PhylteR: efficient identification of outlier sequences in phylogenomic datasets

Aurore Comte, Théo Tricou, Eric Tannier, Julien Joseph, Aurélie Siberchicot, Simon Penel, Rémi Allio, Frédéric Delsuc, Stéphane Dray, Damien de Vienne

PhylteR: efficient identification of outlier sequences in phylogenomic datasets

Aurore Comte[1,2,†], Théo Tricou[3,†], Eric Tannier[3,4], Julien Joseph[3], Aurélie Siberchicot[3], Simon Penel[3], Rémi Allio[5], Frédéric Delsuc[6], Stéphane Dray[3], Damien M. de Vienne[3,*]

[1] French Institute of Bioinformatics (IFB)—South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, Montpellier 34398, France
[2] IRD, CIRAD, INRAE, Institut Agro, PHIM Plant Health Institute, Montpellier University, Montpellier 34398, France
[3] Université de Lyon, Université Lyon 1, UMR CNRS 5558 Laboratoire de Biométrie et Biologie Évolutive, 69622 Villeurbanne, France
[4] Centre de Recherches Inria de Lyon, 68622 Villeurbanne, France
[5] CBGP, INRAE, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier, Montpellier, France
[6] ISEM, Univ. Montpellier, CNRS, IRD, 30495 Montpellier, France
[†] Equal contribution
[*] Corresponding author:
DAMIEN M. DE VIENNE
Université Lyon 1, CNRS
Laboratoire de Biométrie et Biologie Évolutive
Bâtiment Mendel 43 boulevard du 11 Novembre 1918
69622 VILLEURBANNE CEDEX
Phone: +33(0)4 72 43 29 09
E-mail: damien.de-vienne@univ-lyon1.fr

Abstract

In phylogenomics, incongruences between gene trees, resulting from both artifactual and biological reasons, are known to decrease the signal-to-noise ratio and complicate species tree inference. The amount of data handled today in classical phylogenomic analyses precludes manual error detection and removal. However, a simple and efficient way to automate the identification of outlier sequences is still missing.

Here, we present PhylteR, a method that allows a rapid and accurate detection of outlier sequences in phylogenomic datasets, i.e. species from individual gene trees that do not follow the general trend. PhylteR relies on DISTATIS, an extension of multidimensional scaling to 3 dimensions to compare multiple distance matrices at once. In PhylteR, distance matrices obtained either directly from multiple sequence alignments or extracted from individual gene phylogenies represent evolutionary distances between species according to each gene.

On simulated datasets, we show that PhylteR identifies outliers with more sensitivity and precision than a comparable existing method. On a biological dataset of 14,463 genes for 53 species previously assembled for Carnivora phylogenomics, we show (i) that PhylteR identifies as outliers sequences that can be considered as such by other means, and (ii) that the removal of these sequences improves the concordance between the gene trees and the species tree. Thanks to the generation of numerous graphical outputs, PhylteR also allows for the rapid and easy visual characterisation of the dataset at hand, thus aiding in the precise identification of errors.
PhylteR is distributed as an R package on CRAN and as containerized versions (docker and singularity).

Introduction

Supermatrix and supertree approaches are commonly used in phylogenomics to obtain a species tree from a collection of genes. Both methods are similar in their first steps: for a list of taxa of interest, a large collection of single-copy orthologous gene sequences is retrieved and a multiple sequence alignment (MSA) is computed for each cluster of orthologous genes (see von Haeseler 2012 for a comparison of these approaches).
The methods then differ by the strategy employed. In the supermatrix approach, MSAs are concatenated into a supermatrix that is used to build a phylogeny, generally with Maximum Likelihood (ML) or Bayesian methods (such as Phylobayes, Lartillot et al. 2013; or IQTREE, Minh, Schmidt, et al. 2020). In the supertree approach, individual gene trees are built from individual MSAs and a species tree is obtained by combining them all (e.g. ASTRAL, Zhang et al. 2018).
Whatever the method employed, incongruence between inferred individual gene histories and the history of the species carrying these genes negatively impact the quality (accuracy) of the reconstructed species tree (Philippe et al. 2017).
To alleviate this problem, three types of filtering approaches can be used: the pruning of taxa that are unstable among gene trees (the so-called rogue taxa, Aberer et al. 2013), the elimination of problematic orthologous genes families (whose history is uncorrelated with the others), or a more subtle approach consisting in identifying and filtering out only some species in some genes trees (*i.e.* Phylo-MCOA, de Vienne et al. 2012; or TreeShrink, Mai and Mirarab 2018). These last approaches are thought to provide the best compromise between removing problematic signals and keeping the maximum information content.

Here we present PhylteR, a new method that can accurately and rapidly identify outliers in phylogenomics datasets. Unlike Phylo-MCOA (de Vienne et al. 2012), from which it is largely

inspired, it is an iterative process where obvious outliers are removed first, leaving space for better identification of more subtle ones, and leading *in fine* to a finer identification of outliers. Unlike TreeShrink (Mai and Mirarab 2018), it is not based solely on the diameter of unrooted gene trees and is thus more accurate when outliers are not associated with long branches (e.g. topological incongruences). Also, PhylteR relies on the multivariate analysis method DISTATIS (Abdi et al. 2005; Abdi et al. 2012), which is specifically designed, unlike Multiple Co-inertia Analysis (MCOA, Chessel and Hanafi 1996) used in Phylo-MCOA (de Vienne et al. 2012), to compare distance matrices, and is thus more appropriate for the problem at hand.

We tested PhylteR on two types of datasets: a collection of small and simple simulated datasets where outliers were known, and a biological dataset comprising 14,463 genes for up to 53 species previously used for Carnivora phylogenomics (Allio et al. 2021). In this empirical dataset, outliers were of course unknown but "properties" associated to gene sequences can be gathered (see Shen et al. 2016 for a list of such properties). After illustrating the principle of PhylteR on the simulated datasets, we focused on the Carnivora gene sets: we characterised the sequences that were filtered out by PhylteR and we looked at the effect of PhylteR on the overall concordance between the gene trees and the species tree after filtering. We compared the results with those obtained with TreeShrink (Mai and Mirarab 2018), the only other tool to our knowledge with a similar objective that could reasonably be applied on such a large dataset.

We show that PhylteR identifies outlier sequences with more precision and sensitivity than TreeShrink in most cases. For instance, only PhylteR correctly identifies species in gene trees whose phylogenetic placement is not in accordance with its placement in other gene trees, which can result from biological processes such as horizontal gene transfers or hidden paralogy. We also provide strong evidence that the automatic removal of outliers with PhylteR improves the concordance between gene trees and the species tree in greater proportions than TreeShrink (Mai and Mirarab 2018).
We hope that PhylteR could become the standard that was lacking (Philippe et al. 2017) for cleaning datasets prior to phylogenomic analyses.

Material and Methods

Description of the PhylteR method

The PhylteR method, in its entirety, is depicted in Figure 1. It starts with K distance matrices obtained from K genes by either computing pairwise distances (sum of branch lengths) between

species in each gene tree, or directly from each gene multiple sequence alignment (MSA). All the matrices are given the same dimensionality by filling missing data (if any) with the mean value across matrices, and are then normalised by dividing each matrix by either its median or its mean value (default is median). The normalisation by median prevents genes from fast- (resp. slow-) evolving orthologous genes to be erroneously considered as outliers, and appears as a better choice than a normalisation by the mean as it is less affected by outlier values.

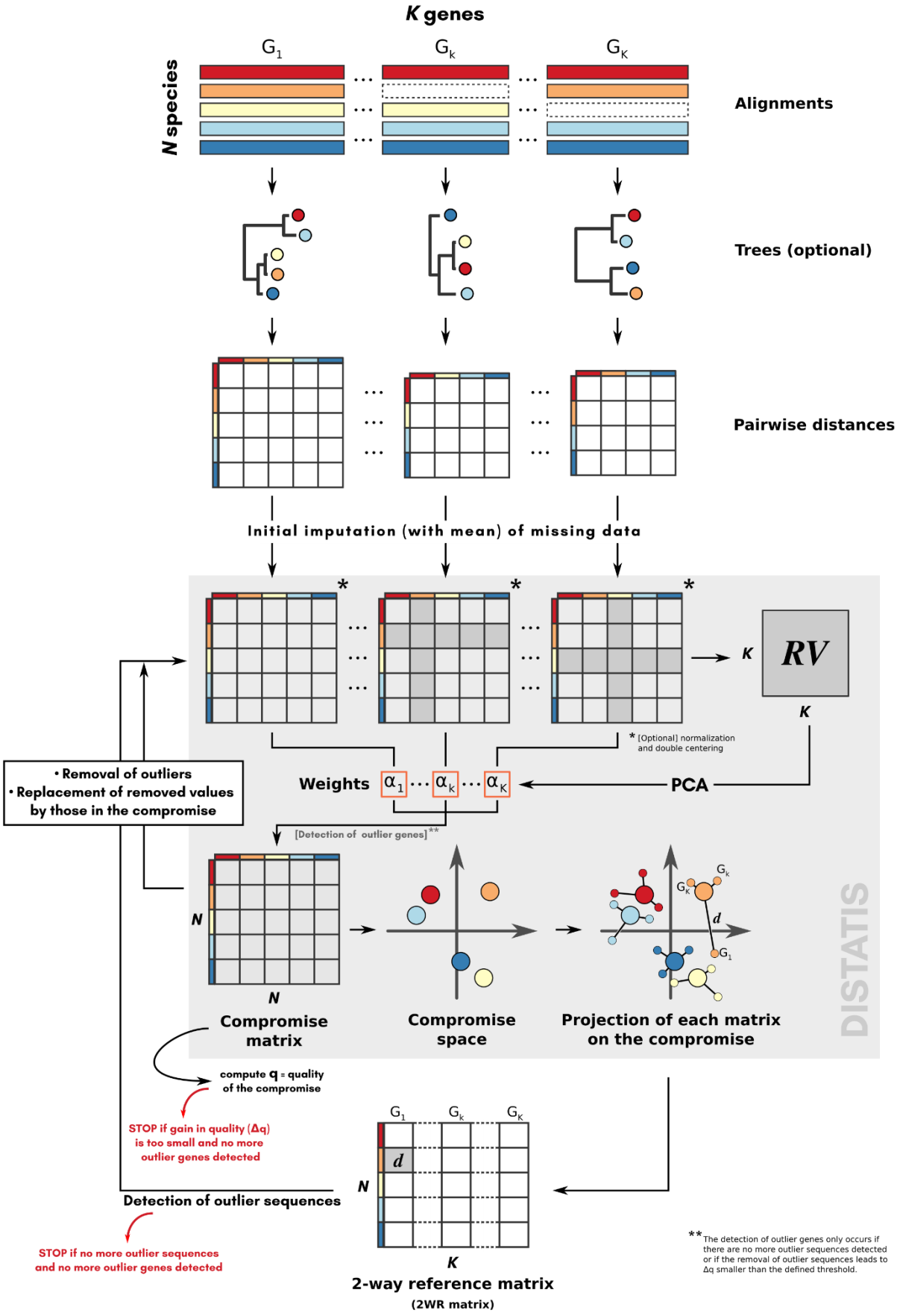

Figure 1. Principle of the PhylteR method for identifying outliers in phylogenomic datasets. The method relies on DISTATIS (grey block), an extension of multidimensional scaling to three dimensions. See text for the detail of the different steps.

From the *K* matrices obtained, an incremental process starts consisting in three main steps detailed in the next sections: (1) comparison of the matrices with the DISTATIS method (Abdi et al. 2005; Abdi et al. 2012), (2) detection of outliers sequences, and (3) evaluation of the impact of removing these outliers on the overall concordance between the matrices. Note that we refer to *outlier sequence* as a single gene for a single species (one sequence in one alignment, or one tip in one gene tree) that does not follow the general trend (i.e. other alignments or gene trees), while *outlier gene* refers to a complete alignment (or a complete gene tree) that does not agree with the other alignments (or gene trees).

These steps are repeated until no more outlier sequence is detected, or until the removal of the identified outlier sequences does not increase the concordance between the matrices more than a certain amount specified by the user. Before finishing the optimization, PhylteR performs a last action consisting in checking whether some outlier genes still exist despite the removal of outlier sequences already performed. These outlier genes correspond to single-copy orthologous genes for which the lack of correlation with others is not due to a few outlier sequences but are globally not following the trend. If outlier genes are discarded there, the optimization restarts as it may have unblocked the detection of other outliers.

