



HAL
open science

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

► To cite this version:

Aurore Comte, Théo Tricou, Eric Tannier, Julien Joseph, Aurélie Siberchicot, et al.. PhylteR: efficient identification of outlier sequences in phylogenomic datasets. *Molecular Biology and Evolution*, 2023, 40 (11), 10.1093/molbev/msad234 . hal-03995366v2

HAL Id: hal-03995366

<https://hal.science/hal-03995366v2>

Submitted on 6 Nov 2023 (v2), last revised 20 Dec 2023 (v3)

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

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



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

1 PhylteR: efficient identification of outlier sequences 2 in phylogenomic datasets

3
4 **Aurore Comte^{1,2,†}, Théo Tricou^{3,†}, Eric Tannier^{3,4}, Julien Joseph³, Aurélie Siberchicot³, Simon
5 Penel³, Rémi Allio⁵, Frédéric Delsuc⁶, Stéphane Dray³, Damien M. de Vienne^{3,*}**

6
7 ¹ French Institute of Bioinformatics (IFB)—South Green Bioinformatics Platform, Bioversity,
8 CIRAD, INRAE, IRD, Montpellier 34398, France

9 ² IRD, CIRAD, INRAE, Institut Agro, PHIM Plant Health Institute, Montpellier University,
10 Montpellier 34398, France

11 ³ Université de Lyon, Université Lyon 1, UMR CNRS 5558 Laboratoire de Biométrie et Biologie
12 Évolutive, 69622 Villeurbanne, France

13 ⁴ Centre de Recherches Inria de Lyon, 68622 Villeurbanne, France

14 ⁵ CBGP, INRAE, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier, Montpellier, France

15 ⁶ ISEM, Univ. Montpellier, CNRS, IRD, 30495 Montpellier, France

16 [†] Equal contribution

17
18 * Corresponding author:

19 DAMIEN M. DE VIENNE

20 Université Lyon 1, CNRS

21 Laboratoire de Biométrie et Biologie Évolutive

22 Bâtiment Mendel 43 boulevard du 11 Novembre 1918

23 69622 VILLEURBANNE CEDEX

24 **Phone:** +33(0)4 72 43 29 09

25 **E-mail:** damien.de-vienne@univ-lyon1.fr

28 **Abstract**

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

34 Here, we present PhylteR, a method that allows a rapid and accurate detection of outlier sequences
35 in phylogenomic datasets, i.e. species from individual gene trees that do not follow the general
36 trend. PhylteR relies on DISTATIS, an extension of multidimensional scaling to 3 dimensions to
37 compare multiple distance matrices at once. In PhylteR, these distance matrices extracted from

1 individual gene phylogenies represent evolutionary distances between species according to each
2 gene.

3 On simulated datasets, we show that PhylteR identifies outliers with more sensitivity and precision
4 than a comparable existing method. We also show that PhylteR is not sensitive to ILS-induced
5 incongruences, which is a desirable feature. On a biological dataset of 14,463 genes for 53 species
6 previously assembled for Carnivora phylogenomics, we show (i) that PhylteR identifies as outliers
7 sequences that can be considered as such by other means, and (ii) that the removal of these
8 sequences improves the concordance between the gene trees and the species tree. Thanks to the
9 generation of numerous graphical outputs, PhylteR also allows for the rapid and easy visual
10 characterisation of the dataset at hand, thus aiding in the precise identification of errors.

11 PhylteR is distributed as an R package on CRAN and as containerized versions (docker and
12 singularity).

13 **Introduction**

14 Supermatrix, supertree and coalescent-based approaches are commonly used in phylogenomics to
15 obtain a species tree from a collection of genes. These methods are similar in their first steps: for a
16 list of taxa of interest, a large collection of single-copy orthologous gene sequences is retrieved and
17 a multiple sequence alignment (MSA) is computed for each cluster of orthologous genes (see von
18 Haeseler 2012 for a comparison of these approaches). The methods then differ by the strategy
19 employed. In the supermatrix approach, MSAs are concatenated into a supermatrix that is used to
20 build a phylogeny, generally with Maximum Likelihood (ML) or Bayesian methods (such as IQ-
21 TREE, Minh, Schmidt, et al. 2020; or Phylobayes, Lartillot et al. 2013). In the supertree and
22 coalescent-based approaches, individual gene trees are built from individual MSAs and a species
23 tree is obtained by combining them all, *e.g.* with MRP (Baum 1992; Ragan 1992; Ronquist 1996),
24 MP-EST (Liu et al. 2010) or ASTRAL (Zhang et al. 2018) to only cite a few.

25 Regardless of the method employed, errors in the individual gene MSAs and errors in the individual
26 gene trees (leading to incongruences with the species tree) can negatively impact the quality
27 (accuracy) of the reconstructed species tree (Philippe et al. 2017).

28 For MSAs, various filtering methods have been developed, categorised into two groups: methods
29 that entirely remove sites (columns) or sequences (rows) from the alignment (trimAl, Capella-
30 Gutiérrez et al. 2009; BMGE, Criscuolo and Gribaldo 2010), and methods that are more *picky* and
31 allow identifying and filtering (or masking) small segments in the alignments (Divvier, Ali et al.
32 2019; HmmCleaner Di Franco et al. 2019; TAPER, Zhang et al. 2021). The latter group of methods
33 was shown to be a better choice for alignment filtering, leading to better gene tree topologies (closer

1 to the species tree) and more consistent terminal branch lengths in gene trees (Ranwez and Chantret
2 2020).

3 For collection of gene trees, filtering methods also exist, and the categorization into two groups of
4 strategies still holds. In the first group are methods that prune rogue taxa, *i.e.* taxa that are unstable
5 among gene trees (RogueNaRok, Aberer et al. 2013) and methods that eliminate orthologous gene
6 families whose history is uncorrelated with the others. In the second group are more *picky* methods
7 that identify and filter out only some species in some genes trees (*i.e.* Phylo-MCOA, de Vienne et
8 al. 2012; or TreeShrink, Mai and Mirarab 2018). Just like for the alignment filtering methods seen
9 above, *picky* approaches are thought to provide the best compromise between removing sequences
10 with conflicting phylogenetic signals and keeping the maximum information content.

11 Filtering (sometimes called trimming) MSAs is now done routinely in phylogenomic pipelines, with
12 methods that can be applied automatically to large datasets (see above). But just because we apply a
13 filter at the MSA level doesn't mean we shouldn't filter gene trees also. Some of the reasons that
14 lead to incongruences between gene trees and the species tree (and among gene trees), *i.e.* gene
15 tree-building errors, undetected paralogy, horizontal gene transfer (HGT) and Incomplete Lineage
16 Sorting (ILS), may not be detectable at the MSA stage. For filtering individual gene trees, reference
17 methods do not exist yet (Philippe et al. 2017). Indeed, identifying species in gene trees whose
18 position is not concordant with their position in the other gene trees (referred to as outliers in de
19 Vienne et al. 2012 and hereafter) is still commonly done by eye (when it is done), which is highly
20 questionable in terms of efficacy, objectivity, and reproducibility.

