the models we use for reconstruction, as PHYLDOG was run with a simple JC69 model Jukes and Cantor (1969) without any model of rate heterogeneity.

7 Use of PHYLDOG on 6966 gene families from 36 mammalian species

Gene families were downloaded from the Ensembl Compara database v. 57. Families containing more than one sequence from the 36 mammalian species under study were considered. Families exceeding 100 sequences, containing at most 2 sequences, or with alignments smaller than 500 bases were discarded. This resulted in 6966 gene families input into PHYLDOG for joint reconstruction of species and gene trees.

8 Correcting for incomplete genome coverage

To correct for artifactual losses due to low sequencing coverage (Table S1) and unannotated genes, we altered branch-wise estimates of gene losses in terminal branches. Assuming the average mammalian genome contains the same number of genes N as in *Homo sapiens*, we reasoned that a genome g with $n \leq N$ genes according to the Ensembl database must have $N - n$ genes missing in its genome assembly. We therefore derived the expected number of losses e expected in g only because of sequencing quality, and added this value e to the average estimate provided by equation 4 for all branches on which no correction is applied. In effect, this decreases the weight of a loss along the terminal branch leading to a low-coverage genome, and diminishes the "poorly sequenced genome" artifact by which low-coverage genomes tend to be grouped together just because they share artifactual gene losses.

9 Estimating gene tree reconstruction accuracy

Gene trees reconstructed by TreeBeST were obtained from Ensembl v57 Flicek et al. (2010). We reconstructed gene trees using PhyML Guindon et al. (2010) with a GTR+4GI model of sequence evolution. Trees reconstructed by PhyML were reconciled according to the parsimony criterion using the algorithm from Zmasek and Eddy (2001).

Supplementary Fig. S14 shows how reconstructed ancestral syntenies can reveal errors in reconstructed gene trees. This figure represents a species with four species 1 to 4, along which 5 gene families have evolved without events of duplication or loss. These gene families are adjacent and in the same order in the four genomes under consideration. To reconstruct the ancestral chromosomal arrangement of these gene families in the species 7 ancestral to all four species, we use the gene trees reconstructed for all 5 gene families. If there has been an error during the reconstruction of the tree for gene family C, then a duplication is erroneously inferred in the history of this gene family. Because of this duplication event, two genes from the gene family C are inferred to be present in the ancestral species 7. This creates an ambiguity in the ancestral chromosomal arrangement, as two genes from gene family 7 could be placed at the same position in the ancestral chromosome. This ambiguity creates a "conflicting adjacency". We use the number of conflicting adjacencies in ancestral chromosomes as a measure of the number of errors in reconstructed gene trees.

More formally, we estimate gene tree accuracy using ancestral synteny as follows. For one ancestor A, the gene content is computed in the following way. Define G_A as a graph whose vertices are the extant genes of species deriving from A. Two vertices are linked by an edge if the two genes are in the same family, and either:

- their common ancestor in the tree is a speciation node mapped to A or a descendant of A, or
- their common ancestor in the tree is a duplication node mapped to a strict descendent of A.

Connected components of G_A are the genes of A. The vertices in a component C are the descendants of the gene C. This method has the advantage of dealing with any reconciliation as input, and not only a LCA reconciliation. In the case of a LCA reconciliation, it corresponds to the parsimonious count of the number of genes in each ancestral species.

We say that two genes are “adjacent” in an extant or ancestral genome if there is no other gene located between the two in their linear arrangement along a chromosome.

We know extant adjacencies from the observed positions of the genes. Ancestral adjacencies are computed by a Dollo-style parsimony: if two adjacencies may derive from an ancestral one, then we infer it. More precisely, the algorithm searches for all quadruples G,H,I,J of extant genes, such that:

- G and H are adjacent in the genome of an extant species S_1
- I and J are adjacent in the genome of an extant species S_2 different from S_1

- G and I are orthologous and their LCA named A_1 belongs to an ancestral species S_A
- H and J are orthologous and their LCA named A_2, distinct from A_1, belongs to S_A

For each such quadruple G,H,I,J, we draw adjacencies between all pairs of ancestral genes X and Y such that:

- $X = A_1$ and $Y = A_2$, or
- X and Y belong to the same ancestral species S_B, which is a descendant of S_A, and X is an ancestor of G, and Y is an ancestor of H, or
- X and Y belong to the same ancestral species S_B, which is a descendant of S_A, and X is an ancestor of I, and Y is an ancestor of J.

