Genome evolution-aware gene trees

Method Manuscript

Emmanuel Noutahi*[1], Magali Semeria*[2],
Manuel Lafond[1], Jonathan Seguin[1],
Bastien Boussau[2], Laurent Guéguen[2],
Nadia El-Mabrouk[1,4], Eric Tannier[2,3,4]

1 - Département d'Informatique (DIRO), Université de Montréal, H3C3J7, Canada;
2 - LBBE, UMR CNRS 5558, Université de Lyon 1, F-69622 Villeurbanne, France;
3 - INRIA Grenoble Rhône-Alpes, F-38334 Montbonnot, France
4 - Corresponding authors mabrouk@iro.umontreal.ca, Eric.Tannier@inria.fr
* - equal contribution

Keywords: genome evolution, gene repertoire, genome rearrangements, phylogeny, trees, ancestral genomes

Abstract

A gene family tree is traditionally inferred from a multiple alignment of homologous sequences according to a model of sequence evolution. Trees for several genes families are thus constructed independently from each other. They often carry unresolutions or bad resolutions. Information for their full resolution may lie in the poorly exploited dependency between gene families, each bringing information for the resolution of the others. We propose to use several kinds of such dependencies in the construction of gene trees: information from a species tree through a model of gene content evolution by duplication, speciation and loss, information from extant synteny through ortholog predictions, and information from ancestral synteny through a model of gene neighborhood evolution. We develop several "correction" techniques, yielding a software package called "RefineTree". We report some tests on simulated data and an application on the full set of gene families from the Ensembl database. We perform a genome-wide analysis of duplication and loss patterns on the history of 65 eukaryote species, including ancestral genes and gene orders of all ancestors along this phylogeny. We show that according to several measures including running time, likelihood, stability of genome content and linearity of ancestral chromosomes, trees corrected by RefineTree are arguably more plausible than the ones stored by Ensembl. We discuss the quality criteria in the light of gene definition as a sequence or as a locus. We extract some cases where a "true" gene tree should depend on this definition.

RefineTree web interface is available at: `http://www-ens.iro.umontreal.ca/~adbit/polytomysolver.html`

Introduction

Several gene tree databases from whole genomes are available, including Ensembl Compara (Vilella et al. 2009), Hogenom (Penel et al. 2009), Phog (Datta et al. 2009), MetaPHOrs (Pryszcz et al. 2011), PhylomeDB (Huerta-Cepas et al. 2011), Panther (Mi et al. 2012). However they are known to contain many errors and uncertainties, in particular for unstable families (Boeckmann et al. 2011), which makes them uneasy to use for accurate ancestral genome inference, orthology detection, or the study of genome dynamics.

For example Ensembl Compara trees, when reconciled with a species tree to annotate gene duplication and loss, systematically and unrealistically overestimate the number of genes in ancestral genomes, and lead to erroneous predictions of ancestral chromosome structures (Boussau et al. 2013). It is a known artifact and a significant number of nodes in the Ensembl gene trees are labelled as "dubious" (Flicek et al. 2014).

Reasons for errors in gene trees are numerous. For example they are dependent on annotation, family clustering and alignment errors. But even if we assume a perfect gene family assignment and alignment, trees are usually constructed from a DNA or protein sequence alignment with a model of substitution (cf. e.g. PHYLIP (Felsenstein 1981, 2005), PhyML (Guindon and Gascuel 2003), RAxML (Stamatakis 2006), MrBayes (Ronquist and Huelsenbeck 2003), PhyloBayes (Lartillot and Philippe 2004)). These models make simplifying assumptions, inducing known systematic artifacts, and algorithms may not properly explore the solution space, but above all, gene sequences often do not contain enough substitutions to resolve all the branches of a phylogeny, or alternatively too many substitutions such that the substitution history is saturated. Therefore trees in databases are usually accompanied with measures of statistical support on their branches, which gives a measure of how confident the inference method is on his choice of the best tree.

It is possible to choose among many statistically equivalent trees, or sharpen the *a posteriori* distribution by modeling gains and losses of genes, inferred by reconciliation of gene and species trees (Szöllősi et al. 2015). Several methods integrate species tree information with sequence information, including TreeBeST (Schreiber et al. 2013), TreeFix (Wu et al. 2013), BBCA (Zimmermann et al. 2014), PhylDog (Boussau et al. 2013), ODT (Szöllősi et al. 2012), ALE (Szöllősi et al. 2013), GSR (Akerborg et al. 2009; Arvestad et al. 2004), SPIMAP (Rasmussen and Kellis 2011), Giga (Thomas 2010), Notung (Durand et al. 2006), MowgliNNI (Nguyen et al. 2013). They all report better gene tree constructions, but leave a large space for improvement or for scaling to genome-wide studies on a large number of species. They also fail to use other information from whole genome evolution such as synteny and chromosome organization. Synteny is often seen as one of the best ways to predict orthology (Jun et al. 2009), a task that is theoretically contained in the phylogeny problem. But, probably due to the complexity of including this information in an evolutionary

model, this has never been really exploited in a phylogenetic context (with the exception of a few pioneering studies applied on yeast genomes as in Wapinski et al. (2007)).

In all sequence-based or integrative methods, tree space exploration is based on local exploration. Moves in gene trees are proposed, and accepted or rejected according to hill-climbing, Metropolis-like criteria, or other statistical or empirical arguments. Moves are proposed at random (typically NNI, SPR, TBR on random branches), or are designed to have a greater chance to be accepted, which is the purpose of gene tree correction methods (Durand et al. 2006; Chen et al. 2000; Gorecki and Eulenstein 2011b,a; Chaudhary et al. 2011; Berglund-Sonnhammer et al. 2006; Doroftei and El-Mabrouk 2011; Swenson et al. 2012; Lafond et al. 2012, 2013; Chauve et al. 2013; Bansal et al. 2014). The construction of a whole genome database using random local search alone is computationally intensive and expected not to scale well as databases grow in size. On the other hand, correction techniques are not usually integrated in an exploration framework. Consequently, database construction pipelines such as TreeBeST (contructing the Ensembl Compara gene trees) have to adopt compromises, exploring limited subsets of tree spaces.

In this article we report the development, improvement, implementation, publicization with a web interface and systematic use of some gene tree correction techniques, providing and exploiting an alternative set of gene trees for the Ensembl Compara gene families. From a starting tree with branch supports constructed independently, we propose the following corrections: (1) contract unsupported branches, and apply ProfileNJ, an extension of the algorithm by Lafond et al. (2012) to solve polytomies according to a species tree, minimizing the cumulative cost of duplications and losses, following Neighbor-Joining principles to choose among the numerous optimal solutions; (2) construct a set of putative orthologs from synteny blocks between genomes (obtained with PhylDiag (Lucas et al. 2014)), and apply ParalogyCorrector (Lafond et al. 2013); (3) construct ancestral chromosomes with DeCo (Bérard et al. 2012), and correct some trees which are responsible for a non linear chromosome structure as in Chauve et al. (2013). Thus we use a range of information in the construction of gene trees, taking into account gene sequence evolution, gene content evolution and chromosome structure evolution.

This set of techniques is a mix of published algorithms, with inedite improvements and generalizations, put together here in a modular piece of software called "RefineTree". ProfileNJ is the main methodological development that we report here. It can be seen as a phylogenetic interpretation of phyletic profiles, that have often been used for molecular evolutionary studies, in the absence of trustable gene trees (Clark et al. 2007; Csurös 2010; Cohen et al. 2012). It also generalizes Neighbor-Joining and the multifurcated gene tree reconciliation with a species tree in a duplication and loss context. We tested ProfileNJ on simulations, and

showed that it achieves results of comparable quality but is several times faster than TreeFix (Wu et al. 2013), which has roughly the same objectives with a random exploration.

We provide the result of a successive application of several modules of RefineTree on the whole set of gene families from the Ensembl database, with PhyML maximum likelihood starting trees. We evaluate the results according to several criteria: (1) likelihood ratio based on the Ensembl alignments between our results and the ones stored at Ensembl; (2) ancestral genome sizes based on a duplication-loss reconciliation; (3) linearity of ancestral chromosomal segments computed with DeCo. RefineTree compares very favorably with the trees stored in Ensembl, and its running time allows it to propose trees for the whole database in a few hours on a desktop computer (not including the starting tree construction).

We make the set of trees and ancestral genomes accessible. We also use the reconstructed trees and ancestral genomes to study genome evolution across all 69 eukaryotic species from the Ensembl database. We provide in particular a whole genome analysis of duplication patterns, pointing at certain branches which seem to show acceleration of duplication or loss processes. We finally discuss the distance to "true" gene trees, in the light of incomplete lineage sorting and gene conversion.

Results

ProfileNJ

Description

The main methodological development is ProfileNJ, a correction technique that can be viewed as a generalization of three different algorithms designed for evolutionary studies:

- The standard Neighbor-Joining (Saitou and Nei 1987) (NJ) method which constructs a tree from a distance matrix between taxa;
- The phyletic profile method which infers ancestral genes from the number of homologous copies in extant taxa, along with duplication and loss costs (Csurös 2010);
- The most parsimonious reconciliation between a multifurcated rooted gene tree and a species tree (Lafond et al. 2012).

These methods have little in common as they solve different problems. But here they are seen as having the same objective: constructing a gene tree. Although phyletic profile methods usually do not explicitly

output a tree, such a tree is implicitly inferred from the gain and loss events in a gene family. The goal of ProfileNJ is to integrate information from the species tree (like in phyletic profiles), from well supported branches of a gene tree (like in reconciliations), and from a distance matrix (like NJ), to complement the information when it is missing.

A run of ProfileNJ is summarized in Figure 1. The input of the method is:

- a rooted species tree S;
- a rooted or unrooted possibly multifurcated gene tree G (which can be obtained by constructing a gene tree from a sequence alignment and contracting unsupported branches);
- a distance matrix d on the leaves of G;
- a weight for duplication and loss events (the default is equal weight for both events).