21
22 Here we present PhylteR, a new phylogenomics filtering method that can accurately and rapidly
23 identify outliers in a collection of gene trees. Unlike Phylo-MCOA (de Vienne et al. 2012), from
24 which it is largely inspired, it is an iterative process where obvious outliers are removed first,
25 leaving space for better identification of more subtle ones, and leading *in fine* to a finer
26 identification of outliers. Unlike TreeShrink (Mai and Mirarab 2018), it is not based solely on the
27 diameter of unrooted gene trees and is thus more accurate when outliers are not associated with long
28 branches (e.g. topological incongruences). Also, PhylteR relies on the multivariate analysis method
29 DISTATIS (Abdi et al. 2005; Abdi et al. 2012), which is specifically designed to compare distance
30 matrices, unlike Multiple Co-inertia Analysis (MCOA, Chessel and Hanafi 1996) used in Phylo-
31 MCOA (de Vienne et al. 2012), and is thus more appropriate for the problem at hand

32
33 We tested PhylteR on two types of datasets: simulated datasets where outliers were known, and a
34 biological dataset comprising 14,463 genes for up to 53 species previously used for Carnivora
35 phylogenomics (Allio et al. 2021). For the simulated datasets, horizontal gene transfers (HGTs)

1 were simulated and recorded, and sequences affected by these HGTs were considered as outliers.
2 The simulations also included various degrees of Incomplete Lineage Sorting (ILS), a phenomenon
3 where within-species polymorphism lasts longer than the time between two successive speciations
4 (Scornavacca and Galtier 2017), leading to incongruences between gene and species trees and
5 among gene trees. The importance of this phenomenon as a source of incongruence in
6 phylogenomic datasets is not clear, but because recent coalescence-based methods are able to
7 handle ILS-induced signal explicitly (e.g. ASTRAL, Zhang et al. 2018), it seemed pertinent to
8 evaluate whether PhylteR was (or not) sensitive to it. In the empirical dataset, outliers were of
9 course unknown but “properties” associated to gene sequences could be gathered (see Shen et al.
10 2016 for a list of such properties), so that enrichment of sequences having some of these properties
11 in the list of outliers could be quantified. Finally, we looked at the effect of PhylteR on the overall
12 concordance between the gene trees and the species tree after filtering. We compared the results
13 with those obtained with TreeShrink (Mai and Mirarab 2018), the only other tool to our knowledge
14 with a similar objective that could reasonably be applied on such a large dataset.

15
16 We show that PhylteR correctly identifies species in gene trees whose phylogenetic placement is
17 not in accordance with its placement in other gene trees, and that this holds even in the presence of
18 incongruence among gene trees due to ILS. We also provide strong evidence that the automatic
19 removal of outliers with PhylteR improves the concordance between gene trees and the species tree
20 in greater proportions than TreeShrink (Mai and Mirarab 2018).

21 We hope that PhylteR could become the standard that was lacking (Philippe et al. 2017) for
22 cleaning datasets prior to species tree reconstruction in phylogenomic pipelines.

23 **Material and Methods**

24 **Description of the PhylteR method**

25 The PhylteR method, in its entirety, is depicted in Figure 1. It starts with K distance matrices
26 obtained from K genes by computing pairwise distances (sum of branch lengths) between species in
27 each gene tree. All the matrices are given the same dimensionality by filling missing data (if any)
28 with the mean value across matrices, and are then normalised by dividing each matrix by either its
29 median or its mean value (default is median). The normalisation by median prevents genes from
30 fast- and slow-evolving orthologous genes to be erroneously considered as outliers, and appears as a
31 better choice than a normalisation by the mean as it is less affected by outlier values.

32

33

1 **Figure 1. Principle of the PhylteR method for identifying outliers in phylogenomic datasets.** The method relies on
2 DISTATIS (grey block), an extension of multidimensional scaling to three dimensions. See text for the detail of the
3 different steps.
4

5
6
7
8 From the K matrices obtained, an incremental process starts consisting in three main steps detailed
9 in the next sections: (1) comparison of the matrices with the DISTATIS method (Abdi et al. 2005;
10 Abdi et al. 2012), (2) detection of outliers sequences, and (3) evaluation of the impact of removing
11 these outliers on the overall concordance between the matrices. Note that we refer to *outlier*
12 *sequence* as a single gene for a single species (one sequence in one alignment, or one tip in one
13 gene tree) that does not follow the general trend (i.e. other alignments or gene trees), while *outlier*
14 *gene* refers to a complete alignment (or a complete gene tree) that does not agree with the other
15 alignments (or gene trees).

16 These steps are repeated until no more outlier sequences are detected, or until the removal of the
17 identified outlier sequences does not increase the concordance between the matrices more than a
18 certain amount specified by the user. Before finishing the optimization, PhylteR performs a last
19 action consisting of checking whether some outlier genes still exist despite the removal of outlier
20 sequences already performed. These outlier genes correspond to single-copy orthologous genes for
21 which the lack of correlation with others is not due to a few outlier sequences but are globally not
22 following the trend. If outlier genes are discarded there, the optimization restarts as it may have
23 unblocked the detection of other outliers.
24

25 ***Comparison of individual gene matrices with DISTATIS***

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

1 definition of the compromise). In a third step, the compromise matrix is submitted to an
 2 eigendecomposition procedure so that species can be represented in a low-dimensional multivariate
 3 space. In this compromise space, species are positioned so that their distances (computed in few
 4 dimensions, see after) represent the best approximations of the original distances contained in the
 5 compromise matrix. We used a broken stick model (Barton and David 1956) to estimate the number
 6 of dimensions (axes) of the compromise space, as this simple method was shown to give a good
 7 approximation of the correct dimensionality of the data with another multivariate approach (Jackson
 8 1993). Then, each individual pairwise distance matrix is projected on the compromise space. This
 9 allows us to obtain a representation of species associated with each gene family. In other words, the
 10 compromise identifies the dissimilarities between species that are common for all genes whereas the
 11 projections of individual distance matrices allow depiction of the peculiarities of each sequence.
 12 Lastly, we compute the distances, in the compromise space, between the position of a species given
 13 by all genes (the compromise) and its position associated to a particular gene family (using the
 14 projection procedure) and filled a gene x species 2-Way Reference matrix (2WR matrix, see figure
 15 1) with these values.

16

17 ***Detection of outlier sequences from DISTATIS results***

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

22

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

23

24 are considered outliers. $Q3$ is the 3rd quartile of the distribution, IQR is its interquartile range and
 25 MC is the medcouple of the distribution (Brys et al. 2004), a measure of skewness bounded between
 26 -1 (left skewed) and +1 (right skewed). The k value is chosen by the user (default is 3), and controls
 27 how stringent the detection of gene outliers is. Small values of k lead to more gene outliers being
 28 detected. The detection of gene outliers is performed after normalisation of the 2WR matrix,
 29 achieved by dividing each row (the default) or each column by its median. This normalisation leads
 30 to an exaggeration of outlier values, easing their identification.

31

32 ***Detection of outlier genes***

33 When no more outlier sequences are found in the 2WR matrix, PhylteR checks whether some genes
 34 are still uncorrelated to others. These outlier genes are detected by finding outlier values in the

1 weight array ($\alpha_1, \alpha_2, \dots, \alpha_K$, see Figure 1). The outlier detection method used is the same as for the
2 outlier sequences of the 2WR matrix (Equation 1) but its stringency can be tuned independently
3 (with parameter k_2 in place of parameter k in Equation 1, defaulting to $k_2 = k = 3$).