Comparison of individual gene matrices with DISTATIS

DISTATIS is a multivariate method designed to evaluate the concordance between *K* distance matrices (*K* orthologous genes) measured on the same *N* species. The principle of DISTATIS is depicted in Figure 1 (grey box). The first step of DISTATIS consists in computing a matrix of RV coefficients (Robert and Escoufier 1976) that measures the similarities between the species pairwise distances present in each matrix. This can be seen as an extension of the correlation matrix (used in principal component analysis) that, instead of measuring the links between a set of variables, evaluates the relationships between a set of tables (gene distance matrices here). In a second step, a compromise distance matrix is built as the average of the *K* distance matrices weighted by the first eigenvector of the matrix of RV coefficients. The compromise represents the best consensus between the *K* distance matrices, as the weights used in the averaging procedure take into account the similarities between them (i.e., more similar distance matrices would have more weights in the definition of the compromise). In a third step, the compromise matrix is submitted to an eigendecomposition procedure so that species can be represented in a low-dimensional multivariate

space. In this compromise space, species are positioned so that their distances (computed in few dimensions, see after) represent the best approximations of the original distances contained in the compromise matrix. We used a broken stick model (Barton and David 1956) to estimate the number of dimensions (axes) of the compromise space, as this simple method was shown to give a good approximation of the correct dimensionality of the data with another multivariate approach (Jackson 1993). Then, each individual pairwise distance matrix is projected on the compromise space. This allows us to obtain a representation of species associated with each gene family. In other words, the compromise identifies the dissimilarities between species that are common for all genes whereas the projections of individual distance matrices allow to depict the peculiarities of each sequence. Lastly, we computed the distances, in the compromise space, between the position of a species given by all genes (the compromise) and its position associated to a particular gene family (using the projection procedure) and filled a gene x species 2-Way Reference matrix (2WR matrix, see figure 1) with these values.

Detection of outlier sequences from DISTATIS results

From the 2-Way Reference matrix (2WR matrix, see figure 1), we apply the method of (Hubert and Vandervieren 2008) to detect all values that are outliers, at the right of the univariate distribution of values. This method is an adjustment of the Tukey method (the classical boxplot) adapted to skewed distribution. In brief all values above

$$Q3 + ke^{3MC}IQR \quad (1)$$

are considered outliers. *Q3* is the 3rd quartile of the distribution, *IQR* is its interquartile range and *MC* is the medcouple of the distribution (Brys et al. 2004), a measure of skewness bounded between -1 (left skewed) and +1 (right skewed). The *k* value is chosen by the user (default is 3), and controls how stringent the detection of gene outliers is. Small values of *k* lead to more gene outliers being detected. The detection of gene outliers is performed after normalisation of the 2WR matrix, achieved by dividing each row (the default) or each column by its median. This normalisation leads to an exaggeration of outlier values, easing their identification.

Detection of outlier genes

When no more outlier sequences are found in the 2WR matrix, PhylteR checks whether some genes are still uncorrelated to others. These outlier genes are detected by finding outlier values in the weight array (α_1, α_2, ..., α_K, see Figure 1). The outlier detection method used is the same than for the

outlier sequences of the 2WR matrix (Equation 1) but its stringency can be tuned independently (with parameter *k2* in place of parameter *k* in Equation 1, defaulting to $k2 = k = 3$).

Exit criteria of the PhylteR iterative process

PhylteR is an iterative process (see Figure 1) with two exit points. The first one is straightforward: if no more outlier sequences are detected in the 2WR matrix, and if no more outlier genes exist (see above), then the process stops. The second one is based on the gain (Δ_q) achieved by removing outlier sequences (i.e. the change in q, the quality of the compromise). If this gain is below a certain threshold (10^{-5} by default), and if no more outlier genes exist, then the process stops.

Evaluation of the PhylteR method

Datasets

We used two types of datasets to evaluate PhylteR and compare it with TreeShrink: a collection of simulated simple examples, and a large Carnivora phylogenomic dataset with 53 species (Allio et al. 2021). These datasets are described below.

- *Simulated datasets (SD1 to SD4):* We generated a collection of simple datasets in order to either illustrate the different steps of the PhylteR process (SD1), or to compare PhylteR and TreeShrink in terms of their ability to detect misplaced species in gene trees (SD2), long-branch outliers in gene trees (SD3) or both types of outliers when mixed (SD4).
  These simulated datasets were obtained with the `simtrees()` function in the R package *phylter* (this publication). The way this function works is as follows: a single phylogenetic tree with a given number of species (*Nsp*) is randomly generated with function `rtree()` from package *ape* v5.6.2 (Paradis and Schliep 2019). This tree is duplicated as many times as the number of orthologous gene families required (*Ngn*). To add variance to branch lengths, a value sampled in a normal distribution with mean 0 and standard deviation *brlen.sd* is added to each branch length of each tree. If the resulting branch length is negative its absolute value is taken. The number of outliers sequences present in the dataset (*Nb.cell.outlier*) and the type of outlier (*out.type*) is chosen. If outliers of type “topology” are simulated, outlier sequences are generated by randomly sampling *Nb.cell.outlier* times a species in a gene tree and moving it to another random location. If outliers sequences of type "brlength" are simulated, outliers are generated by randomly sampling *Nb.cell.outlier* times a species in a gene tree and multiplying its branch length by *bl.mult*. Finally, if both types of outliers are simulated, half of the outliers are assigned to type "topology" and the other half

to type "brlength". In case of an odd number of outliers, an extra outlier of type "topology" is simulated. Table 1 gives the parameters chosen for generating each one of the four datasets (SD1 to SD4). Each dataset was simulated 20 times in order to compute the variance in precision and sensitivity of outlier detection with the two outlier detection methods tested.

Datasets	Species (*Nsp*)	Gene families (*Ngn*)	Std. dev. of branch lengths (*brlen.sd*)	Outliers (*Nb.cell.outlier*)	Type of outliers (*out.type*)	Branch length multiplier (*bl.mult*)
SD1 (x20)	20	25	0.15	10	topology	-
SD2 (x20)	40	100	0.15	10	topology	-
SD3 (x20)	40	100	0.15	10	brlength	20
SD4 (x20)	40	100	0.15	20	both	20

Table 1. Parameters used for the simulation of the four simple example datasets used for PhylteR and TreeShrink comparisons.

- *Carnivora dataset (CD):* We used the raw sequence files (before alignment and filtering) from a previously assembled phylogenomic dataset comprising 14,463 genes for 53 species aimed at resolving the phylogeny of the order Carnivora (Allio et al. 2021). This dataset was obtained by extracting single-copy protein-coding orthologous genes from the genomes of 52 carnivore species, plus the Malayan pangolin (*Manis javanica*) used as outgroup, following the orthology delineation strategy of the OrthoMaM database (Scornavacca et al. 2019). These raw sequence files were aligned and filtered using the OMM_MACSE pipeline (Ranwez et al. 2021), which combines (i) translated nucleotide sequence alignment at the amino acid level with MAFFT (Katoh and Standley 2013), (ii) nucleotide alignment refinement (based on amino acid alignment) with MACSE v2 (Ranwez et al. 2018) to handle frameshifts and non-homologous sequences (Ranwez et al. 2018), and (iii) masking of ambiguously aligned and dubious parts of sequences with HMMcleaner (Di Franco et al. 2019). In the original study (Allio et al. 2021), this Carnivora dataset has been successfully filtered using an early version of PhylteR allowing the removal of outlier sequences and genes generating abnormally long branches. Therefore, it was a good candidate dataset to test the completely redesigned and improved version of PhylteR presented here.

Evaluation of the accuracy of PhylteR outlier detection and comparison with TreeShrink

We evaluated PhylteR's ability to detect outliers that are either correct (when it is possible to test it, with simulated datasets) or meaningful according to the biological information we can gather from the dataset at hand.

We used the first simulated dataset (SD1) for illustration purposes only. For the simulated datasets SD2 to SD4 (Table 1), and for each one of the 20 replicates, we ran PhylteR with default parameters and we counted the number of True Positives (TP, outlier sequences that were simulated and that are retrieved), False Positives (FP, outlier sequences that were not simulated but are identified) and False Negative (FN, outlier sequences that were simulated but are not retrieved). From those, we computed the mean precision (TP/(TP+FP)) and recall (or sensitivity, TP/(TP+FN)) of the outlier identification of PhylteR. For comparison purposes, we performed the same analyses using TreeShrink v1.3.9 (Mai and Mirarab 2018) in place of PhylteR with default parameters for detecting outliers.

For the Carnivora dataset, we have no access to the *true* outliers. It is thus impossible to compute precision and recall on this empirical dataset as done on the simulated ones. Instead, we can compute "features" associated to each gene sequence for each species (*sequence* hereafter), that are, *a priori*, associated with errors or with lack of signal in phylogenomic datasets. We can then evaluate whether the outliers detected by PhylteR are enriched in extreme values for these features, as compared with randomly selected sequences or with outliers identified with TreeShrink. The list of features and the reason for their choice is listed below.

- **Sequence length:** Long sequences were shown to carry more phylogenetic signal than shorter ones (Salichos and Rokas 2013; Shen et al. 2016). To explore the possible enrichment of outliers in short sequences, we computed the length (in bp) of each sequence in each gene MSA, and explored its distribution in outliers.
- **Duplication score**: when a sequence in a gene tree is not orthologous to the others but is a paralog, its localization in the gene tree is likely to be incorrect. To have an insight into the level of "paralogousness" of each sequence in the Carnivora dataset, we compared the Carnivora species tree published in Allio et al. (2021) with each one of the 14,463 gene trees using the reconciliation program *ALEml_undated* (Szöllősi et al. 2015). This tool allows inferring the duplications, losses and transfers experienced by a gene by comparing its history (the gene tree) with that of the species (the species tree). Here we inferred only duplications and losses (transfer rate was forced to be 0), we forced the origination of each gene at the root of the species tree (parameter O_R=10000) and we used default value for all other parameters. We then computed the number of duplications inferred from the root to each tip of each gene tree, and normalised this value by the number of nodes encountered.

This value represents the normalised number of duplications experienced by each sequence, whose distribution in outliers could be evaluated.

- **Hidden paralogy, the KRAB Zinc finger (KZNF) protein family case:** The KZNF super-family is actively duplicating in vertebrates with hundreds of paralogs per genome (Huntley et al. 2006; Liu et al. 2014). Thus, the orthologous relationships between these proteins is expected to be hard to retrieve and the reconstructed orthologous gene families are likely to contain hidden-paralogs. If an outlier detection method is indeed able to remove hidden paralogs, we should see an enrichment of KZNF genes in the list of outliers.
- **Synteny:** Synteny (in our sense) is the link between two genes occurring consecutively on a genome, *i.e.* without any other gene (in the dataset) located between them. One gene then has two synteny linkages. A synteny *break* occurs when two genes are consecutive in one species but their orthologs in another species are not. The direction of transcription (coding strand) is considered, *i.e.* if it has changed it is considered as a break even if the genes appear in the same order. One gene, compared to its ortholog in another species, may then be associated with 0, 1 or 2 breaks. We call genes associated with 2 breaks *syntenic outliers*. Synteny breaks are due to genomic rearrangements (inversions, duplications, translocations, …), but can occur in the data, and in much larger proportion, for many artifactual reasons: annotation errors, assembly errors, or orthology assessment errors. We thus formulate the hypothesis that outlier genes may be more often associated with synteny breaks than randomly sampled genes. To evaluate it, we focused on 14 Carnivora genomes (Table S1) that we compared in a pairwise manner. For each pair we compared the list of syntenic outliers with the list of outliers retrieved by each outlier method tested, and we computed the p-value associated with the observed size of the intersection under the hypothesis that the two sets of outliers are independent.

In order to compare the distributions of values for the different features listed above between outlier detection methods, we needed lists of outliers of comparable size. The number of outliers retrieved with default parameters being very different with the two methods using default parameters (7,183 with PhylteR *vs* 19,643 with TreeShrink, see Table 2), we created two collections of outliers, a **small** and a **large** one (Table 2). For the **small** collection, we selected a value for the parameter q in TreeShrink in order to get a number of outliers as close as possible to the number of outliers obtained with PhylteR default parameters. This was achieved for $q = 0.012$, leading to 7,032 outliers. For the **large** collection, we selected a value of the k (and k = $k2$) parameter in PhylteR leading to a number of outliers as close as possible to the number of outliers detected with TreeShrink default parameters. This was achieved for $k = 1.55$, leading to 20,157 outliers.

Parameters used and number of outliers in each collection and with each outlier detection method are presented in Table 2.

	PhylteR		TreeShrink		Random
Collections	Parameters	# outliers	Parameters	# outliers	# outliers
small	*default*	7,183	q = 0.012	7,032	7,183
large	k = k2 = 1.55	20,157	*default*	19,643	20,157

Table 2. Collections of outliers used to evaluate PhylteR and compare it to TreeShrink. The **small** collection is obtained by tuning the TreeShrink parameters in order to obtain roughly the same number of outliers as with the default parameters of PhylteR. The **large** collection is obtained in the opposite way.