We do not pretend this is the most efficient way of constructing ancestral chromosomes, but it is a good way to assess gene tree quality. Compared to more sophisticated methods (Ouangraoua et al, 2011, Muffato et al, to appear), its advantages are that it is simple to describe, it takes reconciled trees as input without filtering step, and it uses a strict definition of adjacencies, contrary to all other available methods. It is theoretically well suited to our needs, as in the absence of convergent gene order evolution or errors in the trees, the ancestral genes should be arranged in a linear order: no gene should be adjacent to more than two other genes. We expect the effects of convergent gene order evolution to be low and equally confusing for all sets of gene trees. So the "linearity" of the ancestral chromosomes, measured as the number of non conflicting adjacencies, is a good indicator for comparing the quality of two sets of gene trees.

10 Comparison of gene tree accuracy between PhyML, TreeBeST, and PHYLDOG

Some genes present in the Ensembl trees could not be found in the Ensembl alignments. To avoid an unfair comparison between TreeBeST trees and trees reconstructed by PhyML and PHYLDOG, we used only those families that contain exactly the same mammalian genes in all three gene tree sets, resulting in 5039 gene families.

Figs. S9, S8, S10 show numbers of non-conflicting adjacencies in ancestral genomes, ancestral gene contents, and non-conflicting adjacencies normalized by gene contents, reconstructed from gene trees built with PhyML, TreeBeST, and PHYLDOG. We show two results for TreeBeST: one (TreeBeST full reconciliation) in which we have reconciled gene trees according to the Last Common Ancestor (LCA) reconciliation method Zmasek and Eddy (2001), and the other result (TreeBeST) uses reconciliations given in the Ensembl-Compara database (the one presented in the main text). The latter annotate gene duplications only if they are ancestral to two subtrees containing equivalent numbers of species notably. The inclusion of these two results allows us to assess TreeBeST gene tree quality alone with respect to other algorithms ("TreeBeST full reconciliation" result), or in conjunction with the TreeBeST reconciliation algorithm ("TreeBeST" result).

We find that PHYLDOG significantly improves upon TreeBeST in terms of non-conflicting adjacencies (Fig. S9), normalized non-conflicting adjacencies (Fig. S10), and gene contents (Fig. S8). In none of these metrics PHYLDOG is found to be less efficient than other approaches to reconstruct and reconcile gene trees, which means that PHYLDOG always improves upon competing approaches.

11 Measuring alignment quality

We measured alignment quality as the number of sites kept after filtering by Gblocks (Castresana 2000) divided by the total number of sites in the alignment. We ran gblocks using the following options: $-t = d$ $-b1 = numSeqs/2{+}1$ $-b2 = numSeqs/2 + 1$ $-b4 = 2$ $-b3 = 0$ $-b5 = a$, where $numSeqs$ is the number of sequences in the alignment.

12 Numbers of duplications and losses across the mammalian phylogeny

Fig. S15 shows that numbers of duplications and losses per gene family are correlated with each other, and are correlated with the number of genes in the gene families.

Overall losses are more frequent than duplications (Fig. S15, Fig. S16). Fig. S16 suggests that losses behave almost as if they occurred regularly along the tree. As expected, genomes with low sequencing coverage display very high numbers of loss. To reduce this effect, we selected only therian genomes with coverage above 5x. Fig. S17 shows that gene losses are more frequent than gene duplications even among these high-coverage genomes. We find no correlation between the branch-wise numbers of duplications and losses per gene (Spearman rho p-value=0.712). Notably some branches show strong opposite patterns for duplications and losses, such as the branch leading to primates, which is predicted to have undergone many losses and a very small number of duplications.

Overall our inferences open interesting avenues for further investigation with a denser taxonomic sampling and high quality genome sequences.

13 Estimation of the impact of Incomplete Lineage Sorting on gene trees

To evaluate the impact of ILS on gene trees and species trees reconstructed by PHYLDOG, we performed simulations and analyzed a well-documented case of closely-spaced cladogenesis events in our mammalian data set.