5 ***Exit criteria of the PhylteR iterative process***

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

12 **Evaluation of the PhylteR method**

13 ***Datasets***

14 We used three types of datasets to evaluate PhylteR and compare it with TreeShrink: a simple
15 dataset used for illustrative purpose only, a collection of simulated examples obtained with the
16 program SimPhy (Mallo et al. 2016), and a large Carnivora phylogenomic dataset with 53 species
17 (Allio et al. 2021). These datasets are described below.

- 19 ● *Simulated dataset for illustrative purpose:* we generated a small collection of gene trees in
20 order to illustrate the different steps of the PhylteR process. A single phylogenetic tree with
21 20 species was randomly generated with function `rtree()` from package *ape* v5.6.2
22 (Paradis and Schliep 2019). This tree was duplicated 25 times to mimic 25 orthologous gene
23 families. To add variance to branch lengths, a value sampled in a normal distribution with
24 mean 0 and standard deviation 0.15 was added to each branch length of each tree (if the
25 resulting branch length was negative its absolute value was taken). Ten outliers were then
26 generated by randomly sampling 10 times a species in a gene tree and moving it to another
27 random location.
- 29 ● *Simulated datasets:* We simulated collections of gene trees with known outliers in order to
30 evaluate PhylteR and compare it with TreeShrink. We used SimPhy (Mallo et al. 2016), a
31 program that can simulate the evolution of gene families (and thus gene trees) given a
32 species tree under various evolutionary processes including HGT but also ILS. We used, as
33 a species tree, the 53-taxa carnivora tree of Allio et al. (2021), the same as for the biological
34 dataset (next point). To be usable in SimPhy, we transformed the tree to ultrametric with
35 function *chronos* in *ape* (Paradis and Schliep 2019) and we rescaled the branch lengths so

1 that the root-to-tip distance reflected (roughly) the number of generations, *i.e.* 8,899,579
2 generations in this case. This value was obtained by dividing the age of the root of the tree
3 (74 millions years old, (Kumar and Subramanian 2002)) with a rough estimate of the
4 generation time in carnivora (8.315 years if taking the median of the generation times of the
5 species studied in (Kerk et al. 2013)).

6 For each replicate (100 each time), collections of 500 gene trees were simulated by setting
7 the rate of HGT to $1e-8$, the tree-wide substitution rate to $2.2e-9$, and varying the level of
8 ILS by changing the population size: 10 (NO-ILS), 100,000 (LOW-ILS), 200,000
9 (MODERATE-ILS), 500,000 (HIGH-ILS; the detailed commands used for Simphy are
10 given as supplementary method). Then, from the 500 trees obtained, only 100 were retained,
11 randomly sampled among those where at most one horizontal gene transfer occurred (to
12 allow unequivocal identification of outliers), and for which the transfer (if any) changed the
13 topology of the gene tree. The whole process was repeated three times, varying the
14 maximum number of outliers allowed per gene, between 1, 10 and 53 (theoretical max).
15 This allowed exploring the impact of the number of outliers per gene (linked here to the age
16 of the HGT) on the capacity to correctly identify outliers, *i.e.* the species that were changing
17 position relative to the species tree because of HGT. For ILS, it was not possible to identify
18 precisely what species should be considered as outliers or not. We could only look at the
19 impact of ILS on the mean topological distance between the collection of gene trees and the
20 species tree (mean RF distance between 3 for NO-ILS and more than 20 for HIGH-ILS, see
21 Figure S1) and evaluate whether this had an impact or not on the precision and sensibility of
22 our method.

- 23 ● *Carnivora dataset (CD)*: We used the raw sequence files (before alignment and filtering)
24 from a previously assembled phylogenomic dataset comprising 14,463 genes for 53 species
25 aimed at resolving the phylogeny of the order Carnivora (Allio et al. 2021). This dataset was
26 obtained by extracting single-copy protein-coding orthologous genes from the genomes of
27 52 carnivore species, plus the Malayan pangolin (*Manis javanica*) used as outgroup,
28 following the orthology delineation strategy of the OrthoMaM database (Scornavacca et al.
29 2019). These raw sequence files were aligned and filtered using the OMM_MACSE pipeline
30 (Ranwez et al. 2021), which combines (i) translated nucleotide sequence alignment at the
31 amino acid level with MAFFT (Katoh and Standley 2013), (ii) nucleotide alignment
32 refinement (based on amino acid alignment) with MACSE v2 (Ranwez et al. 2018) to
33 handle frameshifts and non-homologous sequences (Ranwez et al. 2018), and (iii) masking
34 of ambiguously aligned and dubious parts of sequences with HMMcleaner (Di Franco et al.
35 2019). In the original study (Allio et al. 2021), this Carnivora dataset was successfully

1 filtered using an early version of PhylteR allowing the removal of outlier sequences and
2 genes generating abnormally long branches. Therefore, it was a good candidate dataset to
3 test the completely redesigned and improved version of PhylteR presented here.
4

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

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

9 We used the first simulated dataset for illustration purposes only. For the other simulated datasets,
10 *i.e.* for each level of ILS, for different maximum numbers of outlier species per gene (1, 10 and 53)
11 and for each one of the 100 replicates, we ran PhylteR with default parameters and we counted the
12 number of True Positives (TP, outliers that were simulated and that are retrieved), False Positives
13 (FP, outliers that were not simulated but are identified) and False Negative (FN, outliers that were
14 simulated but are not retrieved). From those, we computed the mean precision ($TP/(TP+FP)$) and
15 recall (or sensitivity, $TP/(TP+FN)$) of the outlier identification of PhylteR. An estimate of the
16 expected precision and recall when the same number of outliers were randomly sampled was also
17 computed. To evaluate the impact of ILS-induced incongruences between gene trees on the ability
18 of PhylteR to correctly identify outliers, precision and recall were computed and compared between
19 the four levels of ILS simulated (NO-ILS, LOW-ILS, MODERATE-ILS and HIGH-ILS). Finally,
20 for comparison purposes we performed the same analyses using TreeShrink v1.3.9 (Mai and
21 Mirarab 2018) in place of PhylteR with default parameters for detecting outliers.

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

- 29 ● **Sequence length:** Long sequences were shown to carry more phylogenetic signal than
30 shorter ones (Salichos and Rokas 2013; Shen et al. 2016). To explore the possible
31 enrichment of outliers in short sequences, we computed the length (in bp) of each sequence
32 in each gene MSA, and explored its distribution in outliers.
- 33 ● **Duplication score:** when a sequence in a gene tree is not orthologous to the others but is a
34 paralog, its localization in the gene tree is likely to be incorrect. To have an insight into the
35 level of "paralogousness" of each sequence in the Carnivora dataset, we compared the

1 Carnivora species tree published in Allio et al. (2021) with each one of the 14,463 gene trees
2 using the reconciliation program *ALEml_undated* (Szöllősi et al. 2015). This tool allows
3 inferring the duplications, losses and transfers experienced by a gene by comparing its
4 history (the gene tree) with that of the species (the species tree). Here we inferred only
5 duplications and losses (transfer rate was forced to be 0), we forced the origination of each
6 gene at the root of the species tree (parameter O_R=10000) and we used default values for
7 all other parameters. We then computed the number of duplications inferred from the root to
8 each tip of each gene tree, and normalised this value by the number of nodes encountered.
9 This value represents the normalised number of duplications experienced by each sequence,
10 whose distribution in outliers could be evaluated.