Evaluation of the impact of outlier sequences removal on species tree support

It is expected that a tool that accurately removes outliers in phylogenomic datasets should increase the concordance between the gene trees and the species tree. To evaluate this and compare PhylteR with randomly sampled sequences and with TreeShrink-identified outliers, we computed the gene concordance factor (gCF, Minh, Hahn, et al. 2020) as implemented in IQ-TREE version 2.1.3 (Minh, Schmidt, et al. 2020) for every branch in the Carnivora species tree (obtained from Allio et al. 2021). For each branch of the species tree, this factor indicates the percentage of gene trees in which this branch is found (among gene trees where this can be computed, or "decisive" trees, see Minh, Hahn, et al. 2020). gCF was computed according to either the original gene trees (gCF_{init}), or to a list of gene trees obtained after pruning outliers (four sets of gene trees corresponding to the four list of outliers in Table 2).

In order to see the effect of outliers removal on the concordance factor, we computed the difference (ΔgCF) between gCF_{init} and every other gCF, separating the small and the large collections of outliers. Positive values of ΔgCF indicate that a branch is more supported after filtering than before. Comparing ΔgCF between PhylteR and TreeShrink gives an indication of whether, for the same total number of outliers removed, PhylteR performs better than TreeShrink at identifying problematic sequences and increasing the concordance between the species tree and the gene trees.

Results

Illustration of the general principle of PhylteR

The different steps of the PhylteR process (Figure 1) are illustrated on a simple example dataset comprising 25 genes for 20 species, with 10 outliers (Table 1). The main steps are as follows. Individual gene trees are transformed into individual gene matrices that are then combined into a unique *compromise* matrix obtained after weighting each matrix by its concordance with the others: matrices that are poorly correlated with the others have less weight in the creation of the compromise (Figure S1A-E). This matrix is then projected onto a space on which individual matrices are projected as well (Figure 2A and S1F). By computing the distance of each species in each orthologous gene to its reference position in this projection, the two-way reference matrix is obtained (Figure 2B and S1G). It is from this matrix that outlier sequences can be identified and removed.

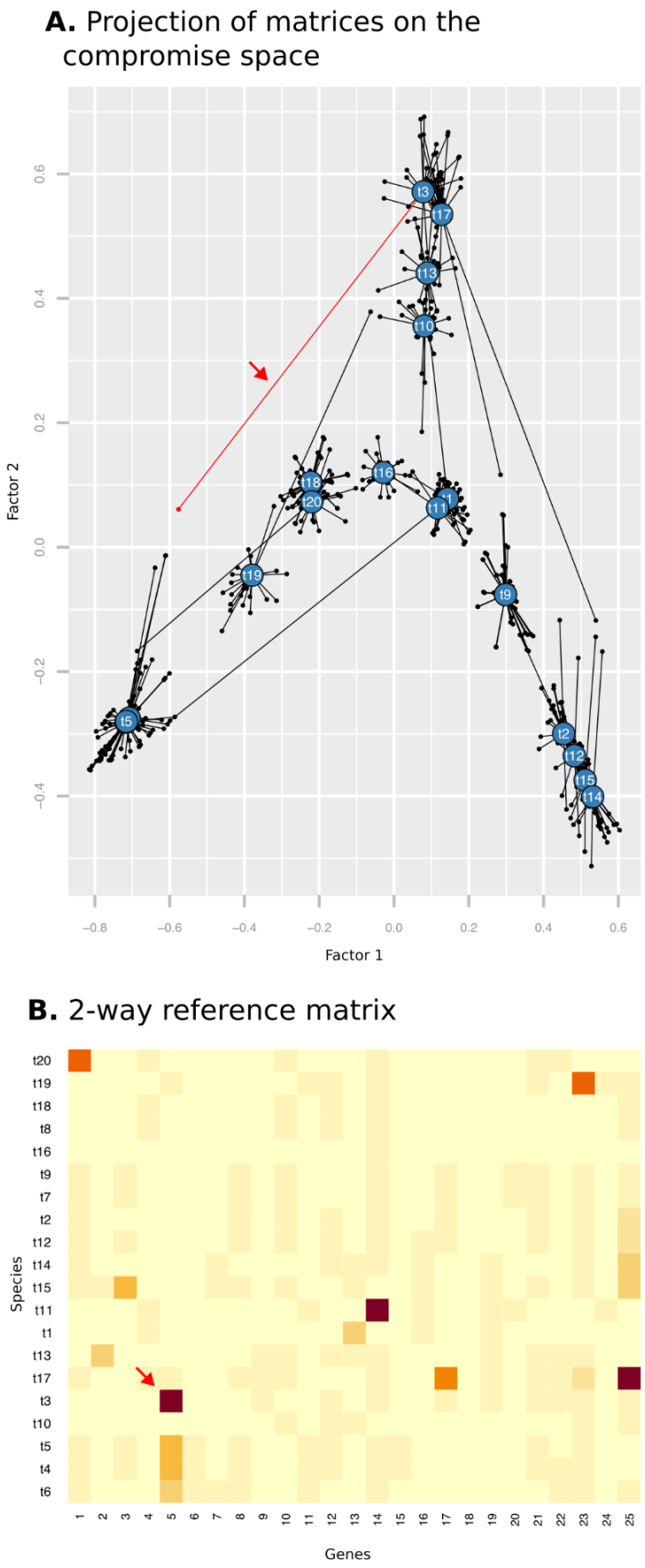

Figure 2. Two objects of the PhylteR process. A: the compromise matrix is projected into a multidimensional space (the two first axes only are represented here). This gives the reference position of each species relative to each other (blue badges with species names on it). Individual gene matrices are projected on the same space (small dots) and the distance between each gene in each species to its reference position is represented by a line. The red line and the red arrow identify species t3 in gene 5. This projection is transformed into a 2D matrix (**B**) by computing the distance between each species in each gene to its reference position (i.e. the length of each line in **A**). The gene × species matrix obtained, that we refer to as the 2-way reference matrix (2WR) is used to detect outliers like the one indicated by the red arrow, corresponding to the red arrow in **A**.

PhylteR performs well on simple examples with all types of outliers

To evaluate the precision and sensitivity of PhylteR, we used it on three simplistic datasets (SD2, SD3 and SD4, table 1). We also computed precision and recall on the same datasets using another method, TreeShrink (Mai and Mirarab 2018). The SD2 simulated dataset contained only outliers of type "topology", i.e. outliers obtained by moving some species to another location in some gene trees. For this type of outliers, PhylteR performs very well, precision and recall being very close to their maximum value 1 (Figure 3, left). On the other hand, TreeShrink performs very badly, usually identifying no correct outliers at all (leading to a mean precision close to 0), but still detecting a large collection of false positives (leading to a low sensitivity). When outliers are obtained by increasing branch lengths for some species in some genes (the SD3 datasets), PhylteR is still performing very well in terms of precision (mean = 1, Figure 3, middle), and better than TreeShrink, mainly because TreeShrink detects many false outliers. In terms of recall now, PhylteR appears slightly lower than TreeShrink. Part of this lower performance, however, can be due to the fact that TreeShrink detects many more outliers than PhylteR. Another reason is that TreeShrink is specifically designed to detect outliers with long branches and that for this specific task, despite many false positives, it seems more sensitive than PhylteR. Finally, when both types of outliers are mixed in a dataset (SD4), PhylteR outperforms TreeShrink (Figure 3, right). This is easily explained by the results of the tests on the two previous datasets.

These simulated datasets are quite simple, but allow us to better understand the main differences between PhylteR and TreeShrink: only PhylteR can detect misplacement of species in gene trees, which is a primordial aspect of outliers in phylogenomic datasets, and PhylteR is much more precise (low number of false positives on the evaluated simulated datasets).

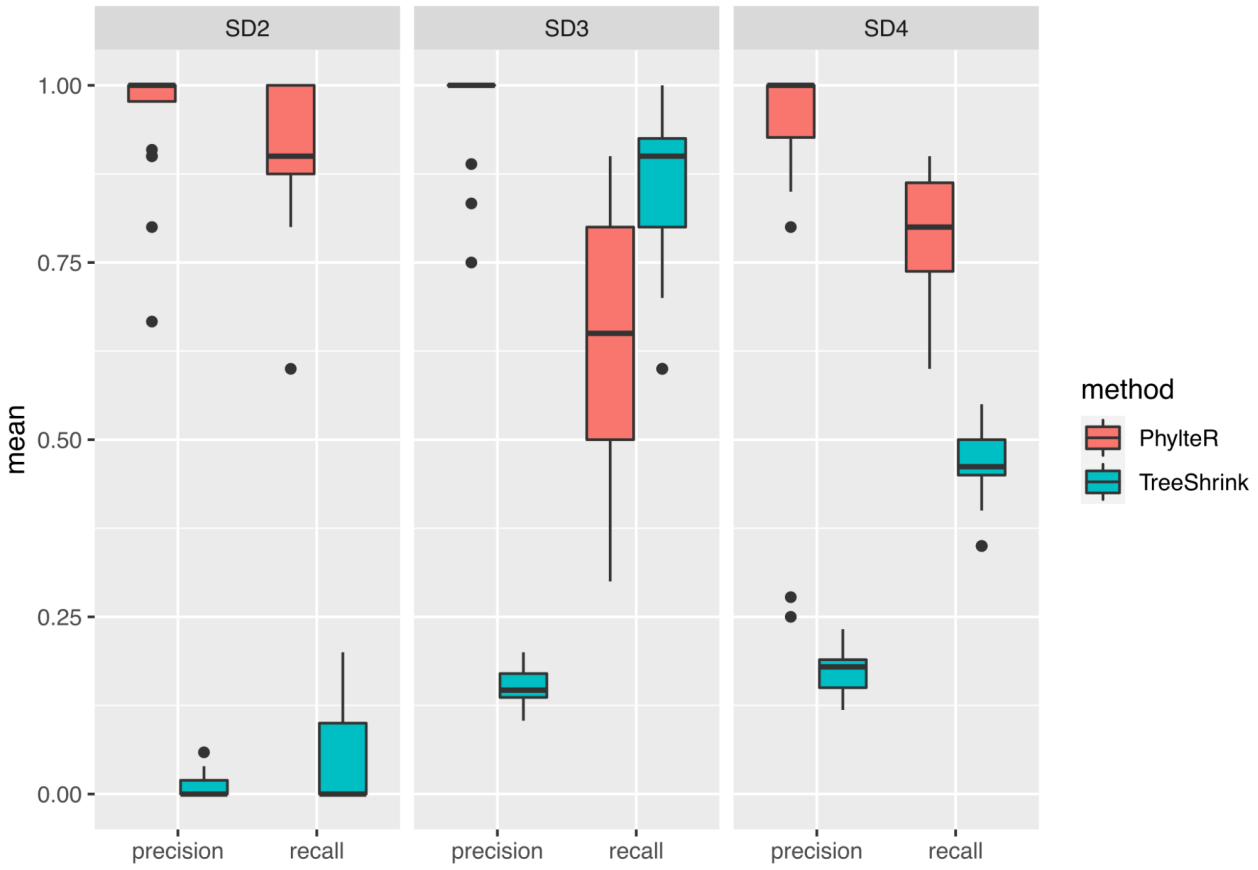

Figure 3. Comparison of the precision and recall (or sensitivity) of the PhylteR and the TreeShrink outlier detection methods for three simulated datasets. SD2: simulated dataset with topological outliers; SD3: simulated dataset with long-branch outliers; SD4: simulated dataset with outliers of both types. The boxplots represent the variance across 20 replicates.

Characterisation of outliers detected with PhylteR on the Carnivora dataset

Outliers in phylogenomic datasets can be of different nature: fast or slow evolving genes in some species, leading to respectively long or short branches in gene trees, or species being placed in aberrant position in some genes because of incomplete lineage sorting (ILS), horizontal gene transfers (HGT), hidden paralogy, saturated signal, compositional bias, long-branch attraction, or other artifactual reasons (Schrempf and Szöllősi 2020).

In the set of 14,463 gene trees analysed by PhylteR, two sets of outliers (7,183 and 20,157 sequences) were identified with PhylteR (with default or tuned parameters, respectively) and 7,032 and 19,643 with TreeShrink (with tuned and default parameters respectively, see Table 2). A simple comparison of the list of outliers of similar sizes revealed that the overlap between the two lists of outliers was quite small (around 20%, Figure 4). This corresponds to about 70% of the outliers detected by PhylteR (resp. TreeShrink) being absent from the list of outliers detected by TreeShrink (resp. PhylteR). This confirms fundamental differences between the two approaches.