13.1 Analysis of simulated data

24 different simulations were performed (see sub-section 6.2) with a range of effective population sizes and duplication rates. Notably, population sizes ranged from values observed in *Drosophila* species (effective population size 1 million),

to 500 times this value. PHYLDOG was run on the resulting alignments to reconstruct gene trees and species trees. Results are reported in Fig. S. 3. The comparison of the reconstructed species tree to the true species tree shows that PHYLDOG fails to recover the correct species tree only for the largest population size. A similar trend is observed for two species-tree reconstruction methods, duptree and iGTP run with the duplication-loss score, (Wehe et al. 2008, Chaudhary et al. 2010), even though contrary to PHYLDOG these methods were provided with the correct gene trees. The comparison of the inferred numbers of duplications to the true numbers of duplications observed in the simulated data shows that accuracy is reduced by increasing effective population size. For instance, for an effective population size of 1 million and 3 duplication rates (1x, 2x, 4x), the reconstructed numbers of duplication are 279, 564, 895 respectively, compared to 234, 515, 843, corresponding to an overestimate between 6% and 20%. For an effective population size of 75 millions, the overestimates are between 107% and 547%. Such overestimates are due to an incorrect interpretation of the gene trees reconstructed by PHYLDOG. These gene trees are reconstructed extremely accurately, as the maximum normalized average Robinson and Foulds distance (Robinson and Foulds 1979) between the reconstructed and simulated trees is 0.004. However, in these gene trees, ILS has introduced differences with the species tree that are incorrectly interpreted by PHYLDOG's model as duplications followed by losses. Interestingly, in these simulations, PHYLDOG reconstructs trees that are closer to the gene trees than to the locus trees, in which population-level effects are ignored (Paired Student test p-value=$7.48.10-8$; Wilcoxon paired test p-value=$1.19.10-7$).

Although we used the exact same simulations as Rasmussen and Kellis (2012), and therefore can compare PHYLDOG's results to the results of the DLCoalRecon method, it is unfair to expect that PHYLDOG should perform as well as DLCoalRecon. Indeed, PHYLDOG does not model coalescent-level processes, and reconstructs gene trees from alignments instead of directly using the correct gene trees. With these handicaps, analyzing a data set for which its model is misspecified, it is no surprise that PHYLDOG performs much worse than DLCoalRecon at recovering the correct number of events. For instance, for the simulation with an effective population size of 25 millions and the duplication rate "1X", PHYLDOG finds 811 duplications, DLCoalRecon 242, for 232 simulations in the real data. We suggest that, because PHYLDOG accurately reconstructs the species tree and the gene trees, running DLCoalRecon on trees reconstructed by PHYLDOG may be a valuable strategy to accurately estimate the number of events in data sets where ILS has been important.

13.2 Analysis of real data

The diversification of *Homo sapiens*, *Pan troglodytes* and *Gorilla gorilla* is a well-documented case of closely-spaced cladogenesis events, a situation in which incomplete lineage sorting (ILS) is known to occur and has been shown to be severe. Recently Scally et al. (2012) estimated the amount of ILS along the human genome using a Hidden Markov Coalescent model. We downloaded estimates of the amount of ILS per gene in the version 67 of Ensembl from

`http://kimura.univ-montp2.fr/jdutheil/Gorilla/`

, and selected gene families for which names could be matched between our version of Ensembl (57) and the newest version of Ensembl used by Scally et al. (2012) (67), and for which the triplet *Homo sapiens*, *Pan troglodytes* and *Gorilla gorilla* was monophyletic in all three reconstruction methods, i.e PhyML, TreeBeST and PHYLDOG. We obtained 1644 gene families. We compared the amount of ILS in genes for which trees differ from the species tree ((*Homo sapiens*, *Pan troglodytes*), *Gorilla gorilla*) versus genes for which trees agree with the species tree (Fig. S18). Among the 1644 genes, PHYLDOG always recovers a gene tree topology identical to the species tree, TreeBeST recovers the species tree in 1035 gene families, and PhyML in 1056 gene families. For both TreeBeST and PhyML the amount of ILS in gene trees that differ from the species tree is higher than for gene trees that match the species tree (ILS for trees similar to the species tree: 0.17310 and 0.17150 respectively; ILS for trees different from the species tree: 0.31740 and 0.33020 respectively; Student tests and Wilcoxon tests p-values $<$ 2.2e-16). These results indicate that gene trees reconstructed by PHYLDOG are virtually not impacted by ILS, whereas Ensembl trees and PhyML trees are. Consequently in this data-set, PHYLDOG trees are closer to the locus trees defined by Rasmussen and Kellis (2012), whereas Ensembl and PhyML trees are probably closer to what these authors define as gene trees.