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

1
 2 In order to compare the distributions of values for the different features listed above between outlier
 3 detection methods, we needed lists of outliers of comparable size. The number of outliers retrieved
 4 with default parameters being very different with the two methods using default parameters (7,183
 5 with PhylteR vs 19,643 with TreeShrink, see Table 1), we created two collections of outliers, a
 6 **small** and a **large** one (Table 1). For the **small** collection, we selected a value for the parameter q in
 7 TreeShrink in order to get a number of outliers as close as possible to the number of outliers
 8 obtained with PhylteR default parameters. This was achieved for $q = 0.012$, leading to 7,032
 9 outliers. For the **large** collection, we selected a value of the k (and $k = k2$) parameter in PhylteR
 10 leading to a number of outliers as close as possible to the number of outliers detected with
 11 TreeShrink default parameters. This was achieved for $k = 1.55$, leading to 20,157 outliers.
 12 Parameters used and number of outliers in each collection and with each outlier detection method
 13 are presented in Table 1.

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

16 **Table 1. Collections of outliers used to evaluate PhylteR and compare it to TreeShrink.** The **small** collection is
 17 obtained by tuning the TreeShrink parameters in order to obtain roughly the same number of outliers as with the default
 18 parameters of PhylteR. The **large** collection is obtained in the opposite way.

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

22 It is expected that a tool that accurately removes outliers in phylogenomic datasets should increase
 23 the concordance between the gene trees and the species tree. To evaluate this and compare PhylteR
 24 with randomly sampled sequences and with TreeShrink-identified outliers, we computed the gene
 25 concordance factor (gCF, Minh, Hahn, et al. 2020) as implemented in IQ-TREE version 2.1.3
 26 (Minh, Schmidt, et al. 2020) for every branch in the Carnivora species tree (obtained from Allio et
 27 al. 2021). For each branch of the species tree, this factor indicates the percentage of gene trees in
 28 which this branch is found (among gene trees where this can be computed, or “decisive” trees, see
 29 Minh, Hahn, et al. 2020). gCF was computed according to either the original gene trees (gCF_{init}), or
 30 to a list of gene trees obtained after pruning outliers (four sets of gene trees corresponding to the
 31 four list of outliers in Table 1).

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

9 Results

10 Illustration of the general principle of PhylteR

11 The different steps of the PhylteR process (Figure 1) are illustrated on a simple example dataset
 12 comprising 25 genes for 20 species, with 10 outliers. The main steps are as follows. Individual gene
 13 trees are transformed into individual gene matrices that are then combined into a unique
 14 *compromise* matrix obtained after weighting each matrix by its concordance with the others:
 15 matrices that are poorly correlated with the others have less weight in the creation of the
 16 compromise (Figure S3A-E). This matrix is then projected onto a space on which individual
 17 matrices are projected as well (Figure 2A and S3F). By computing the distance of each species in
 18 each orthologous gene to its reference position in this projection, the two-way reference matrix is
 19 obtained (Figure 2B and S3G). It is from this matrix that outlier sequences can be identified and
 20 removed.

21
 22 **Figure 2. Two objects of the PhylteR process. A:** the compromise matrix is projected into a multidimensional space
 23 (the two first axes only are represented here). This gives the reference position of each species relative to each other
 24 (blue badges with species names on it). Individual gene matrices are projected on the same space (small dots) and the
 25 distance between each gene in each species to its reference position is represented by a line. The red line and the red
 26 arrow identify species t3 in gene 5. This projection is transformed into a 2D matrix (**B**) by computing the distance
 27 between each species in each gene to its reference position (i.e. the length of each line in **A**). The gene \times species matrix
 28 obtained, that we refer to as the 2-way reference matrix (2WR) is used to detect outliers like the one indicated by the red
 29 arrow, corresponding to the red arrow in **A**.

32 PhylteR performs well on simulated examples and is robust to ILS-induced 33 incongruences

34 To evaluate the precision and sensitivity of PhylteR, we used it on four simulated datasets with
 35 increasing levels of ILS (NO-ILS, LOW-ILS, MODERATE-ILS and HIGH-ILS). We also
 36 computed precision and recall on the same datasets using another method, TreeShrink (Mai and

1 Mirarab 2018). Finally, we computed the expected precision and sensitivity if the same number of
 2 sequences identified as outliers by PhylteR and TreeShrink were randomly selected from all
 3 sequences.

4 The outliers that we considered were single species whose position was moved to a new location in
 5 some gene trees because of HGTs. For this type of outliers, we observe that PhylteR performs well,
 6 precision and recall being close to their maximum value 1 (Figure 3). On the other hand, TreeShrink
 7 performs badly, identifying few correct outliers (leading to a mean precision close to 0), but still
 8 detecting a large collection of false positives (leading to a low sensitivity). When increasing the
 9 maximum number of outlier species in each gene tree to 10 (Figure S2A) or to 53 (Figure S2B), we
 10 observe that both the precision and sensitivity of PhylteR slightly decrease while the ones of
 11 TreeShrink increase, but the difference between both remains in clear advantage of PhylteR (in this
 12 specific setting).

13 Of note, the level of ILS has almost no effect on the precision and sensitivity of the PhylteR (and
 14 TreeShrink, even though negative effect would be hard to see when starting from such low
 15 precision and sensitivity values), except when reaching very high ILS (bottom-right panels in
 16 Figure 3 and Figure S2A and S2B). In other words, even when the mean topological distance
 17 between the gene trees and the species tree is multiplied by more than 5, as is the case between the
 18 NO-ILS and the MODERATE-ILS conditions (Figure S1), the precision and sensitivity of PhylteR
 19 for detecting the outliers simulated by HGTs do not decrease. This suggests that PhylteR does not
 20 consider species that have changed position in some gene trees due to ILS as outliers. This apparent
 21 robustness of PhylteR to ILS can be seen as a desirable feature, e.g. when using species tree
 22 reconstruction tools that explicitly handle ILS such as ASTRAL (Zhang et al. 2018).

23

24

25 **Figure 3. Comparison of the precision and recall (or sensitivity) of the PhylteR and the TreeShrink outlier**
 26 **detection methods for four conditions of Incomplete Lineage Sorting (ILS).** .

27

28 **Characterisation of outliers detected with PhylteR on the Carnivora dataset**

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

34 In the set of 14,463 gene trees analysed by PhylteR, two sets of outliers (7,183 and 20,157
 35 sequences) were identified with PhylteR (with default or tuned parameters, respectively) and 7,032

1 and 19,643 with TreeShrink (with tuned and default parameters respectively, see Table 1). A simple
2 comparison of the list of outliers of similar sizes revealed that the overlap between the two lists of
3 outliers was quite small (around 20%, Figure 4). This corresponds to about 70% of the outliers
4 detected by PhylteR being absent from the list of outliers detected by TreeShrink, and vice versa.
5 This reveals fundamental differences between the two approaches.