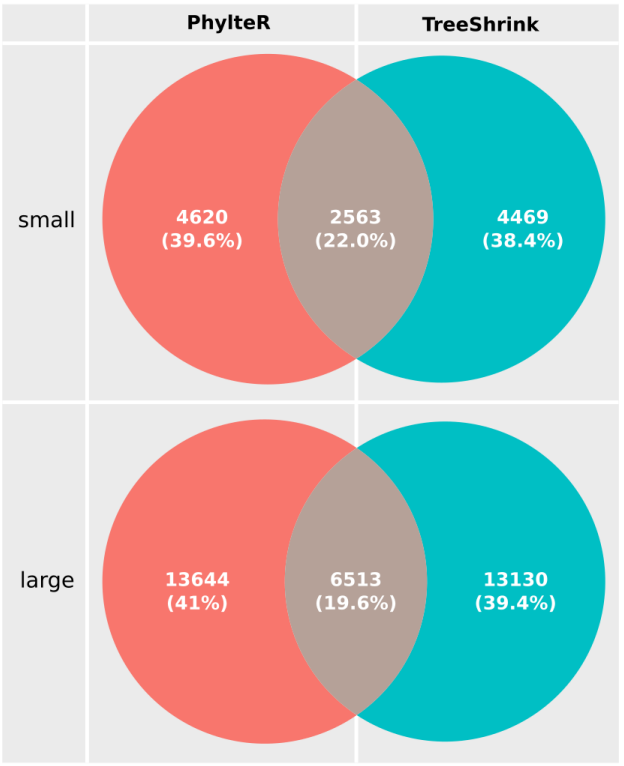

Figure 4. Comparison of the sets of outliers detected by PhylteR (left column) and TreeShrink (write column) on the Carnivora dataset. The two collections of outliers (small and large) correspond to different stringency for the detection of outliers (see Table 2).

To better understand what differs between the outliers detected by PhylteR and those detected by TreeShrink, we compared the distribution values of different features describing these outlier sequences.

First, we observed a significant decrease in sequence length in outlier sequences for both PhylteR and TreeShrink as compared to randomly sampled sequences ($p<2.2e-16$ in both cases and for both collections of outliers, Figure 5A). Sequence lengths were higher in PhylteR outliers than in TreeShrink outliers for the small collection of outliers ($p<2.2e-16$) but the opposite was observed for the large collection of outliers ($p<3.17e-14$). The fact that outliers are enriched in short sequences is thought to be due to the expected correlation between the size of a sequence and the phylogenetic signal it carries. Shorter sequences are more prone to misplacement in phylogenetic trees.

Second, we compared the distribution of duplication scores in the list of outliers produced by PhylteR and TreeShrink (Figure 5B). We observed a clear difference, for both the small and the large collections of outliers between PhylteR outliers and random outliers, but also between PhylteR outliers and TreeShrink outliers: outliers identified by PhylteR are significantly enriched in

sequences that display a higher number of duplications as compared to random or TreeShrink outliers ($p<2.2e-16$ for all comparisons).

This result is in accordance with the results obtained on simulated datasets: PhylteR is good (and much better than TreeShrink) at identifying misplaced species in some gene trees, which is indirectly what the duplication score captures.

One illustration of the difference between PhylteR and TreeShrink in their ability to capture duplicated sequences (and thus probably hidden paralogous) can be given by the study of peculiar proteins, such as the Zinc-finger family (ZNF). This large family of paralogs first duplicated from the gene PRDM9 or PRDM7 in the ancestor of vertebrates (Emerson and Thomas 2009). These genes are involved in the repression of transposable elements and are still actively duplicating. The high number of duplications renders the resolution of the orthology relationship in this gene super-family very challenging. In the Carnivora dataset, the ZNF super-family has been splitted in 168 orthologous gene families (Allio et al. 2021). As expected in case of hidden paralogy, we see an overrepresentation of the genes belonging to these families in the list of outliers, especially in the outliers identified by PhylteR (Figure 5C). Between 3.79% (for the large set) and 7.4% (for the small set) of PhylteR outliers belong to the ZNF family, while these values drop to 1.78% and 1.12% respectively for TreeShrink outliers, and less than 1% for randomly selected sequences (Figure 5C).

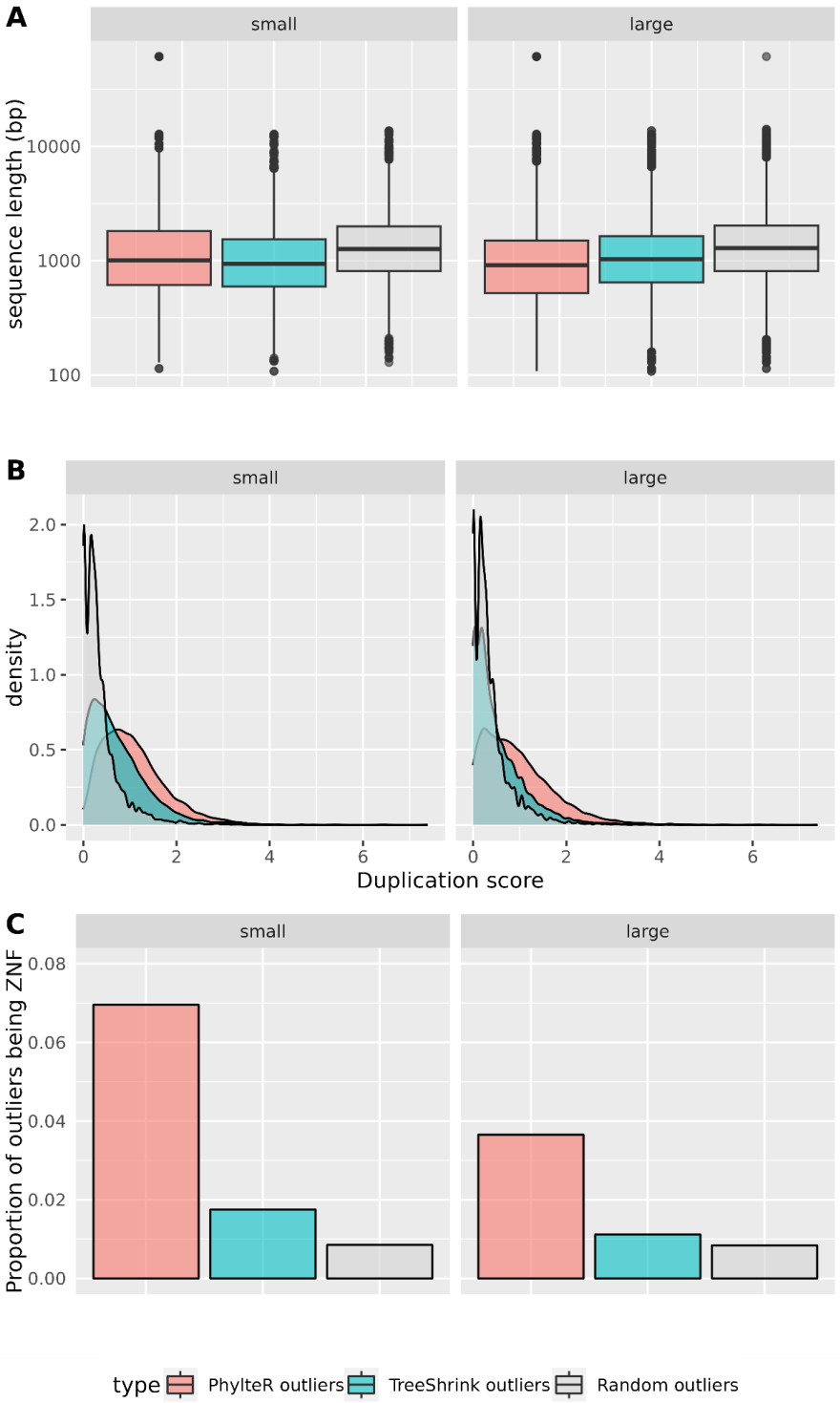

Figure 5. Comparison of distribution values between outliers detected by PhylteR, by TreeShrink, or randomly sampled, for three features associated with outlierness in phylogenomic datasets. A. Distribution of the length (in bp) of the sequence outliers identified by each method. A log scale is used for the y-axis. **B.** Distribution of duplication scores (normalised number of duplications experienced by each sequence) for the outliers identified by each method. **C.** Proportion of outliers being members of the KRAB-ZNF protein family for the outliers identified by each method. The two collections of outliers (small and large) are compared in each case (left and right on each panel).

Third, we compared two by two 14 Carnivora species and identified syntenic outliers (see material and methods). In almost all pairwise comparisons, we found that these syntenic outliers

significantly overlap the outlier sequences detected by Phylter. For example, in the comparison between *Zalophus californianus* and *Suricata suricatta* (illustrated in Figure 6), out of the 5,123 genes common to both species in the dataset, 131 (2.56%) are syntenic outliers (i.e. surrounded by two breaks). In comparison, out of the 47 outlier sequences identified by PhylteR (small list) in either *Zalophus californianus* or *Suricata suricatta*, 38 are syntenic outliers (80.8%), which is significantly more than expected by chance (p-value = 1.5e-43). With TreeShrink (small list) for the same pair of species, only 18.1% (17 out of 94) outlier sequences are syntenic outliers, which is much less than with PhylteR but is still significantly different from what is expected by chance (p-value = 1.36e-10). Similar results were obtained for most of the other pairs of species compared (Figure S2 and Supplementary Tables S2 and S3).

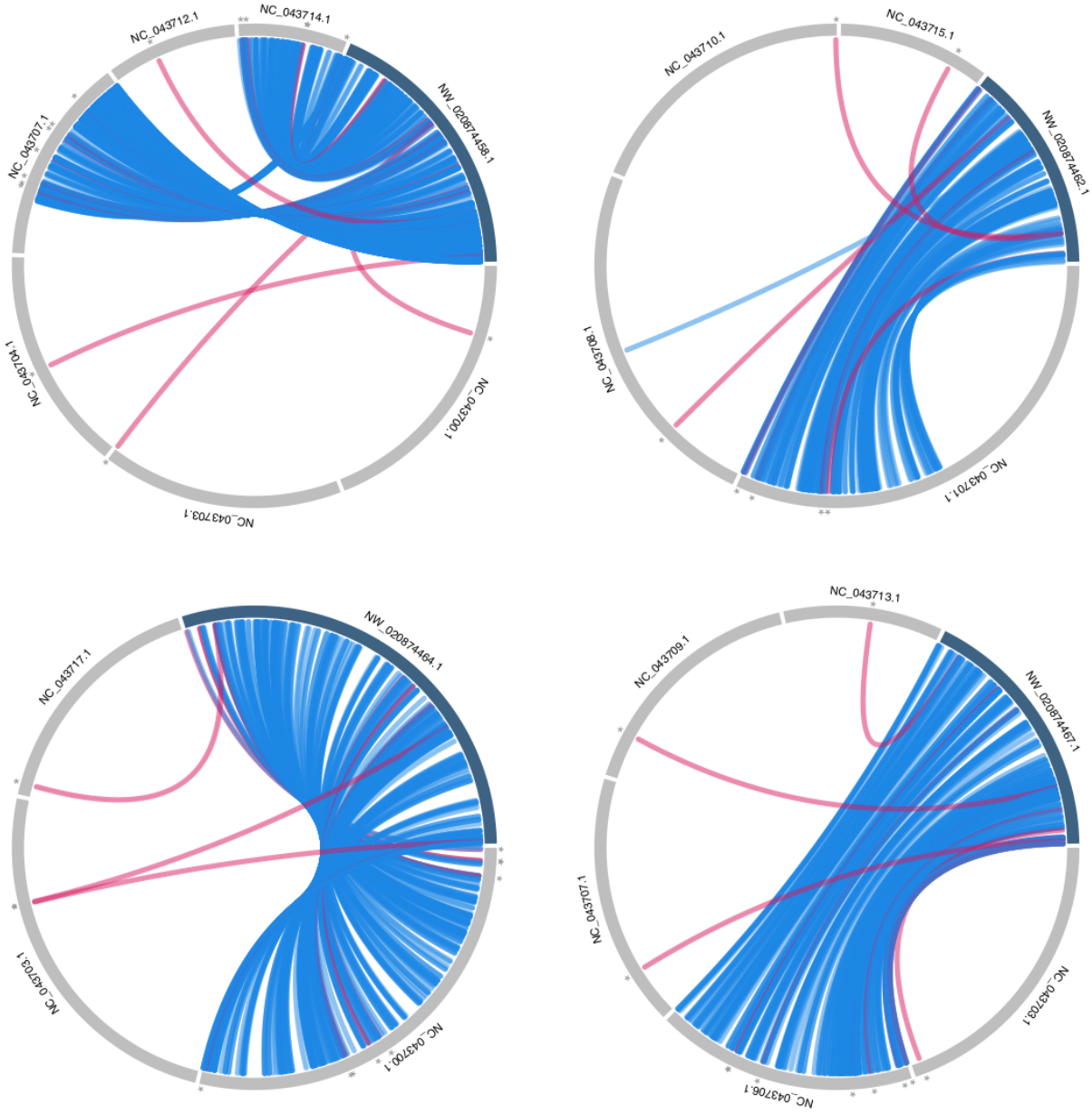

Figure 6: Illustration of the non-syntenic nature of many outliers identified by PhylteR. We represent the comparison of *Zalophus californianus* with *Suricata suricatta* genomes, with Zalophus as a reference (arbitrarily, most other pairs of

species give similar results). On each circle, a reference Zalophus scaffold is represented in dark blue, and all scaffolds for which at least one gene has an ortholog in this scaffold are in light grey. Lines between these scaffolds represent couples of genes annotated as orthologous. Red lines highlight gene outliers detected in *Suricata suricatta*. We observe that they are very often "isolated" genes, *i.e.* syntenic outliers. These genes are thus probably erroneously annotated, erroneously assembled, and their orthology is likely erroneous.