A solution of ProfileNJ is a set of rooted binary gene trees containing all the internal branches of G and minimizing the total weight of duplications and losses resulting from a reconciliation with the species tree. Among all trees verifying an optimal duplication/loss count, if there are several of them, the NJ selection criterion finds one according to the distance matrix. Several trees can nonetheless be output because of the multiplicity of solutions to the duplication/loss count. For each of them, NJ is used to construct the associated tree. The multiplicity of propositions can be seen as an exploration tool, solutions can then be chosen according to different criteria like the likelihood.

Note that if G is a star tree with genes from only one species, ProfileNJ reduces to NJ. If G is a star tree and no distance matrix is given, ProfileNJ gives a "profile tree", minimizing the weight of a duplication and loss scenario. If G is binary but not rooted (in that case d is useless and not required), ProfileNJ can be used to root the tree according to duplication and loss scenarios. ProfileNJ can also be used to reconcile a rooted binary gene tree with a species tree. So ProfileNJ is a phylogenetic tool that generalizes several usually unrelated standard methods.

Efficiency of the NJ criterion

ProfileNJ is an extension of the algorithm by Lafond et al. (2012). The added features are generalizations to unrooted gene trees, arbitrary weights of duplications and losses, and choosing among binary resolutions with the same duplication/loss costs with the NJ criterion.

In order to evaluate the relevance of the NJ criterion, we ran ProfileNJ twice on the same data sets, except that once the distance matrix was computed, using the Ensembl nucleotide alignments, with FastDist

from the FastPhylo package (Khan et al. 2013), and once the distance matrix was random. The starting tree was computed for every family using PhyML on the nucleic alignments, and all branches with aLRT support < 0.95 were contracted. In average 55% of the branches were contracted. A histogram of the full distribution is shown on Figure S1. The species tree is taken from Ensembl.

Then we computed the likelihood of both trees for every family with PhyML. Among the trees for which the likelihood was different (55% of all tested trees), 76% were in favor of the trees built with the FastDist distance matrix, and the log likelihood differences were much larger for those trees, contributing 95% of the total of log likelihood differences.

The comparisons are clearly in favour of the NJ criterion over no criterion at all, while quantitatively there remains a small but non negligible part of the trees for which no criterion (the random distance matrix) gives an unexplained slightly but significantly better likelihood.

Efficiency of the ProfileNJ tree space exploration strategy

ProfileNJ can be used as a tree space exploration tool for the purpose of gene tree correction: given an initial tree with branch support —which can be retrieved from a gene tree database, or alternatively produced with a phylogenetic reconstruction tool— weakly supported branches can be contracted, leading to a multifurcated tree. Then all resolutions of this tree can be explored with ProfileNJ. Remember that there are two reasons for multiplicities of solutions: first because there are several duplication/loss count solutions, second because for one solution, there are several possible trees. The latter is handled by NJ, while the former is used as an exploration and trees are tested and chosen according to their likelihood.

Other tree space search strategies have been proposed for phylogenetic reconstruction, most of them based on random exploration of a tree neighborhood. The most common strategy is to select, in the space of trees obtained from the original one by performing some branch moves (NNI, SPR, TBR), the one best fitting the species tree in terms of reconciliation cost. In this class of algorithms, NOTUNG (Chen et al. 2000) and TreeFix (Wu et al. 2013) are the most closely related to ProfileNJ, with TreeFix being the most recent one. TreeFix generates a tree neighborhood from NNI and SPR moves and explores it randomly using a hill-climbing strategy. Instead we take a deterministic and more targeted approach by focusing on weakly supported branches of the tree, with a possibly deep modification of the tree. The comparison with TreeFix is intended to compare these two tree space exploration methods.

Wu et al. (2013) have compared TreeFix with SPIMAP (Rasmussen and Kellis 2011), showing a similar accuracy and a higher speed for TreeFix. We perform a similar comparison on the same simulated dataset

of 16 fungi. This dataset consists of simulated gene families generated under the SPIMAP model and their corresponding nucleotide alignments, for four different rates of duplication and loss (DL) events: ($1r_D$ - $1r_L$), ($2r_D$ - $2r_L$), ($4r_D$ - $4r_L$) and ($4r_D$ - $1r_L$); where r_D and r_L are respectively the estimated duplication and loss rates for fungi. Comparisons reported in this section are performed on 2575 simulated gene families randomly chosen from the four fungi datasets with different DL rates.

An initial maximum likelihood (ML) tree is constructed for each simulated gene family with RAxML v-8.1.2 (Stamatakis 2006), with the rapid bootstrap algorithm, under the GTR-Γ model and the majority rule consensus tree as bootstopping criterion. A randomly rooted tree is then provided as input to TreeFix (as TreeFix requires the input tree to be rooted), while a multifurcated unrooted tree obtained by contracting the branches with support lower than 95% is provided as input to ProfileNJ. We used default parameters for both programs. Among the set of resolutions output by ProfileNJ, the best supported tree was selected using consel (Shimodaira and Hasegawa 2001) and site-wise likelihood values computed with RAxML (under the GTR-Γ model of nucleotide substitution).

For RAxML, TreeFix and ProfileNJ trees, we measured the Robinson-Foulds (RF) distance to true trees, compared the reconstructed tree with the true tree using site-wise likelihoods (Figure S7), measured the accuracy of the duplication and loss scenarios (Figure S5), the sensitivity of the accuracy to gene family size (Figure S6), the sensitivity to species tree errors (Figure S8), and the running time.

Figure 2 illustrates the results for the RF distance. It shows that sequence-only does not contain enough signal to lead to the true tree, and integrating information from the species tree is necessary. TreeFix and ProfileNJ reconstruct around 75% of true trees, compared with only 10% for RaxML (RF distances were computed for rooted trees with ProfileNJ and TreeFix, and unrooted trees for RaxML, so RF=0 means good topology and root for ProfileNJ and TreeFix). We investigated some cases where they were erroneous, and found that often, the true scenario was not parsimonious in terms of duplications and losses, while TreeFix and ProfileNJ chose too recent duplications in order to avoid losses. An example is given in supplementary material (Figure S4).

The performance of TreeFix and ProfileNJ are similar in terms of distance to the true tree. If we measure the likelihood of the reconstructed tree, RAxML of course gives the best likelihood. Its likelihood is even usually higher than the likelihood of the true tree, but not significantly according to an AU test. Treefix is designed to produce trees which are not significantly different than the ML tree, which we could check: 1.36% of the trees fail the AU test against the ML tree at $\alpha = 0.05$, while the proportion jumps to 9.17% for ProfileNJ. It is noticeable that this has no visible consequence on the distance to the true tree.

Figure 3 shows that ProfileNJ outperforms TreeFix in running-time. The gap in running-time between the two algorithms increases with tree size. This figure also shows that the most time-consuming step in ProfileNJ is the tree selection with consel. For a tree of size 30, ProfileNJ is about four to seven times faster than TreeFix and about 15 times faster without the tree selection step with consel. This includes the construction of the distance matrix. The construction of the initial RaxML tree is not included because it is common to both methods.

Other analyses, including the sensitivity to gene family size, number of duplications and losses, or errors in the species tree are reported in the supplementary material. They show similar tendencies, TreeFix and ProfileNJ have similar performances on all measures except running time, and RAxML has a lower performance except when there are errors in the species tree.

Indeed we also investigated the impact of the species tree used to reconstruct gene trees. We use an incorrect specie tree (Figure S9 (A)) as input for TreeFix and ProfileNJ. We found that the reconstructed gene trees became less accurate than RAxML gene trees. This impact however was limited to the branches that had been rearranged; the rest of the branches in the TreeFix and ProfileNJ gene trees remained more accurate than RAxML gene trees.

RefineTree

Principle

We integrated ProfileNJ in a modular online software, called RefineTree, combining a number of correction techniques, with an easy-to-use interface (see Figure S2 in supplementary material).

Two additional correction techniques are included, that were previously published by our group. They use information from extant or ancient genome organization. The first one uses PhylDiag (Lucas et al. 2014) (see Methods) to compute statistically supported synteny blocks of genes between every pair of genomes from the Ensembl database. We then assume that if several genes are found consecutive in one genome, and their homologs are also found consecutive in the other genome, the common linear arrangement was in the ancestor and the homologous genes are probably orthologous. This hypothesis is incorrect in at least three cases : (1) if the whole block of genes was duplicated, (2) if there is a tandem duplication of a gene followed by a differential loss in the two species, or (3) if a gene is converted by a paralog. To handle these cases, we require that (1) the majority of the homologous genes are indeed predicted as orthologs by phylogeny, (2) the common ancestor of two homologous genes does not lead to two paralog descendants placed in tandem in one species. In case (3), we are in a situation where the loci are orthologous but not the sequences. In that

case we construct the "locus tree" (Rasmussen and Kellis 2012) and trust syntenic information over gene sequence information. Details of these constrains are given in the Method section. Given couples of putative orthologous genes, we use ParalogyCorrector (Lafond et al. 2013) with the output tree from ProfileNJ as input. This method, integrated in RefineTree, constructs the tree which is the closest to the input (in that case, ProfileNJ) tree according to a Robinson Foulds distance, with the constraint that couples of putative orthologs are found orthologs in a reconciled output phylogeny.

The other correction technique integrates information from the linearity of ancestral genomes computed with DeCo (Bérard et al. 2012) (see Methods). Linearity means that genes are linearly ordered along chromosomes, which is true for extant genomes, but not guaranteed in ancestral genomes computed by DeCo. What could seem as a drawback is used here to detect errors in a gene tree: the "Unduplicator" correction (Lafond et al. 2013) algorithm consists in fusing two ancestral copies of a gene when the two copies disrupt the linearity of an ancestral genome. Details can be found in the methods section or the associated publication.

A typical run of RefineTree, integrating all described correction techniques, is illustrated in Figure 4.

Results on Ensembl gene trees

On the whole Ensembl gene family database (version 73, sept 2013), we compared three sets of trees constructed by a modular use of RefineTree, as in Figure 4.

- **Ensembl trees**: Trees stored in the Ensembl database;
- **ProfileNJ trees**: Output trees from ProfileNJ, with as input PhyML trees (where branches with a < 0.95 aLRT support are contracted) and FastDist distance matrices;
- **Synteny trees**: Output trees of either ParalogyCorrector and Unduplicator (the two are computed and the most likely is chosen) with ProfileNJ trees as input, using PhylDiag and DeCo to construct synteny constraints.