6

7 **Figure 4. Comparison of the sets of outliers detected by PhylteR (left column) and TreeShrink (right column) on**
8 **the Carnivora dataset.** The two collections of outliers (small and large) correspond to different stringency for the
9 detection of outliers (see Table 1).

10

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

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

22

23 Second, we compared the distribution of duplication scores in the list of outliers produced by
24 PhylteR and TreeShrink (Figure 5B). We observed a clear difference, for both the small and the
25 large collections of outliers between PhylteR outliers and random outliers, but also between PhylteR
26 outliers and TreeShrink outliers: outliers identified by PhylteR are significantly enriched in
27 sequences that display a higher number of duplications as compared to random or TreeShrink
28 outliers ($p < 2.2e-16$ for all comparisons).

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

32

33 One illustration of the difference between PhylteR and TreeShrink in their ability to capture
34 duplicated sequences (and thus probably hidden paralogues) can be given by the study of peculiar
35 proteins, such as the Zinc-finger family (ZNF). This large family of paralogs first duplicated from
36 the gene PRDM9 or PRDM7 in the ancestor of vertebrates (Emerson and Thomas 2009). These

1 genes are involved in the repression of transposable elements and are still actively duplicating. The
 2 high number of duplications renders the resolution of the orthology relationship in this gene super-
 3 family very challenging. In the Carnivora dataset, the ZNF super-family has been splitted in 168
 4 orthologous gene families (Allio et al. 2021). As expected in case of hidden paralogy, we see an
 5 overrepresentation of the genes belonging to these families in the list of outliers, especially in the
 6 outliers identified by PhylteR (Figure 5C). Between 3.79% (for the large set) and 7.4% (for the
 7 small set) of PhylteR outliers belong to the ZNF family, while these values drop to 1.78% and
 8 1.12% respectively for TreeShrink outliers, and less than 1% for randomly selected sequences
 9 (Figure 5C).

10

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

17

18

19 Third, we compared two by two 14 Carnivora species and identified syntenic outliers (see material
 20 and methods). In almost all pairwise comparisons, we found that these syntenic outliers
 21 significantly overlap the outlier sequences detected by PhylteR. For example, in the comparison
 22 between *Zalophus californianus* and *Suricata suricatta* (illustrated in Figure 6), out of the 5,123
 23 genes common to both species in the dataset, 131 (2.56%) are syntenic outliers (i.e. surrounded by
 24 two breaks). In comparison, out of the 47 outlier sequences identified by PhylteR (small list) in
 25 either *Zalophus californianus* or *Suricata suricatta*, 38 are syntenic outliers (80.8%), which is
 26 significantly more than expected by chance (p-value = 1.5e-43). With TreeShrink (small list) for the
 27 same pair of species, only 18.1% (17 out of 94) outlier sequences are syntenic outliers, which is
 28 much less than with PhylteR but is still significantly different from what is expected by chance (p-
 29 value = 1.36e-10). Similar results were obtained for most of the other pairs of species compared
 30 (Figure S4 and Supplementary Tables S2 and S3).

31

32 **Figure 6. Illustration of the non-syntenic nature of many outliers identified by PhylteR.** We represent the
 33 comparison of *Zalophus californianus* with *Suricata suricatta* genomes, with *Zalophus* as a reference (arbitrarily, most
 34 other pairs of species give similar results). On each circle, a reference *Zalophus* scaffold is represented in dark blue, and
 35 all scaffolds for which at least one gene has an ortholog in this scaffold are in light grey. Lines between these scaffolds
 36 represent couples of genes annotated as orthologous. Red lines highlight gene outliers detected in *Suricata suricatta*.
 37 We observe that they are very often “isolated” genes, i.e. syntenic outliers. These genes are thus probably erroneously
 38 annotated, erroneously assembled, and their orthology is likely erroneous.

39

1 **Impact of filtering outliers on Species Tree support**

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

6 Non-random outlier removal processes are expected to increase gCF scores by discarding sequences
7 representing species in gene trees whose position is not in accordance with their placement in the
8 other gene trees. We looked at the difference in gCF score before and after pruning outliers (Δ gCF)
9 for each branch of the Carnivora species tree. For both PhylteR and TreeShrink, an increase in gene
10 concordance was observed. It was higher with PhylteR than with TreeShrink, indicating a better
11 identification of misplaced species in gene trees for PhylteR. The effect was larger when more
12 outliers were removed (Figure 7, right), the gain in gCF reaching more than 6% for some branches
13 with PhylteR outliers removal (max 5% for TreeShrink). We observed that the gain in concordance
14 was higher for branches that initially had a high gCF, and smaller for poorly supported nodes (plain
15 dots versus circles in Figure 7). This may be due to an easier identification of outliers on a ‘clean’
16 background (many gene trees supporting the same node, leading to high gCF) than on a more noisy
17 one.

18 Note that gCF, which captures topological differences between the gene trees and the species tree,
19 exhibits a notable increase but does not attain its maximum value. This observation might be
20 indicative of some of the incongruences between gene and species trees within the Carnivora
21 dataset to be attributed to ILS. These potential ILS-related incongruences appear not to be identified
22 as outliers by PhylteR, as suggested by the results of our simulations (see above).

23
24 **Figure 7. Effect of filtering outliers in gene trees on the gene concordance factor (gCF) of each branch of the**
25 **Carnivora species tree.** The gain in concordance (Δ gCF, y-axis) is plotted for each branch of the species tree (dots),
26 separating PhylteR (pink) and TreeShrink (blue). Branches are ordered by increasing Δ gCF for the PhylteR outliers.
27 The results for the two collections of outliers (small and large) are displayed side by side.

29 **Discussion**

30 In phylogenomics, incongruence between gene trees, resulting from a myriad of possible technical
31 and analytical issues, or from biological processes, is known to lead to errors in species tree
32 inference (Philippe et al. 2017). A common practice in phylogenomics thus consists of scanning
33 individual gene trees by eye, trying to spot species or group of species weirdly placed in gene trees,
34 suspicious long branches, apparent groups of paralogues, etc. and discarding them prior to the
35 concatenation of the genes (supermatrix approach) or to the assembly of the gene trees into a

1 species tree (supertree and coalescent-based approaches). This hard work is not only time-
2 consuming and laborious, it is also questionable: what is the objectivity in this practice? Is the eye
3 (and the brain) capable of looking at tens of thousands of gene trees at the same time? How
4 reproducible is such a practice? Etc.

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

11 Evaluating a method for its capacity to accurately identify errors in phylogenomics datasets is a
12 difficult task. As for any inference method, we use simulations. However, simulating the processes
13 that result in errors (in our case, outliers in phylogenomics data) has no standard solution: sources
14 of errors are numerous, they combine with each other through all phylogenomic pipelines,
15 sometimes with unpredictable results. So we restricted ourselves to simulating a feature intrinsically
16 detectable by PhylteR, that is, changes in the phylogenetic placements of some species in some
17 gene trees. Further evaluation would involve an independent simulation pipeline, not informed by
18 the hypothesis behind the inference method (Biller et al. 2016), which is by definition outside the
19 scope of the description of the inference method. The simple simulations we performed revealed
20 that outliers corresponding to misplacement of species in a few gene trees was easy to detect with
21 PhylteR but not with TreeShrink. However, this is not surprising *a posteriori*: TreeShrink (Mai and
22 Mirarab 2018) is designed to detect abnormally long branches in collection of gene trees, while we
23 considered here as outliers species that changed position in some gene trees because of horizontal
24 gene transfers; these outliers are not necessarily associated with longer branches.