Impact of filtering outliers on Species Tree support

The gene concordance factor (gCF) is a measure, for a species tree, of how much each one of its branches is supported according to a collection of individual gene trees. A value of 100% means that 100% of the gene trees for which the comparison could be done ("decisive" gene trees in Minh, Hahn, et al. 2020) display this branch.

Non-random outlier removal processes are expected to increase gCF scores by discarding sequences representing species in gene trees whose position is not in accordance with their placement in the other gene trees. We looked at the difference in gCF score before and after pruning outliers (ΔgCF) for each branch of the Carnivora species tree. For both PhylteR and TreeShrink, an increase in gene concordance was observed. It was higher with PhylteR than with TreeShrink, indicating a better identification of misplaced species in gene trees for PhylteR. The effect was larger when more outliers were removed (Figure 7, right), the gain in gCF reaching more than 6% for some branches with PhylteR outliers removal (max 5% for TreeShrink). We observed that the gain in concordance was higher for branches that initially had a high gCF, and smaller for poorly supported nodes (plain dots versus circles in Figure 7).

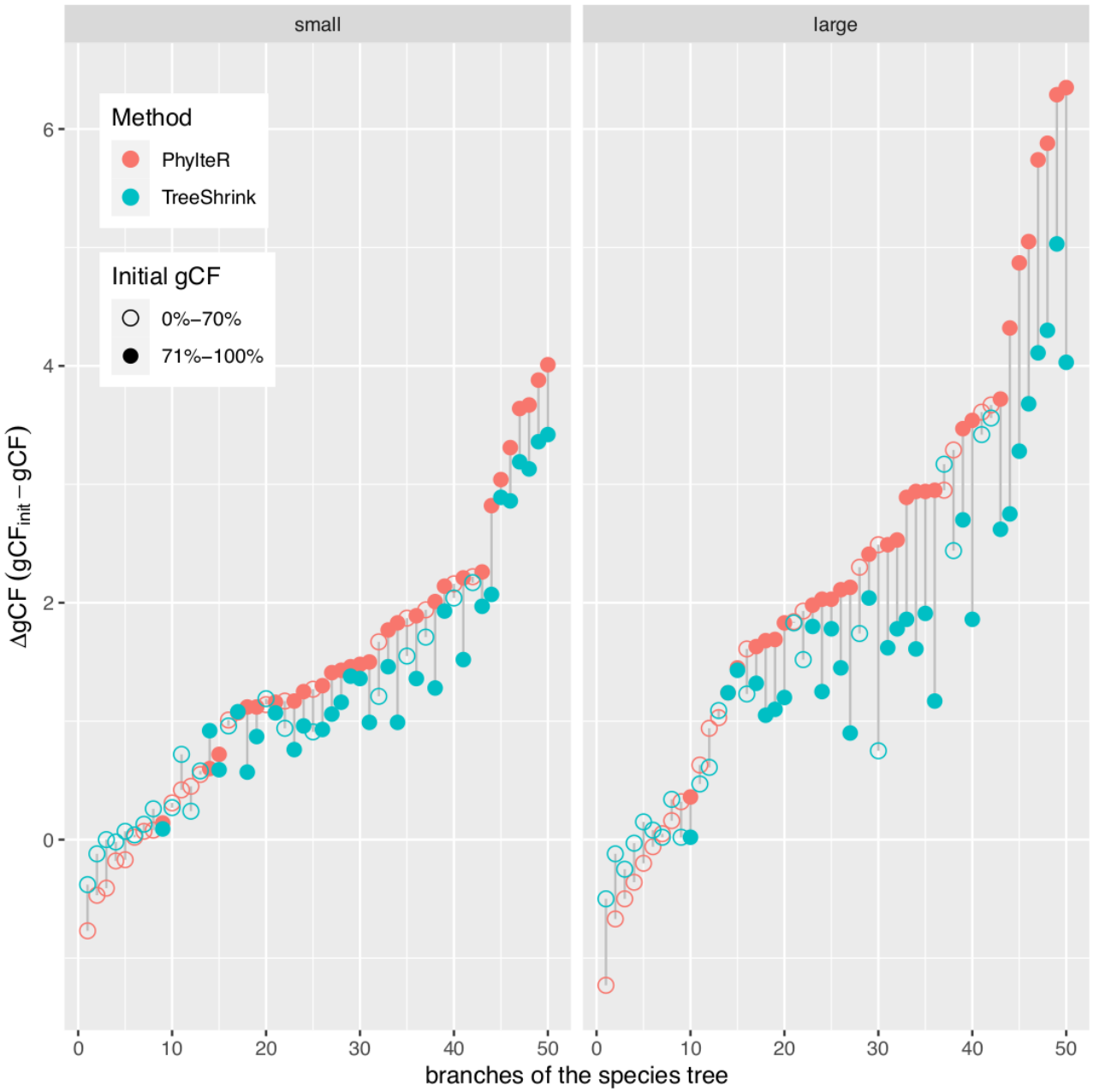

Figure 7. Effect of filtering outliers in gene trees on the gene concordance factor (gCF) of each branch of the Carnivora species tree. The gain in concordance (ΔgCF, y-axis) is plotted for each branch of the species tree (dots), separating PhylteR (pink) and TreeShrink (blue). Branches are ordered by increasing ΔgCF for the PhylteR outliers. The results for the two collections of outliers (small and large) are displayed side by side.

Discussion

In phylogenomics, incongruence between gene trees, resulting from a myriad of possible technical and analytical issues, or from biological processes, is known to lead to errors in species tree inference (Philippe et al. 2017). A common practice in phylogenomics thus consists in scanning individual gene trees by eye, trying to spot "problematic" branches (i.e. species or group of species weirdly placed in gene trees, suspicious long branches, apparent groups of paralogues, etc.) and

discard them prior to the concatenation of the genes (supermatrix approach) or to the combination of the gene trees into a species tree (supertree approach). This hard work is not only time-consuming and laborious, it is also questionable: what is the objectivity in this practice? Is the eye (and the brain) capable of looking at tens of thousands of gene trees at the same time? How reproducible is such a practice? Etc.

Here, with PhylteR, we propose a way of analysing large collections of gene trees by using an automatic method that can simultaneously analyse a large collection of distance matrices (retrieved from gene trees or directly from MSA), identify the common signal between these matrices, and identify elements (outliers) in some of these matrices that are responsible for a decrease in concordance. By using a process where these outliers are automatically and iteratively removed, we propose a new way of efficiently identifying them.

Evaluating a method for its capacity to accurately identify errors in phylogenomics datasets is a difficult task. As for any inference method, we use simulations. However, simulating the processes that result in errors (in our case, outliers in phylogenomics data) has no standard solution: sources of errors are numerous, they combine with each other through all phylogenomic pipelines, sometimes with unpredictable results. So we restricted ourselves to simulating the features intrinsically detectable by PhylteR, that is, changes in branch lengths and topology. Further evaluation would involve an independent simulation pipeline, not informed by the hypothesis behind the inference method (Biller et al. 2016), which is by definition outside the scope of the description of the inference method. The simple simulations we performed revealed that outliers corresponding to misplacement of species in a few gene trees was easy to detect by PhylteR but not by TreeShrink. It appeared, however, that detection of outliers corresponding to long branches in some gene trees (without changes in topology) was slightly more sensitive with TreeShrink than with PhylteR, even though precision was very low with TreeShrink (many false positives).

The way we evaluated PhylteR and compared it with TreeShrink was by looking at some properties associated with gene sequences, and testing possible enrichment of these properties in the list of detected outliers. We observed an enrichment of short sequences, which was anticipated (short sequences carry less phylogenetic signal) and confirmed previous results (Shen et al. 2016).

A notable difference that we observed between PhylteR and TreeShrink, confirming the results on the simple simulated examples, is the duplication score computed here: outliers identified with PhylteR seemed to be highly enriched in gene sequences having experienced more duplications, according to the reconciliation analysis performed. Note, however, that we need to be cautious with this measure: being based solely on a topological comparison between gene and species trees, it cannot distinguish between true paralogy, and other processes (biological or artefactual) leading to a

species in a gene tree to have a position that is not concordant with its position in the other gene trees. Horizontal gene transfers (HGT) or Incomplete Lineage Sorting (ILS), for instance, may lead to high duplication scores according to our approach when none occurred (even though HGT is thought to be anecdotal in the carnivora dataset). Similarly, artefactual reasons such as long branch attraction, annotation error or alignment error can lead to misplacements of species in some gene trees.

A more direct way of testing the ability of PhylteR to detect hidden paralogous sequences was to focus on a specific gene family known to be extremely diverse because of multiple duplication events, the KZNF family (Huntley et al. 2006; Liu et al. 2014). We observed a clear enrichment of sequences belonging to this peculiar family in the list of outlier sequences identified by PhylteR, as compared to those identified by TreeShrink or randomly sampled. This capacity of PhylteR to identify putative paralogs is an important feature, as it was shown earlier that non-orthologous sequences in phylogenomic datasets could have drastic impact on results (Philippe et al. 2017), leading for instance to erroneous branching with high support in the reconstructed species tree in some cases (Philippe et al. 2011).

A final argument that we used to validate PhylteR consisted in exploring the syntenic nature (and lack thereof) of the sequences identified as outliers when comparing the species in a pairwise manner. We observed that outlier sequences were often (much more than expected by chance) syntenic-outliers, i.e. sequences associated with a loss of synteny when comparing the two genomes. This provides two kinds of information: on one side, that the “syntenic outliers” and the “phylogenetic outliers” largely overlap, which proves with an argument orthogonal to all the previous ones, that PhylteR (and TreeShrink to a lesser extent) captures an information about erroneous annotations; on the other side, it suggests that many “syntenic outliers” are due to errors and not to biological processes. “Syntenic outliers” are often filtered out before performing rearrangement analyses, because their position is believed to be artefactual (Lucas and Crollius 2017). However sometimes this outlier position is modelled as the result of a biological process (Dalevi and Eriksen 2008). Our analysis supports this artifactual origin in Carnivora, though some syntenic outliers might proceed from retrotranscription or translocations.

Here we focused on the importance of identifying outlier sequences in phylogenomic datasets in order to remove them prior to phylogenetic inference with supermatrix or supertree methods. But other usage of the tool we present here can be anticipated. For instance, correctly identifying and removing outlier sequences from multiple sequence alignments is crucial when using statistical methods based on the ratio of nonsynonymous over synonymous substitution rates (d_N/d_S ratio) to detect adaptive molecular evolution (see Yang and Bielawski 2000 for a review), or for correctly inferring ancestral sequences (Yang et al. 1995) from sequences of extant species.

Finally, using a tool like PhylteR is not only useful for cleaning the data. The in-depth exploration of the outliers detected and the study of the reasons why they were detected as such can give important insights into the evolutionary history of these sequences, for instance allowing for the identification of horizontally transferred or duplicated genes.

Conclusion

We created PhylteR, a tool to explore phylogenomics dataset and detect outlier gene sequences. Instead of fully removing rogue taxa or full outlier gene family, PhylteR precisely identifies what species in what gene family should be removed to increase concordance between genes. Doing so it accurately spots gene sequences with low phylogenetic signal, genes with saturated signal leading to long branches, paralogous genes, genes associated with synteny breaks and other sequences being dubious in gene phylogenies for any possible reason.