We evaluate the resulting trees with sequence likelihood, ancestral genome content and ancestral chromosome linearity. The results are shown on Figure 5.

The distribution of ancestral gene content sizes is expected to be close to that of extant genomes. Incorrect trees are known to require additional duplications to be reconciled with the species tree, and thus tend to increase the number of genes in ancestral genomes. The linearity of ancestral genomes is expected to be as close as possible to that of the extant genomes as well, with each gene having zero, one or two neighbors,

with a peak at two (the 0s and 1s are due to partially assembled genomes). ProfileNJ trees show a better behaviour than Ensembl trees according to the three measures: more than 2/3 of the trees have a better likelihood than Ensembl trees, the ancestral genome content distribution is much closer to the extant one, and the linearity of chromosomes is higher. So this set of trees, achieving better performance according to sequence evolution, gene content evolution and chromosome evolution, is arguably a better dataset than the one stored in the Ensembl database.

However the content and synteny signals are still distant from what we could expect from true trees. The behavior of synteny trees is interesting from this point of view. Their performance drops in terms of likelihood (Figure 5 (A)), but jumps in terms of the stability of gene content and the linearity of ancestral chromosomes (Figure 5 (B) and (C)). One interpretation is that the synteny corrections, while improving synteny signal, are not yet able to propose reliable gene trees. A reason is that they can break well supported branches to achieve orthology constraints. Branches might be highly misplaced while preserving the ancestral content according to the LCA. We however noticed a correlation between the size of the families (number of genes) and the loss of likelihood in the Synteny trees. Part of the likelihood drop could also be interpreted as an inadequacy of the phylogenetic models to appropriately account for gene families with a high rate of duplications. As observed in our simulations, the true tree is not necessarily the ML tree. Add that likelihood is computed with an alignment which results from a guiding tree which is different from the tested tree. Some synteny trees might therefore be considered as better trees even with these equivocal results.

However there is a third interpretation. Synteny information describes the history of loci (Rasmussen and Kellis 2012), while phylogenetic models describe the evolution of sequences. Loci and sequences often have the same history, but they may differ following gene conversion or incomplete lineage sorting (ILS).

In case of ILS or gene conversion, two different true versions of the gene history are concurrent. In Figure 6) the gene as a locus has a history depicted by the right tree, while the gene as a sequence has a history depicted by the left tree. None of the two are wrong, but they are significantly different. They highlight the ambiguity of the definition of a gene, which yields an ambiguity in its history. Sequence trees will have a high likelihood and mediocre results for gene contents and synteny when constructed from duplication and loss scenarios, while it is the opposite for loci trees. Rasmussen and Kellis (2012) have modeled ILS in sequence and duplications and losses in loci, handling this difference in one case. However, no model is currently able to handle conversion.

Modes of evolution in eukaryotes

With the gene trees we reconstruct all gene contents of ancestral genomes and the way they are organized along ancestral chromosomes. Gene content is computed according to the LCA reconciliation (see Methods), and genome organizations consist in sets of links between consecutive genes. Ancestral genomes are not exactly linearly arranged, but sufficiently close to be often interpreted as chromosomes. We do not linearize them by removing links because the non linearity has diverse causes that we do not wish to mask. But this method also highlights genes or groups of genes evolving together in parts of the tree. For example there are 8488 blocks of co-duplicated genes according to DeCo. Most of them contain only a few number of genes (83% contain 2 genes). The largest blocks are found in the terminal branches leading to *Danio rerio* and *Caenorhabditis elegans*.

As seen in Figure 7, branches of the phylogeny which carry large numbers of duplications are also visible. Patterns of duplications in mammals have been studied by Boussau et al. (2013) with a subset of gene families or in vertebrates by Mahmudi et al. (2013) with a subset of species, but few methods are able to handle whole databases and provide a complete view on the duplication and loss pattern. Figure 7 shows the result for the full genomes of the full phylogeny of the 65 Ensembl species. Branches with a large number of duplications (hot branches) are those leading to vertebrates, which is in agreement with the two rounds of whole genome duplication hypothesis. Interestingly, the speciation event leading to *Petromyzon marinus*, which is usually thought to have diverged after these events (Smith et al. 2013), preceeds the hot branches. This may be in agreement with recent results based on the analysis of Hox clusters in the Japanese lamprey (Mehta et al. 2013). Another hot branch leads to eutherian mammals, which was also found by Boussau et al. (2013) and Mahmudi et al. (2013) with partial data. These two hottest internal branches are exactly the ones found by Mahmudi et al. (2013) using a probabilistic technique, but using only 9 species due to computational cost. Other hot branches are terminal, the hottest being those leading to *Caenorhabditis elegans* and *Danio rerio*. This is possibly due to ongoing dynamics of polymorphic copy number variations. The same tree showing the number of losses is provided in the supplementary material (Figure S10)

Discussion

Possible uses of RefineTree

RefineTree is a gene tree correction toolbox that explores part of the tree space around a given gene tree, using information from a species tree and synteny. As such, it is modular and can be used with variations.

Various ways of contracting the branches of the starting tree can be considered, varying thresholds or choosing specific branches to contract. For example an exploration scheme contracting the branches one by one and applying ProfileNJ can be considered, which would be equivalent to local modifications as in Chaudhary et al. (2011). A more radical modification would be to contract all branches. Other kinds of contraction schemes can be imagined, as contracting branches around "Non Apparent Duplications" (Lafond et al. 2014), or "Dubious duplications" stored in the Ensembl trees.

Notice that moves considered here are not local reversible moves such as NNI, SPR, TBR, that can be used in a Monte Carlo exploration framework with a Metropolis algorithm. However our method could be used to produce a starting tree to speed-up a burn-in step and start a sampling from plausible trees. It might also be useful to guide proposals that would have a good probability to be accepted. These steps would speed-up the convergence, which could be useful as these techniques are known to be rather time consuming on large data.

An integrated model of genome evolution

The corrections and evaluations we propose are not integrated in a mathematical framework of genome evolution. They are fast and intuitive ways to construct, according to a range of different criteria and on a whole genome scale, gene trees that are better rated than the current state of the art. In order to integrate these principles in a model of genome evolution, we would need to model the stability of genome content and the linearity of ancestral chromosomes. While a local version of the former is contained in gene content evolution and can be integrated (Boussau et al. 2013), modeling linearity is more out of reach. A lot of models for genome evolution have linear structures to handle chromosomes (Fertin et al. 2009), but none is able to include duplications and losses at a whole genome scale. For chromosomes defined as a list of local neighborhoods, like here using DeCo, a probability distribution of ancestral genomes according to their linearity, as well as a probabilistic version of DeCo, that could be used in an integrative model, still need to be developed.

Phylogeny and the quest for orthologs

As gene trees contain the most complete information about a gene family history, detecting orthologs or studying gene repertoire evolution should be achieved by interpreting trees. But due to the rate of errors in the current trees stored in databases, orthology is often assessed with a series of techniques including synteny (Sonnhammer et al. 2014) and Reciprocal Best Hits, while the evolution of gene repertoire is often studied with phyletic profile techniques (Cohen et al. 2012). What we present here is a way of integrating those diverse techniques into a phylogenetic framework. Full sets of orthology relations may de derived from our set of trees, while lists are more incomplete when derived from the Ensembl trees.

Not only the gene trees

Using genome evolution in the construction of the gene trees, we get ancestral genomes as a byproduct. They are made of genes and sets of gene adjacencies. They are still too big (in terms of gene number) and too non linear to be fully trusted. This is partly due to incorrect gene trees in our output, or incorrect inferences from DeCo, but also to problems in sequencing, assembling, annotating genomes, clustering families or inferring the species tree. Good methods for finding linear structures from a set of adjacencies exist (Maňuch et al. 2012). Here we rather used non-linearity as a testimony of the flaws of the data and methods used to reconstruct genome evolution.

Although gene trees are "better" with our correction, they are still not good enough. The likelihood drop for synteny correction is indeed surprising, as these corrections lead to ancestral genomes that are closer to gene content and gene neighborhoods of extant genomes. We would need better exploration schemes with integrated models to really trust gene trees on a whole genome database within a deep phylogeny.

Methods

Families, alignments and trees are taken from Ensembl Compara release 73. They were computed with a pipeline called TreeBest, but we simply call them the "Ensembl trees". Trees are rooted and available with branch support and annotation. There are 20529 trees, each corresponding to a gene family, for a total of 1091891 genes taken from 67 species. Information on gene position on chromosomes, scaffolds or contigs is available. See ftp://ftp.ensembl.org/pub/release-73/emf/ensembl-compara/homologies/.

ProfileNJ

A sketch of the algorithm is provided in Figure 1. First, gene tree branches with support values below a user-defined threshold are collapsed to produce a multifurcating tree. In the case of an unrooted gene tree, all internal nodes are successively considered as potential roots. For each such rooted gene tree, each polytomy is considered in turn in a bottom-up traversal of the tree. For one particular polytomy, a binary resolution is constructed in the following way. Similarly to Csurös (2010), counting ancestral genes along a species tree, gene counts are assigned here to branches of the species tree. The algorithm consists in constructing a table M where, for each node (including leaves) u of S and each integer k (limits on k are discussed in Lafond et al. (2012)), $M(u,k)$ is the minimum number of operations (duplications and losses) required to have k genes in the branch of the species tree leading to u. For the root r of S, $M(r,1)$ is the minimum number of duplications and losses resulting from the reconciliation of any binary resolution of the polytomy with the species tree. Table M can be constructed in time linear in the size of S (Lafond et al. 2012).

When M counts have been computed for all polytomies, a backtracking algorithm outputs a vector containing the number of genes per branch of S. Such a vector is associated with a scenario of gains and losses along the species tree.