25 A better way to examine the advantages of a method over another is to explore biological data.

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

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

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

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

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

28 Incomplete Lineage Sorting (ILS) is a known source of incongruence among gene trees and
29 between gene trees and the species tree. This biological process, where ancestral polymorphism is
30 maintained across various speciation events, leads to different portions of the genomes having
31 different evolutionary histories. With simulations we could vary the level of Incomplete Lineage
32 Sorting in the datasets and evaluate the impact it had on the ability of PhylteR (and TreeShrink) to
33 correctly identify outliers. For both methods, we saw no effect of increasing the level of ILS on the
34 precision and sensitivity values. For TreeShrink, it is hard to conclude anything, because the initial
35 values were very low (close to 0) so that any negative effect would have been undetectable. For

1 PhylteR however, where precision and sensitivity were high, this absence of effect reveals that
2 PhylteR does not detect sequences that have experienced ILS as outliers. Whether this is positive or
3 negative can be discussed. On the one hand, it was shown earlier that ILS-related incongruences
4 among gene trees could have a detrimental effect for species tree reconstruction with supermatrix
5 approaches (Degnan and Rosenberg 2006). On the other hand, ILS-induced incongruence is a true
6 biological signal that many species tree inference methods, namely the coalescence-based ones, can
7 now handle (Liu et al. 2010; Zhang et al. 2018). In this context, getting rid of these incongruences
8 may be seen as detrimental, because it removes a meaningful signal that can be accommodated by
9 these methods. However, the real contribution of ILS to gene tree incongruences is something that
10 is rarely measured, but in mammals for instance, it was shown to be rare (Scornavacca and Galtier
11 2017). We can advance two reasons why PhylteR is not sensitive to ILS. First, ILS preferentially
12 affects short branches of the species tree, i.e. speciation events separated by a short amount of time,
13 which leads to a limited effect in the pairwise distance matrices manipulated by PhylteR. Second,
14 when ILS changes the branching pattern of three clades (or species), it is expected that around 50%
15 of each alternative topology to the true one is observed across all gene trees. When the
16 “compromise” matrix is built in the PhylteR pipeline, this signal will thus likely be averaged out.

17
18 Here we focused on the identification of outliers in collection of gene trees in order to remove them
19 prior to phylogenetic inference with supermatrix, supertree or coalescent-based methods. But other
20 usage of the tool we present here can be anticipated. First, because the PhylteR method consists of
21 comparing matrices (in this case phylogenetic distance matrices), it is easy to imagine applying the
22 method without computing gene trees, directly on matrices extracted from multiple sequence
23 alignments (MSA), one matrix per gene. In this sense, comparing PhylteR with MSA-based filtering
24 tools could be a worthwhile follow-up of this work. Second, correctly identifying and removing
25 outliers from phylogenomic datasets could be of interest beyond species tree reconstruction. For
26 instance, it appears to be crucial when using statistical methods based on the ratio of
27 nonsynonymous over synonymous substitution rates (d_N/d_S ratio) to detect adaptive molecular
28 evolution (see Yang and Bielawski 2000 for a review), or for correctly inferring ancestral sequences
29 (Yang et al. 1995) from sequences of extant species. Finally, using a tool like PhylteR is not only
30 useful for cleaning the data. The in-depth exploration of the outliers detected and the study of the
31 reasons why they were detected as such can give important insights into the evolutionary history of
32 these sequences, for instance allowing for the identification of horizontally transferred or duplicated
33 genes.

1 **Conclusion**

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

8 **Acknowledgments**

9 Work was funded by ANR Grant 18-CE02-0007 (Sthoriz) to DMDV, ANR Grant 19-CE45-0010
10 (Evolution) to ET, and European Research Council grant ERC-2015-CoG-683257 (ConvergeAnt
11 project) to FD. This is contribution ISEM 2023-XXX of the Institut des Sciences de l'Evolution de
12 Montpellier. We thank Mélodie Bastian for help with the simulations and two anonymous reviewers
13 for helpful comments.

14 **Software availability**

15 PhylteR is a package written in R language (R Core Team 2023) available on CRAN ([https://cran.r-](https://cran.r-project.org/web/packages/phylter/index.html)
16 [project.org/web/packages/phylter/index.html](https://cran.r-project.org/web/packages/phylter/index.html)) for the latest stable version and on GitHub
17 (<https://github.com/damiendevienne/phylter>) for the latest development version. The latest version
18 of PhylteR is also distributed as a Singularity container
19 (https://cloud.sylabs.io/library/theo.treecou/tool/phylter_singularity) and a docker container
20 (https://hub.docker.com/r/treecoutheo/phylter_docker). Extensive documentation can be found at
21 <https://damiendevienne.github.io/phylter/index.html>.

22 **Data Availability**

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

26 **References**

- 27 Abdi H, O'Toole AJ, Valentin D, Edelman B. 2005. DISTATIS: The analysis of multiple distance
28 matrices. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern
29 Recognition (CVPR'05)-Workshops. IEEE. p. 42–42.
30 Abdi H, Williams LJ, Valentin D, Bennani-Dosse M. 2012. STATIS and DISTATIS: optimum
31 multitable principal component analysis and three way metric multidimensional scaling.
32 *Wiley Interdiscip. Rev. Comput. Stat.* 4:124–167.