Acknowledgments

Work was funded by ANR Grant 18-CE02-0007 (Sthoriz) to DMDV, ANR Grant 19-CE45-0010 (Evoluthon) to ET, and European Research Council grant ERC-2015-CoG-683257 (ConvergeAnt project) to FD. This is contribution ISEM 2023-XXX of the Institut des Sciences de l'Evolution de Montpellier.

Software availability

PhylteR is available on CRAN (https://cran.r-project.org/web/packages/phylter/index.html) for the latest stable version and on GitHub (https://github.com/damiendevienne/phylter) for the latest development version. A version of PhylteR is also available as a Singularity container.

Data Availability

The documented code of PhylteR is available at https://github.com/damiendevienne/phylter along with a thorough documentation. All data and scripts used in this study are available on the dedicated GitHub repository available at https://github.com/damiendevienne/phylter-data/.

References

Abdi H, O'Toole AJ, Valentin D, Edelman B. 2005. DISTATIS: The analysis of multiple distance matrices. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops. IEEE. p. 42–42.

Abdi H, Williams LJ, Valentin D, Bennani-Dosse M. 2012. STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling. *Wiley Interdiscip. Rev. Comput. Stat.* 4:124–167.

Aberer AJ, Krompass D, Stamatakis A. 2013. Pruning Rogue Taxa Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice. *Syst. Biol.* 62:162–166.

Allio R, Tilak M-K, Scornavacca C, Avenant NL, Kitchener AC, Corre E, Nabholz B, Delsuc F. 2021. High-quality carnivoran genomes from roadkill samples enable comparative species delineation in aardwolf and bat-eared fox.Perry GH, Perry GH, editors. *eLife* 10:e63167.

Barton D, David F. 1956. Some notes on ordered random intervals. *J. R. Stat. Soc. Ser. B Methodol.* 18:79–94.

Biller P, Knibbe C, Beslon G, Tannier E. 2016. Comparative genomics on artificial life. In: Conference on Computability in Europe. Springer. p. 35–44.

Brys G, Hubert M, Struyf A. 2004. A Robust Measure of Skewness. *J. Comput. Graph. Stat.* 13:996–1017.

Chessel D, Hanafi M. 1996. Analyses de la co-inertie de $ K $ nuages de points. *Rev. Stat. Appliquée* 44:35–60.

Dalevi D, Eriksen N. 2008. Expected gene-order distances and model selection in bacteria. *Bioinformatics* 24:1332–1338.

Di Franco A, Poujol R, Baurain D, Philippe H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol. Biol.* 19:1–17.

Emerson RO, Thomas JH. 2009. Adaptive Evolution in Zinc Finger Transcription Factors. *PLoS Genet.* 5:e1000325.

von Haeseler A. 2012. Do we still need supertrees? *BMC Biol.* 10:13.

Hubert M, Vandervieren E. 2008. An adjusted boxplot for skewed distributions. *Comput. Stat. Data Anal.* 52:5186–5201.

Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, Kim J, Gordon L, Branscomb E, Stubbs L. 2006. A comprehensive catalog of human KRAB-associated zinc finger genes: Insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res.* 16:669–677.

Jackson DA. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 74:2204–2214.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780.

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Syst. Biol.* 62:611–615.

Liu H, Chang L-H, Sun Y, Lu X, Stubbs L. 2014. Deep Vertebrate Roots for Mammalian Zinc Finger Transcription Factor Subfamilies. *Genome Biol. Evol.* 6:510–525.

Lucas JM, Crollius HR. 2017. High precision detection of conserved segments from synteny blocks. *PLOS ONE* 12:e0180198.

Mai U, Mirarab S. 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* 19:272–272.

Minh BQ, Hahn MW, Lanfear R. 2020. New Methods to Calculate Concordance Factors for Phylogenomic Datasets. *Mol. Biol. Evol.* 37:2727–2733.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* 37:1530–1534.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35:526–528.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLOS Biol.* 9:e1000602.

Philippe H, de Vienne DM, Ranwez V, Roure B, Baurain D, Delsuc F. 2017. Pitfalls in supermatrix phylogenomics. *Eur. J. Taxon.* 283:1–25.

Ranwez V, Chantret N, Delsuc F. 2021. Aligning Protein-Coding nucleotide sequences with MACSE. In: Multiple Sequence Alignment. Springer. p. 51–70.

Ranwez V, Douzery EJ, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.* 35:2582–2584.

Robert P, Escoufier Y. 1976. A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient. *J. R. Stat. Soc. Ser. C Appl. Stat.* 25:257–265.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497:327–331.

Schrempf D, Szöllősi G. 2020. The sources of phylogenetic conflicts. *Phylogenetics Genomic Era*:3–1.

Scornavacca C, Belkhir K, Lopez J, Dernat R, Delsuc F, Douzery EJP, Ranwez V. 2019. OrthoMaM v10: Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes. *Mol. Biol. Evol.* 36:861–862.

Shen X-X, Salichos L, Rokas A. 2016. A Genome-Scale Investigation of How Sequence, Function, and Tree-Based Gene Properties Influence Phylogenetic Inference. *Genome Biol. Evol.* 8:2565–2580.

Szöllősi GJ, Davín AA, Tannier E, Daubin V, Boussau B. 2015. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20140335.

de Vienne DM, Ollier S, Aguileta G. 2012. Phylo-MCOA: A Fast and Efficient Method to Detect Outlier Genes and Species in Phylogenomics Using Multiple Co-inertia Analysis. *Mol. Biol. Evol.* 29:1587–1598.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15:496–503.

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

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19:153.

Supplementary Material

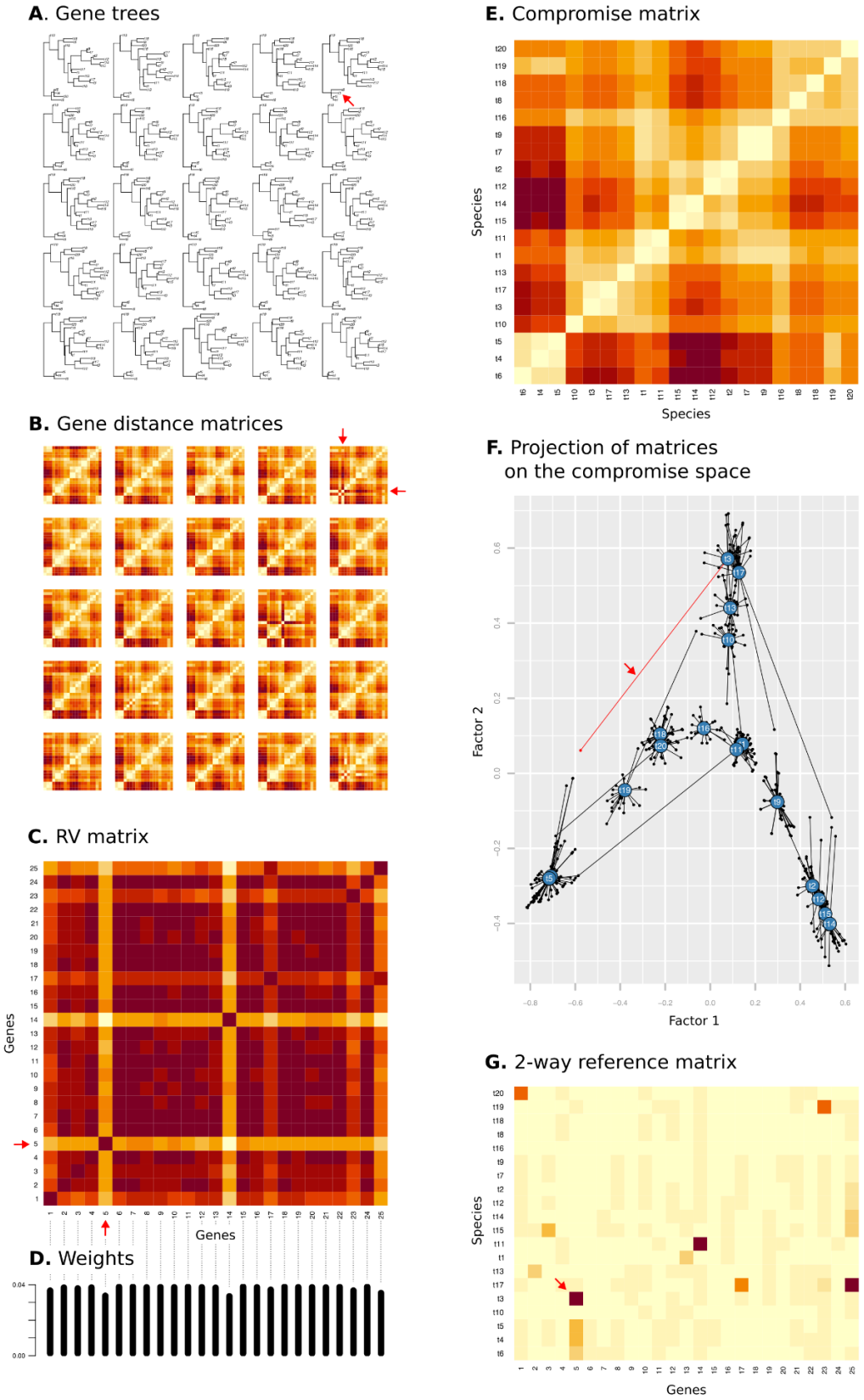

Figure S1. Illustration of the different steps of the PhylteR process depicted in Figure 1 (Main text). The red arrow identifies on each step, one of the outliers of the dataset, namely species t3 in gene 5.

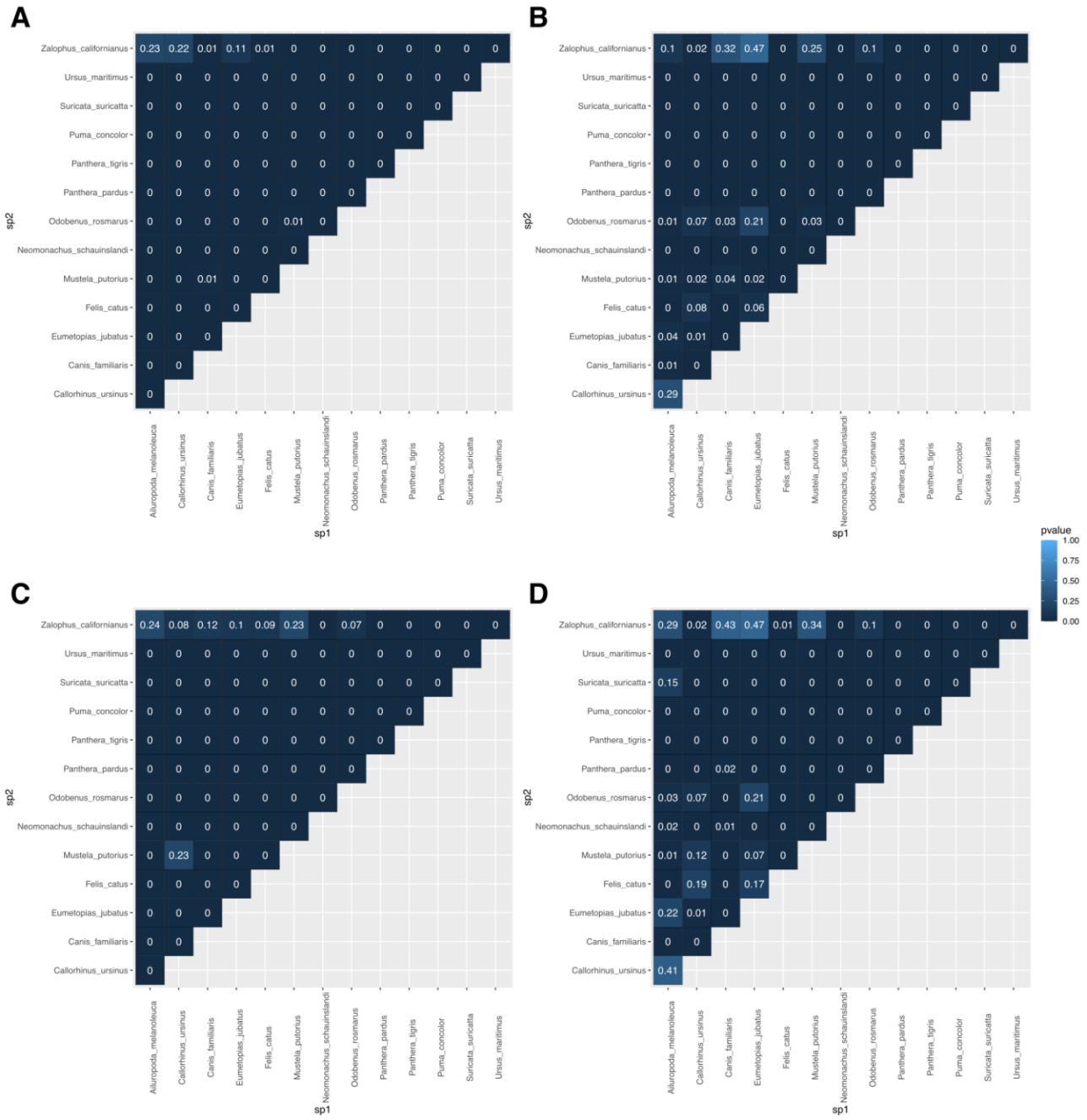

Figure S2. Analysis of the synteny breaks in the list of outliers. Each heatmap represents all pairwise comparisons between the 14 species of interest. For each comparison, the p-value associated with the probability of getting the observed number of syntenic outliers in the list of outliers is indicated. The first row (A and B) are the results with PhylteR outliers, the second row (C and D) are for TreeShrink outliers. The two columns represent the two sizes of outlier lists, small (A and C) and large (B and D).