For example, the solution vector *count* output from Step 4 in Figure 1 accounts for two duplications. The first duplication, located on the branch leading to b, is deduced from the fact that extant genome b contains three gene copies, while *count* indicates two copies on the branch leading to b; the second one, located on the branch leading to d, is deduced from the fact that two genes are assigned respectively to the branches leading to a and b, while only one gene is assigned to the branch leading to d. The vector *count* does not however involve any information allowing to know which of the three genes b_1, b_2, b_3 are implicated in the first duplication. For each such solution *count*, there are many possible compatible binary gene trees.

We construct here one tree with an NJ criterion, constrained by the solution in terms of number of genes, as follows: a leaf of a gene tree is constructed for each gene present in the subtree. We call D the distances issued from the input distance matrix. At the beginning D induces a metric space on the leaves of G. As in the NJ algorithm, genes at the internal nodes will become elements for the metric space as well.

Let X be an internal node of the species tree with children X_1 and X_2. Suppose X, X_1, X_2 respectively have x, x_1, x_2 genes according to the computed solution. Suppose by induction that the elements of X_1 and X_2 are in the metric space, but not the ones of X. If $x_1 > x$ or $x_2 > x$, the difference is explained by duplications. This duplication history is constructed in a first phase. So first suppose $x < x_i$. Choose, according to the NJ criterion, the couple of elements a, b from X_i which minimizes

$$Q(a,b) = (n-2)D(a,b) - \sum_{i \neq a} D(a,i) - \sum_{i \neq b} D(b,i). \quad (1)$$

Replace a and b by a new element ab in X_i, update the distances and the tree like in the NJ algorithm. This makes x_i decrease until we have $x \geq x_1$ and $x \geq x_2$.

So in a second phase we can suppose that the duplication history has been constructed and we suppose that $x \geq x_1$ and $x \geq x_2$. Suppose also without loss of generality that $x_1 < x_2$. This means that there are x_1 pairs of orthologs to couple. For this choose a from X_1 and b from X_2 which again minimize (1), and replace a and b by an element $x \in X$. When X_1 is empty replace every element from X_2 by one element in X.

At this step some elements of X might not be in the metric space. They correspond to an internal branch of the gene tree. Then it is easy to construct an element of the metric space by applying the NJ updating step on the fixed gene tree (for a fixed subtree there is not selection step)

At the end of these procedures all elements of X correspond to an element of the metric space, so an iteration is possible, to the next node of the species tree.

Note that if every species contains at most one gene in the family and if the starting gene tree is a star then the algorithm simply constructs the species tree as a gene tree. Differences can lie in internal branches that are in the starting tree but not in the species tree, and the main difficulty of the problem is the presence of several genes from the same family in some species. Eventually, if all genes of a family belong to the same species and the starting tree is a star, then the solution is simply an NJ tree. In that way, we really generalize NJ, ancestral gene counts and gene tree species tree reconciliation in a single method.

Use of ProfileNJ on Ensembl

PhyML was used with default parameters to compute maximum likelihood trees from the protein multiple alignments from Ensembl. An aLRT support was computed, and all branches with aLRT < 0.95 were contracted. FastDist was run on DNA alignments to provide a distance matrix. Then ProfileNJ was run with the command (an example is given for the first family).

```
ProfileNJ -s Compara.73.species_tree  \\
          -g data/famille_1.start_tree \\
          -d data/famille_1.dist \\
          -o data/famille_1.tree \\
          -n -r best -c nj --slimit 1 --plimit 1 --firstbest --cost 1 0.99999
```

We tested the sensitivity of the method to the choice of the threshold parameter for contracting unsupported branches. The threshold is a trade-off between the amount of change in a tree and the probability that the resulting tree is rejected. Too small values would avoid exploring a large space around the starting tree while high values would lead to low likelihood trees. It has to be settled empirically. For example .80 was considered an acceptable threshold in some genomic studies (Abby et al. 2012).

Ancestral Genomes (gene content and order) from the LCA Reconciliation

LCA reconciliation is used to infer ancestral gene contents, one family at a time. It consists in labeling every node x of the gene tree with a node of the species tree corresponding to the last common ancestor of all extant species containing a gene which is a descendant of x (including x itself if x is a leaf). Then every internal node x is labeled with an event: a duplication if the species label of x is equal to the label of one of its children, and a speciation otherwise.

The LCA reconciliation induces sets of ancestral genes: for a species S (extant or ancestral), draw a graph in which every leaf of a gene tree which maps to S or one of its descendant is a node. Then draw an edge between two homologous genes x and y if their last common ancestor in the gene tree maps to a proper descendant of S, or to S but is a speciation node. Connected components of this graph are the ancestral genes in S: there is exactly one ancestral gene per component, and its descendants are the nodes of the component.

We organized the ancestral genes in the genomes using DeCo (Bérard et al. 2012). This algorithm aims at constructing the neighborhoods between ancestral genes, but starts by inferring ancestral gene contents.

This LCA method assumes the trees to be binary and does not take branch support and node annotation into account. In particular, the algorithm ignores the fact that a branch may be uncertain due to a weak support, or that a node may be labeled as dubious as in Ensembl. However, part of our goal is precisely to resolve unsupported and dubious parts of gene trees, by considering the validity of the obtained ancestral genomes.

Testing the linearity of ancestral genomes with DeCo

DeCo (Bérard et al. 2012) computes ancestral gene neighborhoods that are highly dependant on the shape of the considered gene tree. Indeed, adjacencies in extant genomes, *i.e.* the immediate proximity of two consecutive genes, are taken as input and putative adjacencies in ancestral genomes are constructed by a parsimony principle minimizing the number of gains and losses of adjacencies. As two contemporaneous

adjacencies are supposed to evolve independently one from the other, the linearity of extant genomes, *i.e.* the property that one gene never has more than two neighbors linked by an adjacency, does not guarantee the linearity of ancestral ones.

The apparent weakness of this feature is in fact a strength to evaluate the quality of gene trees. Indeed, a high part of the non linearity of ancestral genomes is not due to the inadequacy of the software itself, but to the quality of the input data. Indeed it has been remarked that a significant improvement in the linearity of ancestral genomes was obtained by constructing gene trees according to more complete models (Boussau et al. 2013; Patterson et al. 2013).

Note that in extant genomes, no gene can have more than two neighbors, and most genes have two. But many genes have 1 or 0, because of the poor assembly of some genomes, many contigs contain one or a few genes.

Information from extant synteny

First we ran PhylDiag as follows, for each pair of genomes. Files genome_1, genome_2 and ancestral_genes respectively contain the ordered list of genes from each genome, and the list of families clustering the genes as in the Ensembl database.

```
phylDiag.py genome_1 genome_2 ancestral_genes \\
    -gapMax=2 -pThreshold=0.00000005 \\
    -filterType=InBothSpecies -multiprocess \\
    -minChromLength=2 >syntenyblocks_1_2
```

The statistical threshold is calculated in order to minimize the number of false positives, taking into account the number (2211) of comparisons between pairs of species and the expected number (500) of synteny blocks for each comparison ($0.05/(2211 * 500) \approx 5e-8$).

For each synteny block found by PhylDiag, we kept only the genes that had one single exemplar in the two blocks from both species. We counted the number of such pairs of genes, and refered to an LCA reconciliation of the output trees of ProfileNJ to check that most pairs are orthologs (their common ancestor is labeled by a speciation). We discarded the blocks that did not fit this condition. This discards possible block duplications.

For the remaining blocks, and for each couple of uniquely represented genes a and b, we required that the LCA node X of a and b in the reconciled ProfileNJ tree is not a supported duplication: let X_1 and X_2

be the two children of the node X labeled as a duplication (so X_1 and X_2 are in the same species as X), the genes a and b are not kept as putative orthologs if one of the branches XX_1 and XX_2 has a high support (> 0.95), and there are two genes, x_1 and x_2, which respectively descend from X_1 and X_2, which are located on the same genome. This discards possible tandem duplications in the block, followed by differential losses of copies.

The output trees from ProfileNJ as well as the filtered pairs of putative orthologs were givent as input to ParalogyCorrector, which finds the tree that is as close as possible to the input tree in terms of RF distance, such that in an LCA reconciliation, all pairs of putative orthologs have an LCA node annotated as a speciation.

Information from ancestral synteny

From the results of DeCo on the output gene trees produced by ProfileNJ, we used an "unduplication" principle as in (Chauve et al. 2013) everytime we found that an ancestral gene x had three neighbors a, b, c, two of them (say a, b) arising from a duplication node d in a single gene tree. In that case, we rearranged the four grand children of d so that the clade under d has an LCA which is annotated as a speciation in the LCA reconciliation. See an insight into its functionning in Figure 8.

Likelihood ratio tests

We computed the likelihood of all trees according to the HKY85 model with PhyML on nucleotide alignments. To test the significance of a likelihood difference, we computed the AU (Appoximately unbiased) tests with Concel. They consist in bootstraping the sites of an alignment, each site having a likelihood according to several trees. Then a probability is associated to each tree from this bootstrap, according to the number of replicates which place it above the others in terms of the bootstrapped likelihood. Unless otherwise stated, we use "significantly" better for a likelihood with a AU value > 0.95.

Data access

The 2575 simulated gene families used for our simulation represent a subset of the original SPIMAP simulated fungi datasets (see http://compbio.mit.edu/spimap/). Those data and the RAxML trees constructed from sequence alignment are available. We also provide the two sets of 20529 trees, as an output from ProfileNJ and with the additional synteny-aware corrections. All softwares are freely accessible for academic purpose,

under a GPL license.

Acknowledgments

MS, LG, BB and ET were supported by the French Agence Nationale de la Recherche (ANR) through Grant ANR- 10-BINF-01-01 "Ancestrome". EN, ML, JS and NEM were supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the "Fonds de recherche du Québec Nature et technologies" (FRQNT) of Quebec. Computations were made on the supercomputer "Briarée" from Université de Montréal, managed by Calcul Québec and Compute Canada.

Disclosure Declaration

None.