13.3 Conclusion: effect of ILS on PHYLDOG's accuracy for reconstructing gene family evolution

Simulation analyses based on 12 Drosophila genomes with a high substitution rate show that PHYLDOG reconstructs gene trees accurately. However, because PHYLDOG's model of gene family evolution only includes duplications and losses, it improperly interprets instances of ILS as events of duplication followed by losses. This misinterpretation of ILS events probably explains the results obtained on the primate data set, where gene trees reconstructed by PHYLDOG seem to be immune to ILS, contrary to results obtained on the simulation. In primates, the number of substitutions supporting one gene tree against another is probably lower than in our simulations with a high substitution rate. In this context, in primates the likelihood of a gene family is maximized by choosing a gene tree that avoids events of duplications and losses that are supported by few substitution events, whereas in our simulations with a high substitution rate the likelihood is maximized by choosing the gene tree supported by the events of substitutions, even though this is penalized by events of duplication and loss.

Expectedly the effect of ILS on the gene trees reconstructed by PHYLDOG depends on the strength of the signal supporting ILS in the sequences. Depending on the strength of the signal, gene trees reconstructed by PHYLDOG fall between the locus trees, if the signal is weak, as in the primate data set, and the gene tree, if the signal is strong, as in the *Drosophila* simulations. Complementing PHYLDOG's model with a coalescent model should provide it with the means to distinguish between events of ILS and events of duplications, as has been done in the DLCoalRecon method (Rasmussen and Kellis 2012).

References and Notes

Akerborg O, Sennblad B, Arvestad L, and Lagergren J. 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences* **106**: 5714–5719.

Bartholomay AF. 1958. On the linear birth and death processes of biology as Markoff chains. *Bulletin of Mathematical Biophysics* **20**: 97–118.

Boussau B and Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Systematic Biology* **55**: 756–768.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **17**: 540–552.

Chaudhary R, Bansal M, Wehe A, Fernandez-Baca D, and Eulenstein O. 2010. igtp: A software package for large-scale gene tree parsimony analysis. *BMC Bioinformatics* **11**: 574.

Chen K, Durand D, and Farach-Colton M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *Journal of computational biology : a journal of computational molecular cell biology* **7**: 429–447.

Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, et al.. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203–218.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**: 368–376.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, et al.. 2010. Ensembl 2011. *Nucleic Acids Research* .

Galtier N and Gouy M. 1998. Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution* **15**: 871–879.

Gascuel O. 1997. Bionj: an improved version of the nj algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* **14**: 685–695.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, and Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**: 307–321.

Guindon S and Gascuel O. 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology* **52**: 696–704.

Hasegawa M, Kishino H, and Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**: 160–174.

Huelsenbeck JP, Bollback JP, and Levine AM. 2002. Inferring the root of a phylogenetic tree. *Systematic Biology* **51**: 32–43.

Jukes TH and Cantor CR. 1969. Evolution of protein molecules. In *Mammalian protein metabolism, III* (ed. HN Munro), pp. 21–132. google.com, New York.

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, and Blaxter ML. 2009. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Research* **19**: 1195–1201.

Kumar S. 2005. Molecular clocks: four decades of evolution. *Nature Reviews Genetics* **6**: 654–662.

Nei M and Graur D. 1984. Extent of protein polymorphism and the neutral mutation theory. *Evol. Biol* **17**: 73–118.

Rasmussen MD and Kellis M. 2007. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Research* **17**: 1932–1942.

Rasmussen MD and Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research* .

Robinson D and Foulds L. 1979. Comparison of weighted labeled trees. In *Isomorphic factorisations VI: Automorphisms, combinatorial mathematics, No. 748 in Lecture Notes in Mathematics* (eds. AF Horadam and WD Wallis), pp. 119–126. Springer, Berlin.

Scally A, Dutheil J, Hillier L, Jordan G, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al.. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**: 169–175. 10.1038/nature10842.

Wehe A, Bansal MS, Burleigh JG, and Eulenstein O. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics (Oxford, England)* **24**: 1540–1541.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39**: 306–314.

Yang Z and Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution* **46**: 409–418.

Yang Z and Roberts D. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Molecular Biology and Evolution* **12**: 451–458.

Yap VB and Speed T. 2005. Rooting a phylogenetic tree with nonreversible substitution models. *BMC Evolutionary Biology* **5**: 2.

Zmasek CM and Eddy SR. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics (Oxford, England)* **17**: 821–828.