- 1 Aberer AJ, Krompass D, Stamatakis A. 2013. Pruning Rogue Taxa Improves Phylogenetic
2 Accuracy: An Efficient Algorithm and Webservice. *Syst. Biol.* 62:162–166.
- 3 Ali RH, Bogusz M, Whelan S. 2019. Identifying Clusters of High Confidence Homologies in
4 Multiple Sequence Alignments. *Mol. Biol. Evol.* 36:2340–2351.
- 5 Allio R, Tilak M-K, Scornavacca C, Avenant NL, Kitchener AC, Corre E, Nabholz B, Delsuc F.
6 2021. High-quality carnivoran genomes from roadkill samples enable comparative species
7 delineation in aardwolf and bat-eared fox. Perry GH, Perry GH, editors. *eLife* 10:e63167.
- 8 Barton D, David F. 1956. Some notes on ordered random intervals. *J. R. Stat. Soc. Ser. B Methodol.*
9 18:79–94.
- 10 Baum BR. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and
11 the desirability of combining gene trees. *TAXON* 41:3–10.
- 12 Biller P, Knibbe C, Beslon G, Tannier E. 2016. Comparative genomics on artificial life. In:
13 Conference on Computability in Europe. Springer. p. 35–44.
- 14 Brys G, Hubert M, Struyf A. 2004. A Robust Measure of Skewness. *J. Comput. Graph. Stat.*
15 13:996–1017.
- 16 Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment
17 trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.
- 18 Chessel D, Hanafi M. 1996. Analyses de la co-inertie de K nuages de points. *Rev. Stat.*
19 *Appliquée* 44:35–60.
- 20 Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new
21 software for selection of phylogenetic informative regions from multiple sequence
22 alignments. *BMC Evol. Biol.* 10:210.
- 23 Dalevi D, Eriksen N. 2008. Expected gene-order distances and model selection in bacteria.
24 *Bioinformatics* 24:1332–1338.
- 25 Degnan JH, Rosenberg NA. 2006. Discordance of Species Trees with Their Most Likely Gene
26 Trees. *PLOS Genet.* 2:e68.
- 27 Di Franco A, Poujol R, Baurain D, Philippe H. 2019. Evaluating the usefulness of alignment
28 filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol. Biol.*
29 19:1–17.
- 30 Emerson RO, Thomas JH. 2009. Adaptive Evolution in Zinc Finger Transcription Factors. *PLoS*
31 *Genet.* 5:e1000325.
- 32 von Haeseler A. 2012. Do we still need supertrees? *BMC Biol.* 10:13.
- 33 Hubert M, Vandervieren E. 2008. An adjusted boxplot for skewed distributions. *Comput. Stat. Data*
34 *Anal.* 52:5186–5201.
- 35 Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, Kim J, Gordon L, Branscomb E,
36 Stubbs L. 2006. A comprehensive catalog of human KRAB-associated zinc finger genes:
37 Insights into the evolutionary history of a large family of transcriptional repressors. *Genome*
38 *Res.* 16:669–677.
- 39 Jackson DA. 1993. Stopping rules in principal components analysis: a comparison of heuristical and
40 statistical approaches. *Ecology* 74:2204–2214.
- 41 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
42 improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780.
- 43 Kerk M van de, Kroon H de, Conde DA, Jongejans E. 2013. Carnivora Population Dynamics Are as
44 Slow and as Fast as Those of Other Mammals: Implications for Their Conservation. *PLOS*
45 *ONE* 8:e70354.
- 46 Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci. U.*
47 *S. A.* 99:803–808.
- 48 Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: Phylogenetic Reconstruction
49 with Infinite Mixtures of Profiles in a Parallel Environment. *Syst. Biol.* 62:611–615.
- 50 Liu H, Chang L-H, Sun Y, Lu X, Stubbs L. 2014. Deep Vertebrate Roots for Mammalian Zinc
51 Finger Transcription Factor Subfamilies. *Genome Biol. Evol.* 6:510–525.
- 52 Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species

- 1 trees under the coalescent model. *BMC Evol. Biol.* 10:302.
- 2 Lucas JM, Crollius HR. 2017. High precision detection of conserved segments from synteny blocks.
3 *PLOS ONE* 12:e0180198.
- 4 Mai U, Mirarab S. 2018. TreeShrink: fast and accurate detection of outlier long branches in
5 collections of phylogenetic trees. *BMC Genomics* 19:272–272.
- 6 Mallo D, De Oliveira Martins L, Posada D. 2016. SimPhy : Phylogenomic Simulation of Gene,
7 Locus, and Species Trees. *Syst. Biol.* 65:334–344.
- 8 Minh BQ, Hahn MW, Lanfear R. 2020. New Methods to Calculate Concordance Factors for
9 Phylogenomic Datasets. *Mol. Biol. Evol.* 37:2727–2733.
- 10 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R.
11 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the
12 Genomic Era. *Mol. Biol. Evol.* 37:1530–1534.
- 13 Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary
14 analyses in R. *Bioinformatics* 35:526–528.
- 15 Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011.
16 Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLOS*
17 *Biol.* 9:e1000602.
- 18 Philippe H, de Vienne DM, Ranwez V, Roure B, Baurain D, Delsuc F. 2017. Pitfalls in supermatrix
19 phylogenomics. *Eur. J. Taxon.* 283:1–25.
- 20 R Core Team. 2023. R: A Language and Environment for Statistical Computing. Vienna, Austria: R
21 Foundation for Statistical Computing Available from: <https://www.R-project.org/>
- 22 Ragan MA. 1992. Matrix representation in reconstructing phylogenetic relationships among the
23 eukaryotes. *Biosystems* 28:47–55.
- 24 Ranwez V, Chantret N, Delsuc F. 2021. Aligning Protein-Coding nucleotide sequences with
25 MACSE. In: Multiple Sequence Alignment. Springer. p. 51–70.
- 26 Ranwez V, Chantret NN. 2020. Strengths and limits of multiple sequence alignment and filtering
27 methods.
- 28 Ranwez V, Douzery EJ, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: toolkit for the
29 alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.*
30 35:2582–2584.
- 31 Robert P, Escoufier Y. 1976. A Unifying Tool for Linear Multivariate Statistical Methods: The RV-
32 Coefficient. *J. R. Stat. Soc. Ser. C Appl. Stat.* 25:257–265.
- 33 Ronquist F. 1996. Matrix representation of trees, redundancy, and weighting. *Syst. Biol.* 45:247–
34 253.
- 35 Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic
36 signals. *Nature* 497:327–331.
- 37 Schrempf D, Szöllösi G. 2020. The sources of phylogenetic conflicts. *Phylogenetics Genomic*
38 *Era*:3–1.
- 39 Scornavacca C, Belkhir K, Lopez J, Derrat R, Delsuc F, Douzery EJP, Ranwez V. 2019.
40 OrthoMaM v10: Scaling-Up Orthologous Coding Sequence and Exon Alignments with
41 More than One Hundred Mammalian Genomes. *Mol. Biol. Evol.* 36:861–862.
- 42 Scornavacca C, Galtier N. 2017. Incomplete Lineage Sorting in Mammalian Phylogenomics. *Syst.*
43 *Biol.* 66:112–120.
- 44 Shen X-X, Salichos L, Rokas A. 2016. A Genome-Scale Investigation of How Sequence, Function,
45 and Tree-Based Gene Properties Influence Phylogenetic Inference. *Genome Biol. Evol.*
46 8:2565–2580.
- 47 Szöllösi GJ, Davín AA, Tannier E, Daubin V, Boussau B. 2015. Genome-scale phylogenetic
48 analysis finds extensive gene transfer among fungi. *Philos. Trans. R. Soc. B Biol. Sci.*
49 370:20140335.
- 50 de Vienne DM, Ollier S, Aguilera G. 2012. Phylo-MCOA: A Fast and Efficient Method to Detect
51 Outlier Genes and Species in Phylogenomics Using Multiple Co-inertia Analysis. *Mol. Biol.*
52 *Evol.* 29:1587–1598.

- 1 Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol.*
2 *Evol.* 15:496–503.
- 3 Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid
4 sequences. *Genetics* 141:1641–1650.
- 5 Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree
6 reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19:153.
- 7 Zhang C, Zhao Y, Braun EL, Mirarab S. 2021. TAPER: Pinpointing errors in multiple sequence
8 alignments despite varying rates of evolution. *Methods Ecol. Evol.* 12:2145–2158.