Figure legends

Figure 1: ProfileNJ at a glance. The input is the species tree S, an unrooted gene tree G with multifurcations and a distance matrix d. Step 1. Each internal node of the gene tree is considered in turn for the root. Here the considered root is highlighted in red in G. Step 2. In a bottom-up traversal of the tree, each polytomy P (non-binary node) is considered in turn. Step 3. With P and S as input, a dynamic programming table M is constructed. $M(u, k)$ denotes the minimum weight scenario of duplications and losses required to have k genes in the branch of the species tree leading to u. Each entry is constructed from neighboring entries as in (Lafond et al. 2012). Step 4. All minimum weight duplication and loss *count* solutions are obtained by backtracking in M. Step 5. One *count* solution might correspond to many binary trees. The Neighbor joining (NJ) procedure computes the one that best agrees with the distance matrix. The final completely resolved tree is given bottom left.

Figure 2: Topology accuracy of RAxML, TreeFix and ProfileNJ trees, mesured by RF distance with the true tree, on $\sim$ 2500 simulated trees from the fungal dataset. We use a sample of trees simulated under four different DL rate : ($1r_D$ - $1r_L$), ($2r_D$ - $2r_L$), ($4r_D$ - $4r_L$) and ($4r_D$ - $1r_L$). Percentage of reconstructed trees (y-axis) with a given RF distance (x-axis) to the true tree. TreeFix and ProfileNJ have a similar reconstruction accuracy (75% of trees match the true trees) while the input tree (RAxML) have the lowest accuracy. The graph is cut on the right, but contains more than 99% of the data.

Figure 3: Runtime of TreeFix and ProfileNJ for increasing size of gene tree.

Figure 4: A general view on RefineTree when run on the Ensembl Compara gene families. An example is given for a species tree S of four fish species, a gene family of six genes (a gene is represented by the picture of the species it belongs to, and two paralogs belonging to the same species are distinguished by a different frame color), a rooted gene tree G (although it can be unrooted in general) with branch support, and a given threshold for branch contraction. Data framed in black are the input and those framed in blue are the output of the correction algorithm labeling the edge linking the considered frames. Black arrows depict the use we make of RefineTree on the Ensembl gene trees. The green arrow and the green "or" are alternative uses avoiding one or both of the correction tools ParalogyCorrector and Unduplicator. Any framed set of data can be alternatively provided to the pipeline as input. For example, orthology constraints obtained from various sources can be directly provided as input to ParalogyCorrector. The method for inferring orthology constraints from synteny blocks is described in the text.

Figure 5: Sequence likelihood, ancestral genome content and ancestral chromosome linearity for ProfileNJ, Synteny and Ensembl trees: **(A)** Proportion of trees with a significantly better likelihood computed with PhyML. AU tests were computed for the three trees for each family, and if the tree at the first rank was significantly better than the second, it was stored as the best likelihood, and if not, it was stored as "no significant difference at the first rank". **(B)** Gene content computed with DeCo. Gene content has one value for each node of the phylogeny of 65 species, except for extant genomes, for which it has one value for each leaf. **(C)** Genome linearity computed with DeCo. Genome linearity is represented by a graph, whose x axis is the number of neighbors a gene can have, and the y axis shows the proportion of genes having this number of neighbors. Parameters from extant genomes are given as a reference in (B) and (C). Statistics for ancestral genomes are assumed better when close to the extant ones.

Figure 6: A probable example of ILS visible on a subtree of an ensembl gene family. The monophyly of the chimpanze and gorilla genes (ENSPTRP00000033018 and ENSGGOP00000011432) is well supported by the sequences (left tree, constructed by PhyML, with aLRT supports), while synteny argues for orthology of both with the human genes (ENSP00000414208 and ENSP00000378687) (right tree, constructed by ProfileNJ followed by ParalogyCorrector), so that a scenario of duplication and losses compatible with the left tree is unlikely.

Figure 7: Numbers of duplications in the eukaryote phylogeny, estimated with reconciled ProfileNJ trees from PhyML starting trees on the whole Ensembl Compara database, version 73.

Figure 8: The unduplication principle (figure from Chauve et al. (2013)). A non linearity is detected in an ancestral genome (gene g has three neighbors). Two of its neighbors g_1 and g_2 are issued from a possibly dubious duplication labeled node. The tree is rearranged so that its root is labeled with a speciation instead of a duplication. In the resulting configuration g'_1 and g'_2 are in two different species, so that g can have only one neighbor in this family, and linearity is recovered.

Figures

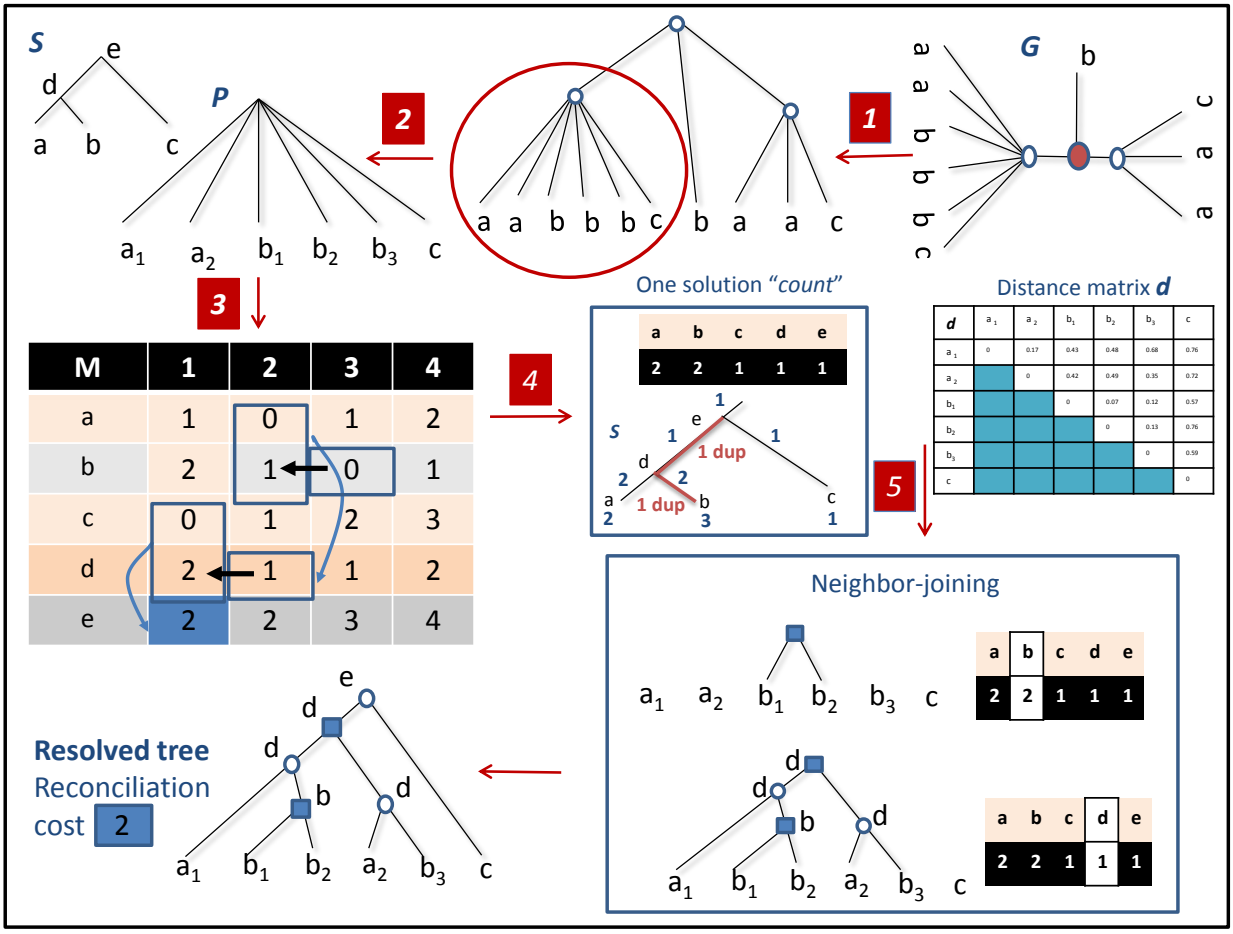

Figure 1

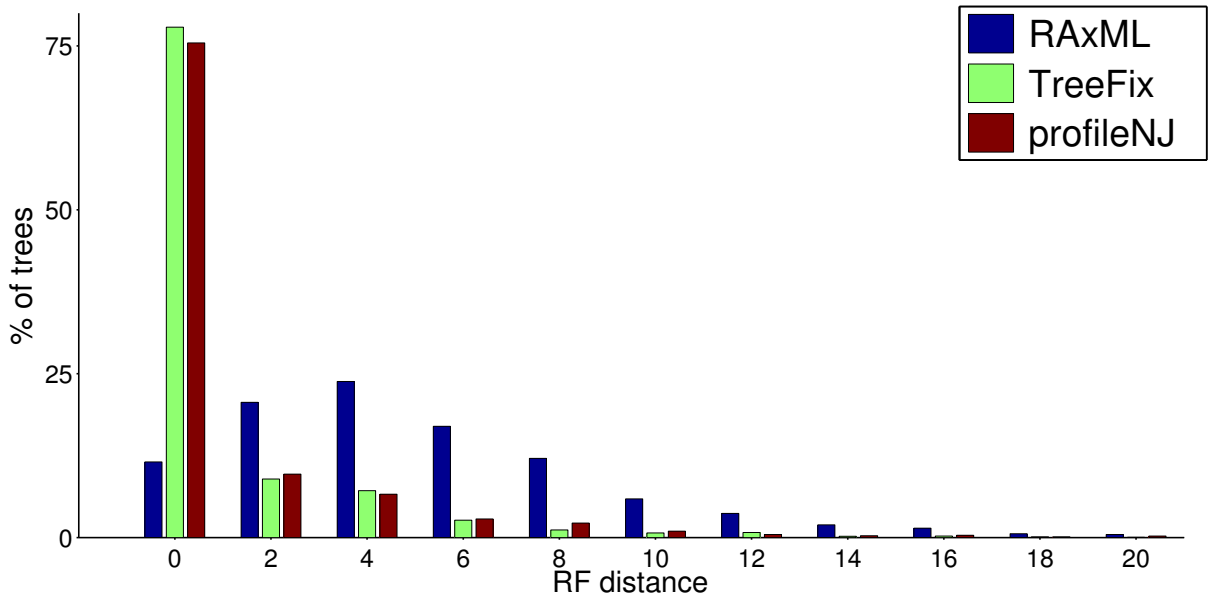
RAxML
TreeFix
profileNJ
% of trees
75
50
25
0
0
2
4
6
8
10
12
14
16
18
20
RF distance