Table S1. Genomes used for synteny breakage analysis.

Species	Family	Accession	AssemblyName	# scaffolds
Ailuropoda melanoleuca	Ursidae	GCF_000004335.2	AilMel_1.0	1913
Callorhinus ursinus	Otariidae	GCF_003265705.1	ASM326570v1	146
Canis lupus	Canidae	GCF_000002285.3	CanFam3.1	59
Eumetopias jubatus	Otariidae	GCF_004028035.1	ASM402803v1	323
Felis catus	Felidae	GCF_000181335.3	Felis_catus_9.0	25
Mustela putorius	Mustelidae	GCF_000215625.1	MusPutFur1.0	457
Neomonachus schauinslandi	Phocidae	GCF_002201575.1	ASM220157v1	273
Odobenus rosmarus	Odobenidae	GCF_000321225.1	Oros_1.0	1170
Panthera pardus	Felidae	GCF_001857705.1	PanPar1.0	289
Panthera tigris	Felidae	GCF_000464555.1	PanTig1.0	505
Puma concolor	Felidae	GCF_003327715.1	PumCon1.0	48
Suricata suricatta	Herpestidae	GCF_006229205.1	meerkat_22Aug2017_6uvM2_HiC	25
Ursus maritimus	Ursidae	GCF_000687225.1	UrsMar_1.0	314
Zalophus californianus	Otariidae	GCF_900631625.1	zalCal2.2	27

Table S2. Comparisons of genomes for the synteny analysis with PhylteR outlier lists.

			Phylter Outliers (small list)			Phylter outliers (large list)		
Species pair	**Genes**	**Syntenic outliers**	**Total**	**syntenic**	**P-value**	**Total**	**Syntenic**	**P-value**
Ailuropoda melanoleuca - Callorhinus ursinus	6493	558	56	14	1.986E-04	182	29	7.696E-04
Ailuropoda melanoleuca - Canis familiaris	12192	753	247	33	2.215E-05	740	66	1.484E-03
Ailuropoda melanoleuca - Eumetopias jubatus	7543	682	68	17	8.122E-05	223	36	3.977E-04
Ailuropoda melanoleuca - Felis catus	11229	656	144	19	7.040E-04	450	46	1.409E-04
Ailuropoda melanoleuca - Mustela putorius	11720	770	116	17	1.458E-03	445	56	1.740E-06
Ailuropoda melanoleuca - Neomonachus schauinslandi	12335	797	142	29	2.358E-08	454	53	1.854E-05
Ailuropoda melanoleuca - Odobenus rosmarus	12214	913	147	26	3.071E-05	460	56	1.847E-04
Ailuropoda melanoleuca - Panthera pardus	12123	733	150	22	9.565E-05	514	52	1.682E-04
Ailuropoda melanoleuca - Panthera tigris	11725	840	165	66	3.106E-33	483	99	1.735E-22
Ailuropoda melanoleuca - Puma concolor	9907	894	119	49	3.124E-21	364	93	2.337E-21
Ailuropoda melanoleuca - Suricata suricatta	10368	730	126	70	6.374E-48	378	92	2.302E-27
Ailuropoda melanoleuca - Ursus maritimus	11924	771	145	54	9.758E-28	400	78	3.807E-19
Ailuropoda melanoleuca - Zalophus californianus	5506	586	47	7	2.288E-01	159	20	2.450E-01
Callorhinus ursinus - Canis familiaris	6979	146	143	13	7.985E-06	409	18	2.083E-03
Callorhinus ursinus - Eumetopias jubatus	6660	105	40	5	3.756E-04	113	9	6.521E-05
Callorhinus ursinus - Felis catus	6309	92	72	8	8.385E-06	205	10	7.430E-04

			Phylter Outliers (small list)			Phylter outliers (large list)		
Species pair	**Genes**	**Syntenic outliers**	**Total**	**syntenic**	**P-value**	**Total**	**Syntenic**	**P-value**
Callorhinus ursinus - Mustela putorius	6616	182	50	7	3.901E-04	182	7	2.346E-01
Callorhinus ursinus - Neomonachus schauinslandi	7147	186	73	15	3.757E-10	207	18	6.373E-06
Callorhinus ursinus - Odobenus rosmarus	7061	355	75	11	1.199E-03	207	22	6.725E-04
Callorhinus ursinus - Panthera pardus	7003	161	86	11	3.722E-06	261	22	1.039E-07
Callorhinus ursinus - Panthera tigris	6617	212	82	40	5.197E-39	222	53	6.342E-33
Callorhinus ursinus - Puma concolor	6086	246	58	31	9.747E-29	173	46	3.465E-26
Callorhinus ursinus - Suricata suricatta	5953	139	58	40	4.026E-54	165	46	6.236E-39
Callorhinus ursinus - Ursus maritimus	6760	160	68	29	3.304E-30	165	35	1.819E-24
Callorhinus ursinus - Zalophus californianus	4782	71	17	1	2.249E-01	70	3	8.492E-02
Canis familiaris - Eumetopias jubatus	8091	190	182	17	1.134E-06	485	26	6.111E-05
Canis familiaris - Felis catus	11863	216	254	16	1.623E-05	734	28	1.584E-04
Canis familiaris - Mustela putorius	12383	301	239	13	5.692E-03	728	36	3.754E-05
Canis familiaris - Neomonachus schauinslandi	13119	358	271	29	3.072E-10	808	48	2.678E-07
Canis familiaris - Odobenus rosmarus	12955	517	271	29	1.299E-06	806	65	3.656E-08
Canis familiaris - Panthera pardus	12919	313	287	24	1.293E-07	849	40	3.868E-05
Canis familiaris - Panthera tigris	12405	407	286	75	2.215E-47	795	108	2.913E-39
Canis familiaris - Puma concolor	10525	444	217	53	1.990E-26	631	87	7.318E-24

			Phylter Outliers (small list)			Phylter outliers (large list)		
Species pair	**Genes**	**Syntenic outliers**	**Total**	**syntenic**	**P-value**	**Total**	**Syntenic**	**P-value**
Canis familiaris - Suricata suricatta	10948	276	244	83	8.032E-74	674	100	6.369E-52
Canis familiaris - Ursus maritimus	12614	355	276	61	9.473E-38	746	83	1.835E-28
Canis familiaris - Zalophus californianus	5918	106	119	7	5.254E-03	367	10	1.201E-01
Eumetopias jubatus - Felis catus	7318	132	77	9	8.797E-06	231	13	2.539E-04
Eumetopias jubatus - Mustela putorius	7666	251	69	12	1.957E-06	226	16	2.972E-03
Eumetopias jubatus - Neomonachus schauinslandi	8287	241	95	27	5.158E-20	247	29	1.003E-10
Eumetopias jubatus - Odobenus rosmarus	8178	422	89	19	9.234E-08	236	27	7.787E-05
Eumetopias jubatus - Panthera pardus	8109	205	107	18	1.354E-10	312	28	3.966E-09
Eumetopias jubatus - Panthera tigris	7685	273	99	43	6.526E-37	258	55	1.328E-28
Eumetopias jubatus - Puma concolor	7099	293	78	36	8.046E-30	219	48	1.202E-22
Eumetopias jubatus - Suricata suricatta	6905	197	72	48	7.701E-59	212	56	2.380E-40
Eumetopias jubatus - Ursus maritimus	7853	211	102	42	3.399E-40	236	50	7.035E-32
Eumetopias jubatus - Zalophus californianus	5390	91	33	2	1.063E-01	104	4	9.814E-02
Felis catus - Mustela putorius	11436	235	143	12	3.785E-05	459	24	2.488E-05
Felis catus - Neomonachus schauinslandi	12019	286	165	18	7.406E-08	493	29	6.411E-06
Felis catus - Odobenus rosmarus	11890	443	153	17	5.489E-05	474	40	1.085E-06
Felis catus - Panthera pardus	11887	176	149	14	4.142E-08	408	23	3.412E-08

			Phylter Outliers (small list)			Phylter outliers (large list)		
Species pair	**Genes**	**Syntenic outliers**	**Total**	**syntenic**	**P-value**	**Total**	**Syntenic**	**P-value**
Felis catus - Panthera tigris	11482	258	161	41	3.468E-32	396	55	1.303E-28
Felis catus - Puma concolor	9687	303	115	30	6.281E-20	306	53	2.375E-25
Felis catus - Suricata suricatta	10140	210	121	56	9.431E-64	351	67	2.597E-47
Felis catus - Ursus maritimus	11628	227	161	39	3.804E-32	448	48	1.002E-22
Felis catus - Zalophus californianus	5268	72	57	4	7.342E-03	177	5	9.343E-02
Mustela putorius - Neomonachus schauinslandi	12565	389	141	17	1.647E-06	474	33	1.164E-05
Mustela putorius - Odobenus rosmarus	12441	522	139	13	5.622E-03	475	33	3.103E-03
Mustela putorius - Panthera pardus	12334	323	153	17	5.161E-07	523	34	9.247E-07
Mustela putorius - Panthera tigris	11905	415	154	58	3.658E-45	494	81	4.598E-33
Mustela putorius - Puma concolor	10064	428	108	38	2.150E-25	370	62	2.780E-21
Mustela putorius - Suricata suricatta	10496	303	115	65	2.327E-71	372	78	1.550E-46
Mustela putorius - Ursus maritimus	12097	327	148	44	4.341E-34	447	62	2.157E-27
Mustela putorius - Zalophus californianus	5537	174	37	6	9.131E-04	159	7	2.327E-01
Neomonachus schauinslandi - Odobenus rosmarus	13241	568	155	29	1.555E-11	462	46	8.479E-08
Neomonachus schauinslandi - Panthera pardus	13160	357	190	28	2.287E-13	586	46	7.056E-11
Neomonachus schauinslandi - Panthera tigris	12635	413	188	65	1.386E-49	539	88	5.334E-38
Neomonachus schauinslandi - Puma concolor	10711	449	146	56	9.144E-40	421	83	3.797E-34

Species pair	Genes	Syntenic outliers	Phylter Outliers (small list)			Phylter outliers (large list)		
			Total	syntenic	P-value	Total	Syntenic	P-value
Neomonachus schauinslandi - Suricata suricatta	11126	328	134	72	4.327E-76	420	85	1.469E-48
Neomonachus schauinslandi - Ursus maritimus	12855	380	172	57	9.094E-45	473	66	7.579E-27
Neomonachus schauinslandi - Zalophus californianus	6052	163	54	15	5.205E-12	183	18	1.637E-06
Odobenus rosmarus - Panthera pardus	12988	538	181	24	4.466E-07	583	47	9.022E-06
Odobenus rosmarus - Panthera tigris	12447	613	189	72	1.344E-45	540	100	3.563E-32
Odobenus rosmarus - Puma concolor	10558	665	142	56	6.060E-31	415	93	1.767E-28
Odobenus rosmarus - Suricata suricatta	11012	521	137	69	2.882E-55	421	91	1.960E-36
Odobenus rosmarus - Ursus maritimus	12672	554	172	54	5.486E-32	479	72	9.482E-21
Odobenus rosmarus - Zalophus californianus	5973	351	57	10	1.561E-03	183	16	7.072E-02
Panthera pardus - Panthera tigris	12541	381	183	55	5.716E-40	480	79	1.201E-36
Panthera pardus - Puma concolor	10607	428	158	49	1.305E-30	395	78	2.561E-33
Panthera pardus - Suricata suricatta	10955	290	141	69	1.152E-72	403	83	3.834E-52
Panthera pardus - Ursus maritimus	12660	323	189	49	6.084E-36	541	62	3.828E-24
Panthera pardus - Zalophus californianus	5938	135	70	8	1.671E-04	230	15	2.045E-04
Panthera tigris - Puma concolor	10394	439	119	44	1.932E-30	334	69	1.787E-29
Panthera tigris - Suricata suricatta	10583	331	141	83	1.168E-90	385	105	1.344E-72
Panthera tigris - Ursus maritimus	12504	363	165	51	8.915E-39	477	74	6.926E-34

			Phylter Outliers (small list)			Phylter outliers (large list)		
Species pair	**Genes**	**Syntenic outliers**	**Total**	**syntenic**	**P-value**	**Total**	**Syntenic**	**P-value**
Panthera tigris - Zalophus californianus	5593	203	67	41	1.732E-43	200	48	1.454E-27
Puma concolor - Suricata suricatta	9020	359	120	78	1.140E-81	322	99	1.704E-63
Puma concolor - Ursus maritimus	10452	413	128	49	4.479E-36	365	72	2.534E-31
Puma concolor - Zalophus californianus	5164	209	44	24	8.684E-23	161	37	6.387E-19
Suricata suricatta - Ursus maritimus	10741	270	145	83	3.697E-98	394	93	3.550E-67
Suricata suricatta - Zalophus californianus	5123	132	47	33	1.339E-43	149	42	4.734E-34
Ursus maritimus - Zalophus californianus	5717	158	69	37	5.993E-41	171	42	1.287E-29

Table S3. Comparisons of genomes for the synteny analysis with TreeShrink outlier lists.