Figure 2

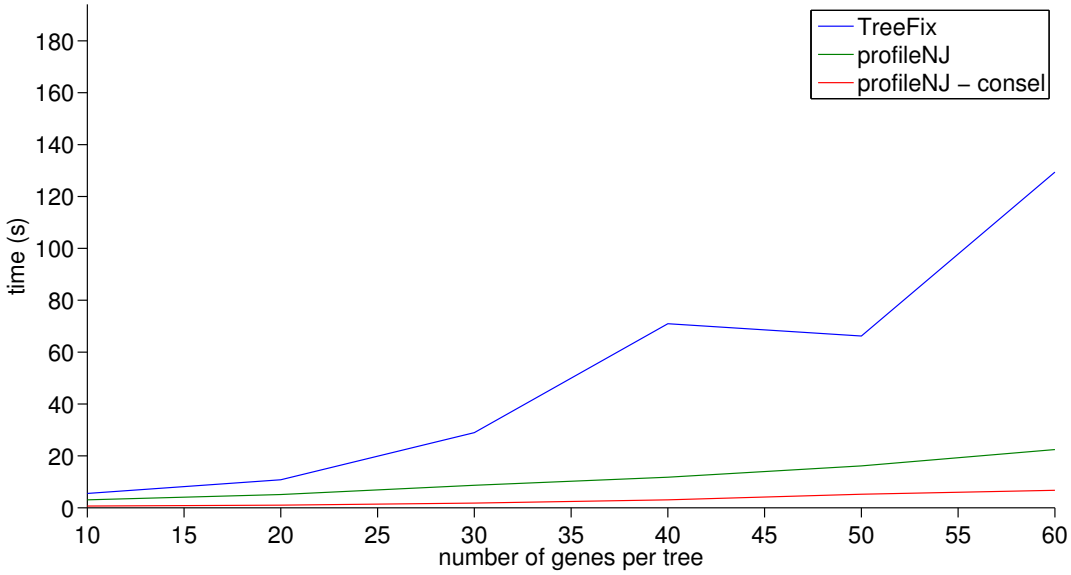
TreeFix
profileNJ
profileNJ – consel
time (s)
0
20
40
60
80
100
120
140
160
180
10
15
20
25
30
35
40
45
50
55
60
number of genes per tree

Figure 3

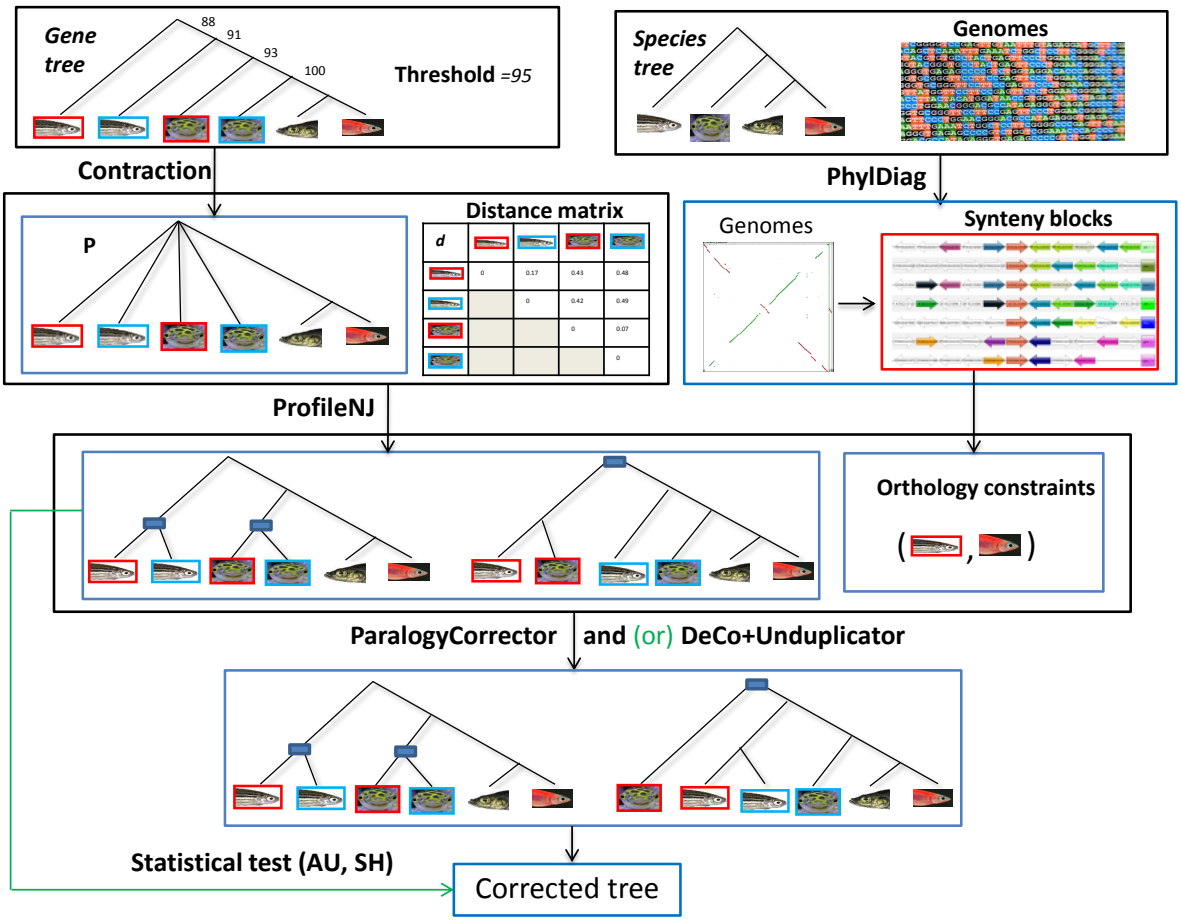
Gene tree
88
91
93
100
Threshold =95
Species tree
Genomes
Contraction
PhylDiag
P
Distance matrix
Genomes
Synteny blocks
ProfileNJ
Orthology constraints
(,)
ParalogyCorrector
and (or) DeCo+Unduplicator
Statistical test (AU, SH)
Corrected tree

Figure 4

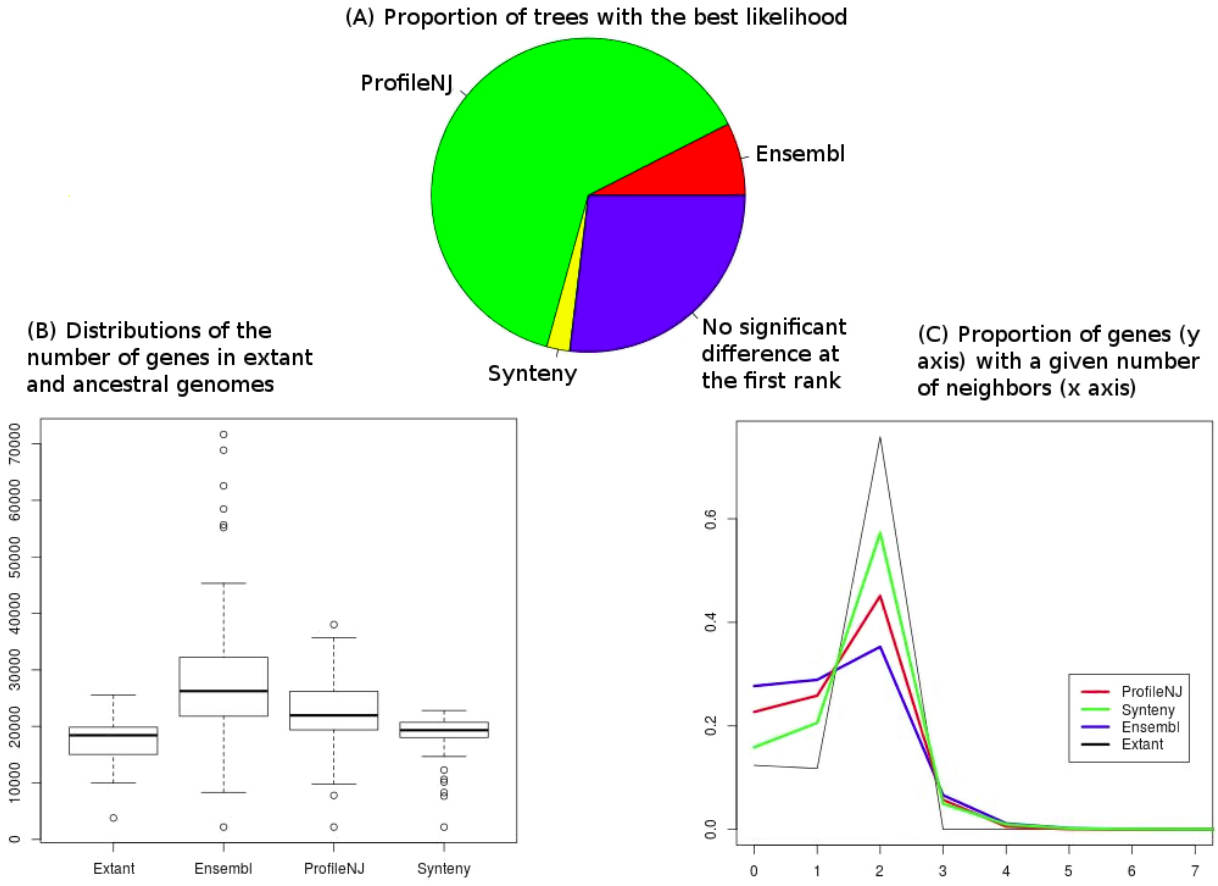
(A) Proportion of trees with the best likelihood
ProfileNJ
Ensembl
No significant difference at the first rank
Synteny
(B) Distributions of the number of genes in extant and ancestral genomes
0
10000
20000
30000
40000
50000
60000
70000
Extant
Ensembl
ProfileNJ
Synteny
(C) Proportion of genes (y axis) with a given number of neighbors (x axis)
0.0
0.2
0.4
0.6
0
1
2
3
4
5
6
7
ProfileNJ
Synteny
Ensembl
Extant

Figure 5

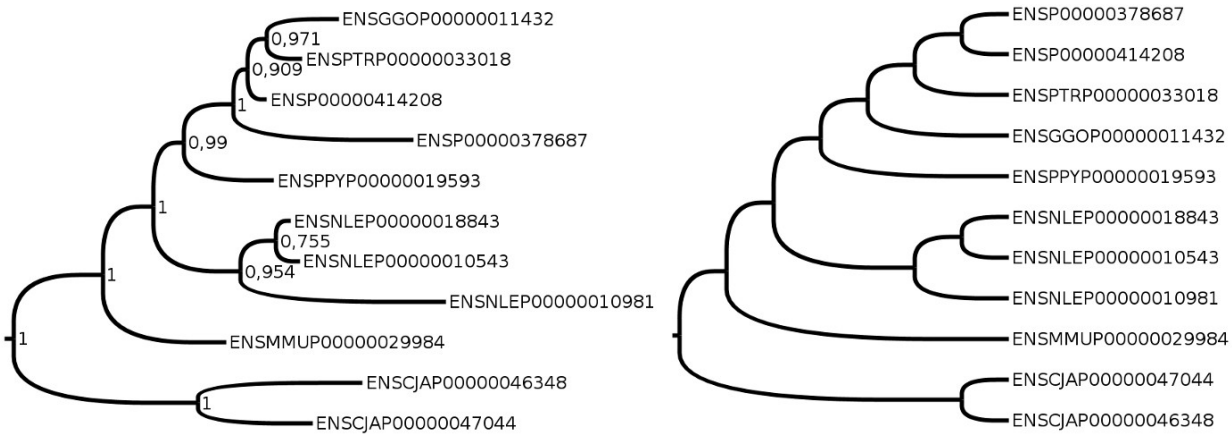
ENSGGOP00000011432
0,971
ENSPTRP00000033018
0,909
ENSP00000414208
1
ENSP00000378687
0,99
ENSPPYP00000019593
1
ENSNLEP00000018843
0,755
ENSNLEP00000010543
1
0,954
ENSNLEP00000010981
1
ENSMMUP00000029984
ENSCJAP00000046348
1
ENSCJAP00000047044
ENSP00000378687
ENSP00000414208
ENSPTRP00000033018
ENSGGOP00000011432
ENSPPYP00000019593
ENSNLEP00000018843
ENSNLEP00000010543
ENSNLEP00000010981
ENSMMUP00000029984
ENSCJAP00000047044
ENSCJAP00000046348

Figure 6

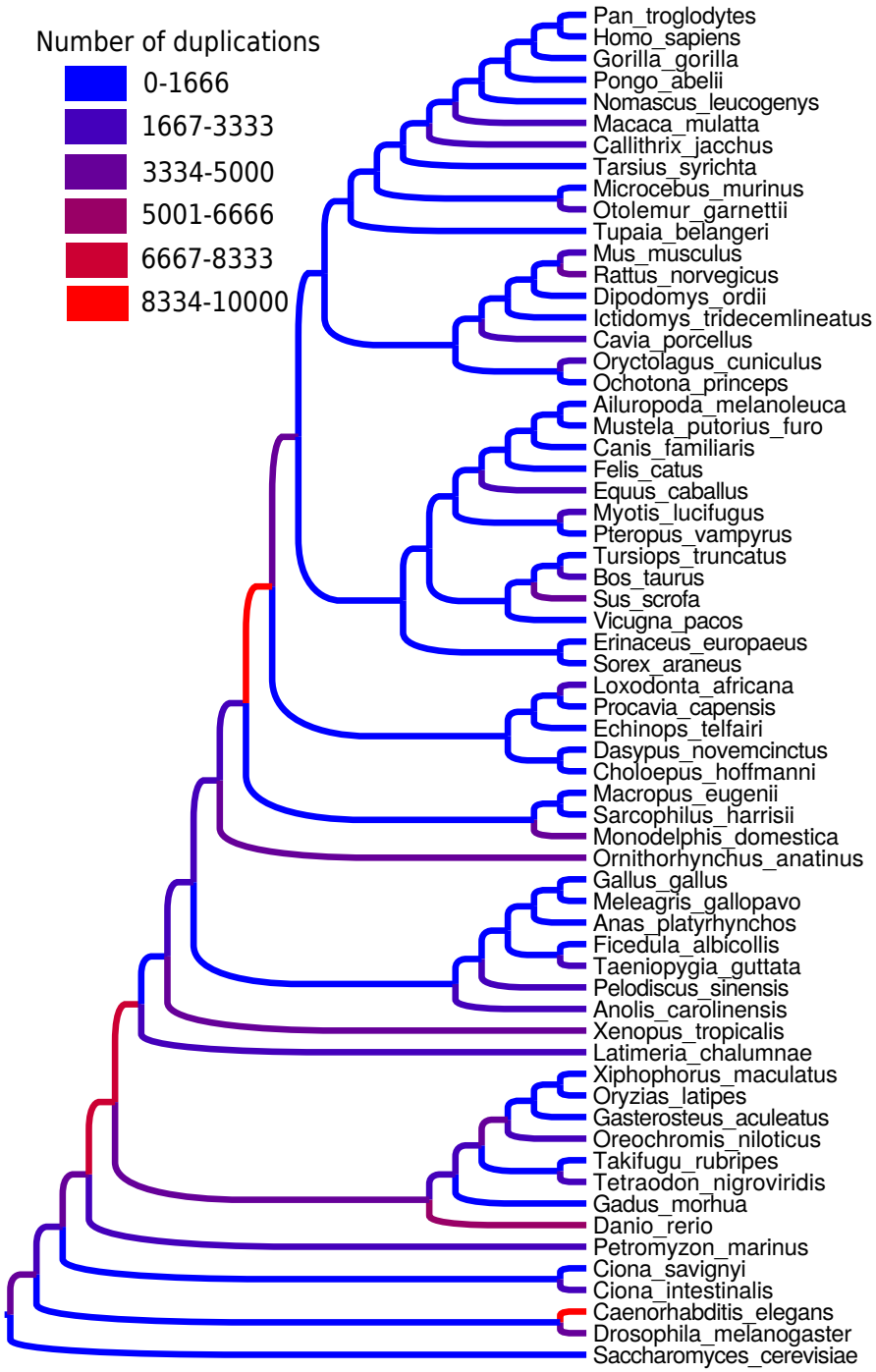
Number of duplications
0-1666
1667-3333
3334-5000
5001-6666
6667-8333
8334-10000
Pan_troglodytes
Homo_sapiens
Gorilla_gorilla
Pongo_abelii
Nomascus_leucogenys
Macaca_mulatta
Callithrix_jacchus
Tarsius_syrichta
Microcebus_murinus
Otolemur_garnettii
Tupaia_belangeri
Mus_musculus
Rattus_norvegicus
Dipodomys_ordii
Ictidomys_tridecemlineatus
Cavia_porcellus
Oryctolagus_cuniculus
Ochotona_princeps
Ailuropoda_melanoleuca
Mustela_putorius_furo
Canis_familiaris
Felis_catus
Equus_caballus
Myotis_lucifugus
Pteropus_vampyrus
Tursiops_truncatus
Bos_taurus
Sus_scrofa
Vicugna_pacos
Erinaceus_europaeus
Sorex_araneus
Loxodonta_africana
Procavia_capensis
Echinops_telfairi
Dasypus_novemcinctus
Choloepus_hoffmanni
Macropus_eugenii
Sarcophilus_harrisii
Monodelphis_domestica
Ornithorhynchus_anatinus
Gallus_gallus
Meleagris_gallopavo
Anas_platyrhynchos
Ficedula_albicollis
Taeniopygia_guttata
Pelodiscus_sinensis
Anolis_carolinensis
Xenopus_tropicalis
Latimeria_chalumnae
Xiphophorus_maculatus
Oryzias_latipes
Gasterosteus_aculeatus
Oreochromis_niloticus
Takifugu_rubripes
Tetraodon_nigroviridis
Gadus_morhua
Danio_rerio
Petromyzon_marinus
Ciona_savignyi
Ciona_intestinalis
Caenorhabditis_elegans
Drosophila_melanogaster
Saccharomyces_cerevisiae