Species pair	Genes	Syntenic outliers	TreeShrink Outliers (small list)			TreeShrink outliers (large list)		
			Total	Syntenic outliers	P-value	Total	Syntenic outliers	P-value
Ailuropoda melanoleuca - Canis familiaris	12192	753	215	22	1.362E-02	803	71	1.248E-03
Ailuropoda melanoleuca - Eumetopias jubatus	7543	682	111	16	4.073E-02	225	24	2.240E-01
Ailuropoda melanoleuca - Felis catus	11229	656	144	22	3.013E-05	351	34	2.526E-03
Ailuropoda melanoleuca - Mustela putorius	11720	770	204	22	1.486E-02	781	67	1.382E-02
Ailuropoda melanoleuca - Neomonachus schauinslandi	12335	797	168	24	1.990E-04	332	32	1.509E-02
Ailuropoda melanoleuca - Odobenus rosmarus	12214	913	198	25	6.764E-03	352	36	3.365E-02
Ailuropoda melanoleuca - Panthera pardus	12123	733	188	27	2.352E-05	347	34	3.668E-03
Ailuropoda melanoleuca - Panthera tigris	11725	840	209	75	7.103E-34	447	93	1.248E-21
Ailuropoda melanoleuca - Puma concolor	9907	894	165	59	1.268E-21	367	83	1.038E-15
Ailuropoda melanoleuca - Suricata suricatta	10368	730	176	29	1.382E-05	693	56	1.514E-01
Ailuropoda melanoleuca - Ursus maritimus	11924	771	173	44	1.373E-15	384	56	6.694E-09
Ailuropoda melanoleuca - Zalophus californianus	5506	586	76	12	1.046E-01	164	20	2.916E-01
Callorhinus ursinus - Canis familiaris	6979	146	121	8	3.720E-03	404	18	1.818E-03
Callorhinus ursinus - Eumetopias jubatus	6660	105	47	4	6.185E-03	47	4	6.185E-03
Callorhinus ursinus - Felis catus	6309	92	70	3	8.176E-02	103	3	1.896E-01

Species pair	Genes	Syntenic outliers	TreeShrink Outliers (small list)			TreeShrink outliers (large list)		
			Total	Syntenic outliers	P-value	Total	Syntenic outliers	P-value
Callorhinus ursinus - Mustela putorius	6616	182	100	7	1.985E-02	358	14	1.155E-01
Callorhinus ursinus - Neomonachus schauinslandi	7147	186	79	9	1.930E-04	79	9	1.930E-04
Callorhinus ursinus - Odobenus rosmarus	7061	355	72	7	6.878E-02	72	7	6.878E-02
Callorhinus ursinus - Panthera pardus	7003	161	95	10	5.871E-05	95	10	5.871E-05
Callorhinus ursinus - Panthera tigris	6617	212	122	39	1.987E-29	180	42	1.483E-25
Callorhinus ursinus - Puma concolor	6086	246	113	41	2.366E-29	176	48	7.046E-28
Callorhinus ursinus - Suricata suricatta	5953	139	101	14	6.715E-08	298	20	1.611E-05
Callorhinus ursinus - Ursus maritimus	6760	160	112	24	4.920E-17	169	26	1.105E-14
Callorhinus ursinus - Zalophus californianus	4782	71	37	3	1.699E-02	37	3	1.699E-02
Canis familiaris - Eumetopias jubatus	8091	190	153	11	9.194E-04	480	28	6.433E-06
Canis familiaris - Felis catus	11863	216	174	11	3.393E-04	642	28	1.506E-05
Canis familiaris - Mustela putorius	12383	301	251	11	4.312E-02	1117	43	1.640E-03
Canis familiaris - Neomonachus schauinslandi	13119	358	214	18	2.340E-05	683	30	6.916E-03
Canis familiaris - Odobenus rosmarus	12955	517	243	16	3.445E-02	704	48	1.903E-04
Canis familiaris - Panthera pardus	12919	313	229	16	1.465E-04	686	26	1.593E-02
Canis familiaris - Panthera tigris	12405	407	257	71	1.649E-46	751	91	7.629E-29
Canis familiaris - Puma concolor	10525	444	207	61	1.913E-35	664	81	8.506E-19

Species pair	Genes	Syntenic outliers	TreeShrink Outliers (small list)			TreeShrink outliers (large list)		
			Total	Syntenic outliers	P-value	Total	Syntenic outliers	P-value
Canis familiaris - Suricata suricatta	10948	276	224	26	6.989E-11	991	41	1.025E-03
Canis familiaris - Ursus maritimus	12614	355	264	46	8.006E-24	771	65	7.198E-16
Canis familiaris - Zalophus californianus	5918	106	110	3	3.151E-01	348	7	4.325E-01
Eumetopias jubatus - Felis catus	7318	132	83	4	6.249E-02	121	4	1.744E-01
Eumetopias jubatus - Mustela putorius	7666	251	125	9	2.119E-02	433	20	7.451E-02
Eumetopias jubatus - Neomonachus schauinslandi	8287	241	100	13	5.783E-06	100	13	5.783E-06
Eumetopias jubatus - Odobenus rosmarus	8178	422	94	7	2.109E-01	94	7	2.109E-01
Eumetopias jubatus - Panthera pardus	8109	205	115	10	6.337E-04	115	10	6.337E-04
Eumetopias jubatus - Panthera tigris	7685	273	145	45	5.308E-31	209	48	1.743E-26
Eumetopias jubatus - Puma concolor	7099	293	126	41	1.264E-26	200	48	1.691E-24
Eumetopias jubatus - Suricata suricatta	6905	197	123	18	9.139E-09	367	25	3.989E-05
Eumetopias jubatus - Ursus maritimus	7853	211	150	31	1.774E-19	216	32	1.288E-15
Eumetopias jubatus - Zalophus californianus	5390	91	37	1	4.685E-01	37	1	4.685E-01
Felis catus - Mustela putorius	11436	235	186	16	1.380E-06	669	33	2.382E-06
Felis catus - Neomonachus schauinslandi	12019	286	142	14	7.113E-06	204	16	2.941E-05
Felis catus - Odobenus rosmarus	11890	443	169	16	5.713E-04	230	21	1.391E-04
Felis catus - Panthera pardus	11887	176	129	14	6.433E-09	183	16	1.212E-08

Species pair	Genes	Syntenic outliers	TreeShrink Outliers (small list)			TreeShrink outliers (large list)		
			Total	Syntenic outliers	P-value	Total	Syntenic outliers	P-value
Felis catus - Panthera tigris	11482	258	146	41	4.405E-34	269	44	7.307E-26
Felis catus - Puma concolor	9687	303	123	39	3.180E-29	246	45	1.416E-22
Felis catus - Suricata suricatta	10140	210	150	22	3.894E-13	564	28	1.410E-05
Felis catus - Ursus maritimus	11628	227	185	35	5.896E-25	328	41	5.936E-22
Felis catus - Zalophus californianus	5268	72	59	5	1.170E-03	89	5	7.070E-03
Mustela putorius - Neomonachus schauinslandi	12565	389	201	20	4.234E-06	665	38	1.918E-04
Mustela putorius - Odobenus rosmarus	12441	522	228	16	3.082E-02	694	54	7.614E-06
Mustela putorius - Panthera pardus	12334	323	222	21	3.696E-07	688	44	3.174E-08
Mustela putorius - Panthera tigris	11905	415	238	61	2.653E-36	780	94	1.266E-27
Mustela putorius - Puma concolor	10064	428	201	50	3.057E-25	677	75	6.000E-15
Mustela putorius - Suricata suricatta	10496	303	194	30	3.455E-14	906	56	2.985E-08
Mustela putorius - Ursus maritimus	12097	327	243	35	3.285E-16	776	58	8.841E-13
Mustela putorius - Zalophus californianus	5537	174	81	4	2.498E-01	300	11	3.424E-01
Neomonachus schauinslandi - Odobenus rosmarus	13241	568	171	21	1.364E-05	171	21	1.364E-05
Neomonachus schauinslandi - Panthera pardus	13160	357	188	22	7.913E-09	188	22	7.913E-09
Neomonachus schauinslandi - Panthera tigris	12635	413	215	67	9.955E-48	308	72	8.459E-42

Species pair	Genes	Syntenic outliers	TreeShrink Outliers (small list)			TreeShrink outliers (large list)		
			Total	Syntenic outliers	P-value	Total	Syntenic outliers	P-value
Neomonachus schauinslandi - Puma concolor	10711	449	185	65	2.539E-43	288	74	1.210E-38
Neomonachus schauinslandi - Suricata suricatta	11126	328	190	25	3.098E-10	594	37	1.265E-05
Neomonachus schauinslandi - Ursus maritimus	12855	380	213	44	4.467E-25	303	48	5.249E-22
Neomonachus schauinslandi - Zalophus californianus	6052	163	61	9	3.119E-05	61	9	3.119E-05
Odobenus rosmarus - Panthera pardus	12988	538	209	28	3.912E-08	209	28	3.912E-08
Odobenus rosmarus - Panthera tigris	12447	613	243	74	5.050E-39	336	80	1.199E-33
Odobenus rosmarus - Puma concolor	10558	665	202	66	1.437E-30	309	82	1.373E-30
Odobenus rosmarus - Suricata suricatta	11012	521	210	29	1.881E-07	608	47	5.724E-04
Odobenus rosmarus - Ursus maritimus	12672	554	239	47	1.508E-18	327	52	3.018E-16
Odobenus rosmarus - Zalophus californianus	5973	351	68	7	1.027E-01	68	7	1.027E-01
Panthera pardus - Panthera tigris	12541	381	194	66	5.268E-52	265	69	1.357E-45
Panthera pardus - Puma concolor	10607	428	174	64	3.513E-45	262	72	6.301E-41
Panthera pardus - Suricata suricatta	10955	290	176	26	7.913E-13	582	34	1.129E-05
Panthera pardus - Ursus maritimus	12660	323	229	47	1.580E-29	319	50	1.141E-25
Panthera pardus - Zalophus californianus	5938	135	74	8	2.472E-04	74	8	2.472E-04
Panthera tigris - Puma concolor	10394	439	173	66	1.585E-46	316	80	3.867E-41

Species pair	Genes	Syntenic outliers	TreeShrink Outliers (small list)			TreeShrink outliers (large list)		
			Total	Syntenic outliers	P-value	Total	Syntenic outliers	P-value
Panthera tigris - Suricata suricatta	10583	331	210	60	9.452E-42	676	74	2.804E-22
Panthera tigris - Ursus maritimus	12504	363	234	62	1.196E-42	408	69	5.657E-34
Panthera tigris - Zalophus californianus	5593	203	110	46	1.011E-38	159	50	5.493E-35
Puma concolor - Suricata suricatta	9020	359	178	60	1.957E-40	581	75	1.176E-20
Puma concolor - Ursus maritimus	10452	413	193	67	4.736E-46	355	78	1.725E-37
Puma concolor - Zalophus californianus	5164	209	96	37	8.113E-28	156	45	1.527E-27
Suricata suricatta - Ursus maritimus	10741	270	212	43	2.693E-27	695	57	6.877E-16
Suricata suricatta - Zalophus californianus	5123	132	94	15	1.088E-08	270	23	2.635E-07
Ursus maritimus - Zalophus californianus	5717	158	107	28	1.474E-20	159	29	1.239E-16