Figure 7

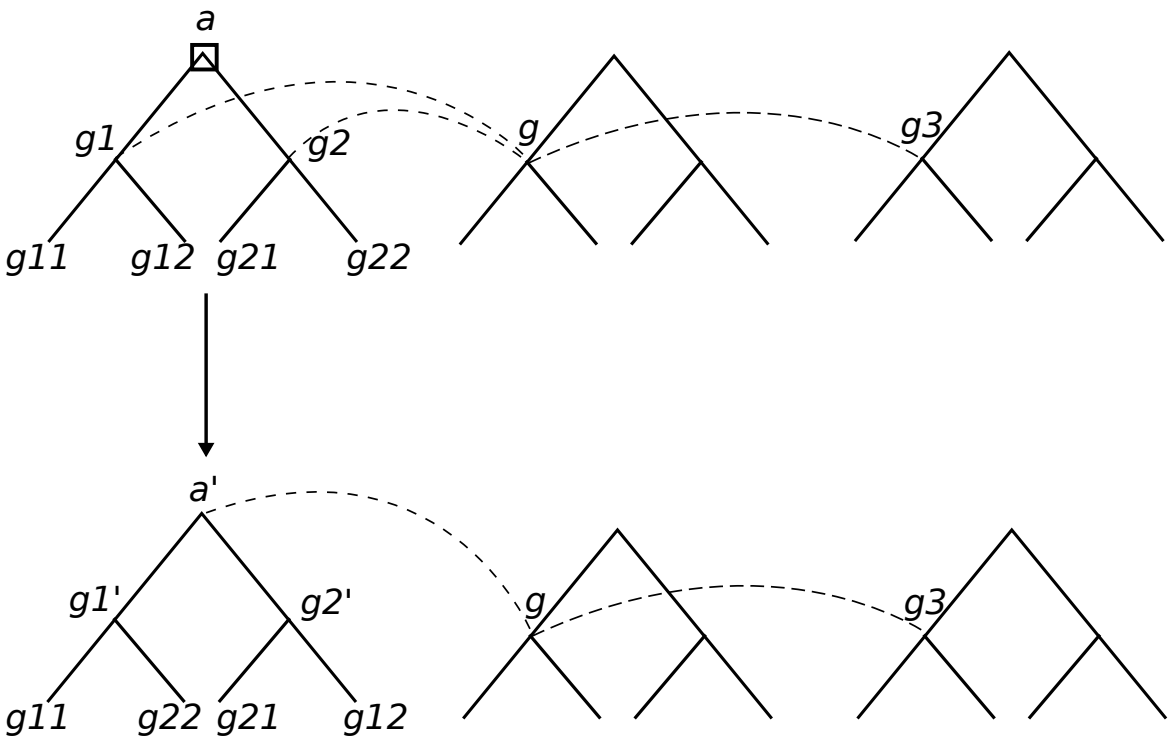

Figure 8

References

Abby SS, Tannier E, Gouy M, and Daubin V. 2012. Lateral gene transfer as a support for the tree of life. *Proceedings of the National Academy of Sciences USA* **109**: 4962–4967.

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 USA* **106**: 5714–5719.

Arvestad L, Berglund A, Lagergren J, and Sennblad B. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *RECOMB*, pp. 326-335.

Bansal MS, Wu YC, Alm EJ, and Kellis M. 2014. Improved gene tree error-correction in the presence of horizontal gene transfer. *Bioinformatics* .

Bérard S, Gallien C, Boussau B, Szöllősi GJ, Daubin V, and Tannier E. 2012. Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics* **28**: i382–i388.

Berglund-Sonnhammer A, Steffansson P, Betts M, and Liberles D. 2006. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *Journal of Molecular Evolution* **63**: 240-250.

Boeckmann B, Robinson-Rechavi M, Xenarios I, and Dessimoz C. 2011. Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief Bioinform* **12**: 423–435.

Boussau B, Szöllősi G, Duret L, Gouy M, Tannier E, and Daubin V. 2013. Genome-scale coestimation of species and gene trees. *Genome Research* **23**: 323-330.

Chaudhary R, Burleigh J, and Eulenstein O. 2011. Efficient error correction algorithms for gene tree reconciliation based on duplication, duplication and loss, and deep coalescence. *BMC Bioinformatics* **13**: S11.

Chauve C, El-Mabrouk N, Guéguen L, Semeria M, and Tannier E. 2013. Duplication, rearrangement and reconciliation: A follow-up 13 years later. In *Models and Algorithms for Genome Evolution* (eds. C Chauve, N El-Mabrouk, and E Tannier), pp. 47–62. Springer, London.

Chen K, Durand D, and Farach-Colton M. 2000. Notung: Dating gene duplications using gene family trees. *Journal of Computational Biology* **7**: 429–447.

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

Cohen O, Ashkenazy H, Burstein D, and Pupko T. 2012. Uncovering the co-evolutionary network among prokaryotic genes. *Bioinformatics* **28**: i389–i394.

Csurös M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**: 1910–1912.

Datta R, Meacham C, Samad B, Neyer C, and Sjölander K. 2009. Berkeley phog: Phylofacts orthology group prediction web server. *Nucleic Acids Research* **37**: W84-W89.

Doroftei A and El-Mabrouk N. 2011. Removing noise from gene trees. In *WABI*, volume 6833 of *LNBI/LNBI*, pp. 76-91.

Durand D, Haldórsson B, and Vernot B. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *Journal of Computational Biology* **13**: 320–335.

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

Felsenstein J. 2005. PHYLIP(phylogeny inference package). Version 3.6. distributed by the author, Seattle (WA): Department of Genome Sciences, University of Washington.

Fertin G, Labarre A, Rusu I, and Vialette ETT. 2009. *Combinatorics of Genome Rearrangements*. MIT press.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al.. 2014. Ensembl 2014. *Nucleic Acids Research* **42**: D749–D755.

Gorecki P and Eulenstein O. 2011a. Algorithms: simultaneous error-correction and rooting for gene tree reconciliation and the gene duplication problem. *BMC Bioinformatics* **13**: S14.

Gorecki P and Eulenstein O. 2011b. A linear-time algorithm for error-corrected reconciliation of unrooted gene trees. In *ISBRA*, volume 6674 of *LNBI*, pp. 148-159. Springer-Verlag.

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

Huerta-Cepas J, Capella-Gutierrez S, Pryszcz L, Denisov I, Kormes D, Marcet-Houben M, and Gabald'on T. 2011. Phylomedb v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Research* **39**: D556-D560.

Jun J, Mandoiu II, and Nelson CE. 2009. Identification of mammalian orthologs using local synteny. *BMC Genomics* **10**: 630.

Khan MA, Elias I, Sjölund E, Nylander K, Guimera RV, Schobesberger R, Schmitzberger P, Lagergren J, and Arvestad L. 2013. Fastphylo: fast tools for phylogenetics. *BMC Bioinformatics* **14**: 334.

Lafond M, Chauve C, Dondi R, and El-Mabrouk N. 2014. Polytomy refinement for the correction of dubious duplications in gene trees. *Bioinformatics* **30**: i519–i526.

Lafond M, Semeria M, Swenson K, Tannier E, and El-Mabrouk N. 2013. Gene tree correction guided by orthology. *BMC Bioinformatics* **14 (supp 15)**.

Lafond M, Swenson K, and El-Mabrouk N. 2012. An optimal reconciliation algorithm for gene trees with polytomies. In *LNCS*, volume 7534 of *WABI*, pp. 106-122.

Lartillot N and Philippe H. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* **21**: 1095–1109.

Lucas JM, Muffato M, and Roest Crollius H. 2014. Phyldiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees. *BMC Bioinformatics* **15**: 268.

Mahmudi O, Sjöstrand J, Sennblad B, and Lagergren J. 2013. Genome-wide probabilistic reconciliation analysis across vertebrates. *BMC Bioinformatics* **14 Suppl 15**: S10.

Maňuch J, Patterson M, Wittler R, Chauve C, and Tannier E. 2012. Linearization of ancestral multichromosomal genomes. *BMC Bioinformatics* **13 Suppl 19**: S11.

Mehta TK, Ravi V, Yamasaki S, Lee AP, Lian MM, Tay BH, Tohari S, Yanai S, Tay A, Brenner S, et al.. 2013. Evidence for at least six hox clusters in the japanese lamprey (lethenteron japonicum). *Proceedings of the National Academy of Sciences* **110**: 16044–16049.

Mi H, Muruganujan A, and Thomas P. 2012. Panther in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research* **41**: D377-D386.

Nguyen TH, Ranwez V, Pointet S, Chifolleau AMA, Doyon JP, and Berry V. 2013. Reconciliation and local gene tree rearrangement can be of mutual profit. *Algorithms for Molecular Biology* **8**: 12.

Patterson M, Szöllősi G, Daubin V, and Tannier E. 2013. Lateral gene transfer, rearrangement, reconciliation. *BMC Bioinformatics* **14 Suppl 15**: S4.

Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, and Perrière G. 2009. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* **10 Suppl 6**: S3.

Pryszcz L, Huerta-Cepas J, and Gabaldón T. 2011. MetaPhOrs: orthology nd paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Research* **39**: e32.

Rasmussen M and Kellis M. 2011. A bayesian approach for fast and accurate gene tree reconstruction. *Molecular Biology and Evolution* **28**: 273- 290.

Rasmussen MD and Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research* **22**: 755–765.

Ronquist F and Huelsenbeck J. 2003. MrBayes3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572- 1574.

Saitou N and Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**: 406-425.

Schreiber F, Patricio M, Muffato M, Pignatelli M, and Bateman A. 2013. Treefam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Research* Doi: 10.1093/nar/gkt1055.

Shimodaira H and Hasegawa M. 2001. Consel: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**: 1246–1247.

Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, Yandell MD, Manousaki T, Meyer A, Bloom OE, et al.. 2013. Sequencing of the sea lamprey (petromyzon marinus) genome provides insights into vertebrate evolution. *Nature genetics* **45**: 415–421.

Sonnhammer ELL, Gabaldón T, Sousa da Silva AW, Martin M, Robinson-Rechavi M, Boeckmann B, Thomas PD, Dessimoz C, and the Quest for Orthologs consortium. 2014. Big data and other challenges in the quest for orthologs. *Bioinformatics* p. btu492.

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

Swenson KM, Doroftei A, and El-Mabrouk N. 2012. Gene tree correction for reconciliation and species tree inference. *Algorithms for Molecular Biology* **7**: 31.

Szöllősi G, Boussau B, Abby S, Tannier E, and Daubin V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences USA* **109**: 17513- 17518.

Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, and Daubin V. 2013. Efficient exploration of the space of reconciled gene trees. *Systematic Biology* **62**: 901–912.

Szöllősi GJ, Tannier E, Daubin V, and Boussau B. 2015. The inference of gene trees with species trees. *Systematic Biology* **64**: e42–e62.

Thomas P. 2010. GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics* **11**: 312.

Vilella A, Severin J, Ureta-Vidal A, Heng L, Durbin R, and Birney E. 2009. EnsemblCompara gene trees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research* **19**: 327-335.

Wapinski I, Pfeffer A, Friedman N, and Regev A. 2007. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics* **23**: i549–i558.

Wu Y, Rasmussen M, Bansal M, and Kellis M. 2013. TreeFix: Statistically informed gene tree error correction using species trees. *Systematic Biology* **62**: 110- 120.

Zimmermann T, S M, and Warnow T. 2014. BBCA: Improving the scalability of BEAST using random binning. *BMC Genomics* p. S11. Proceedings of RECOMB-CG.