9
10

ACCEPTED MANUSCRIPT

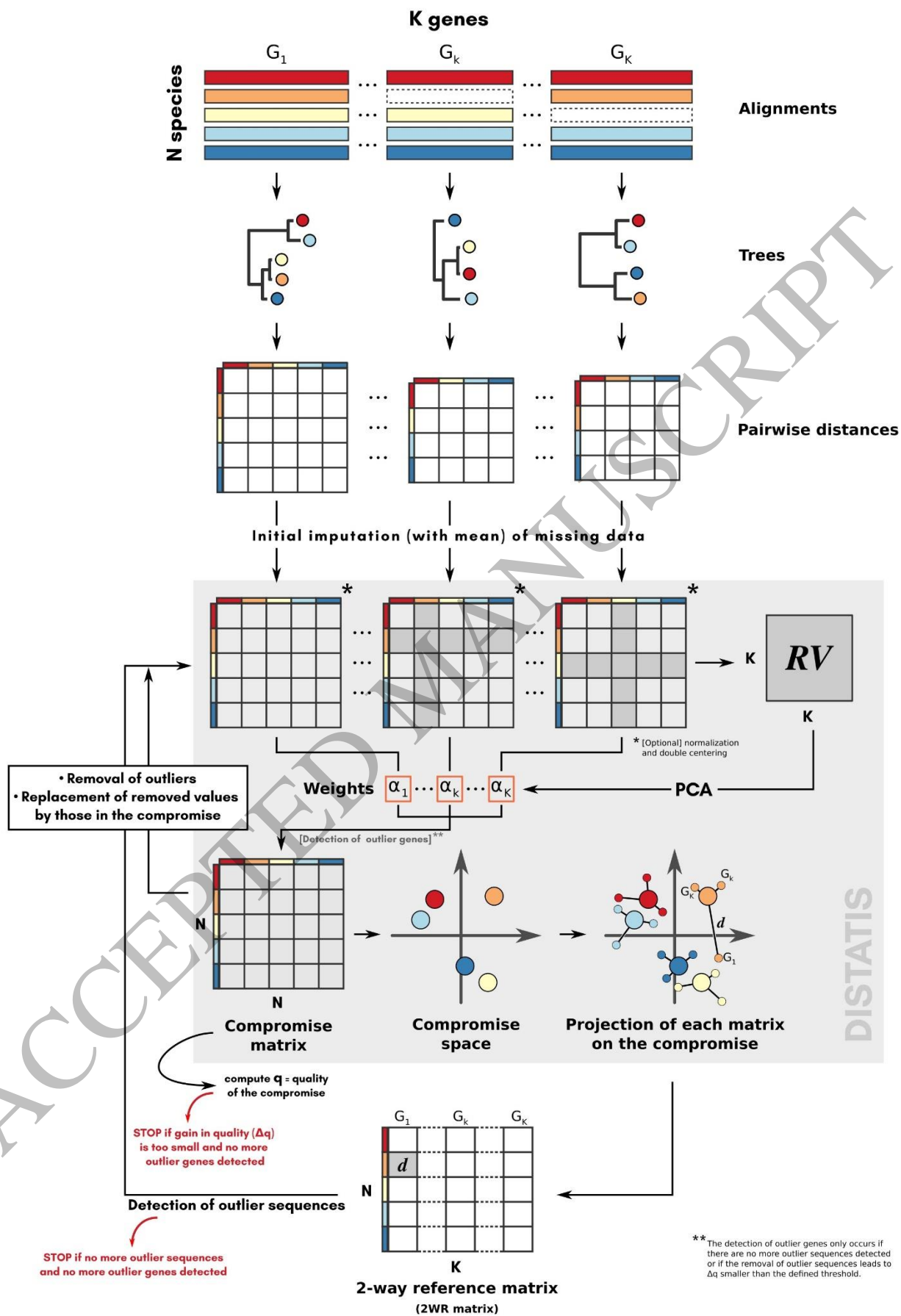
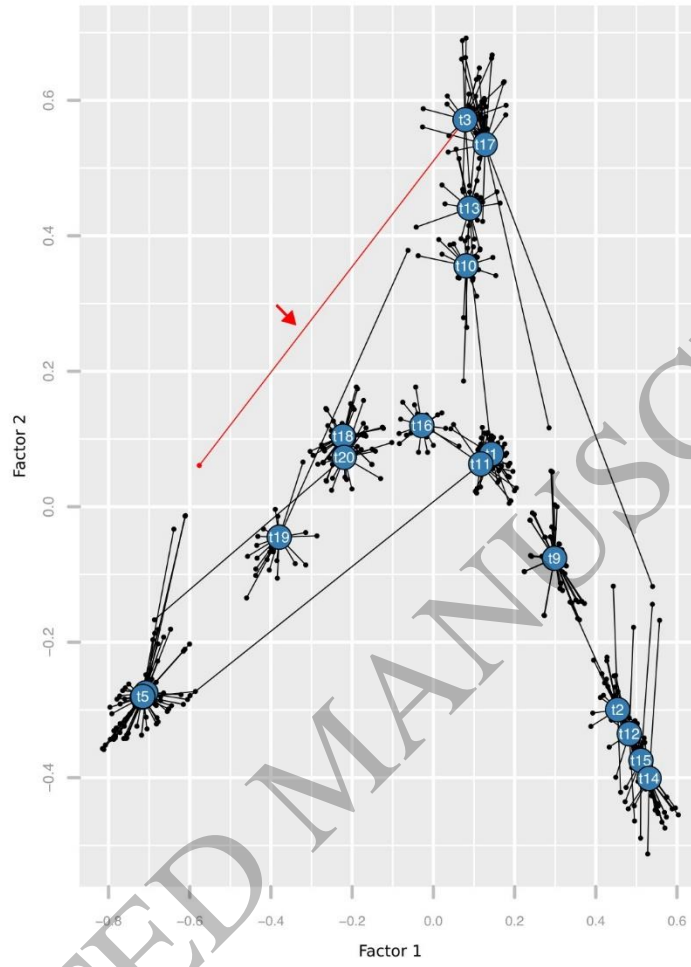


Figure 1
210x297 mm (x DPI)

1
2
3

A. Projection of matrices on the compromise space



B. 2-way reference matrix

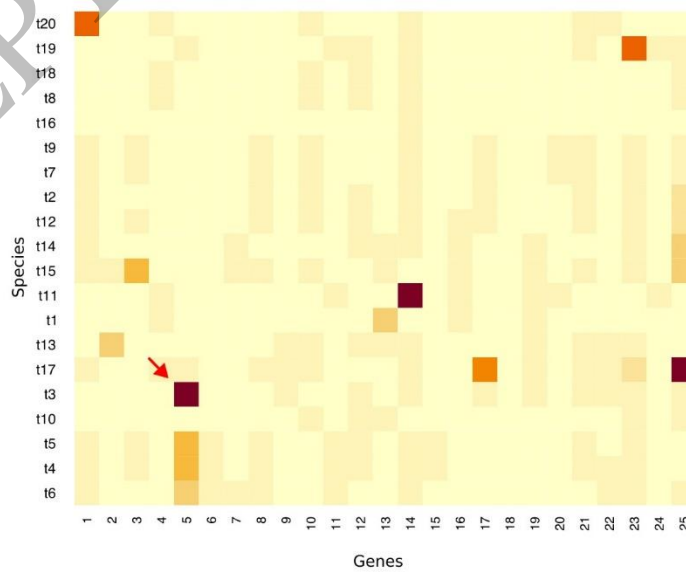


Figure 2
105x234 mm (x DPI)

1

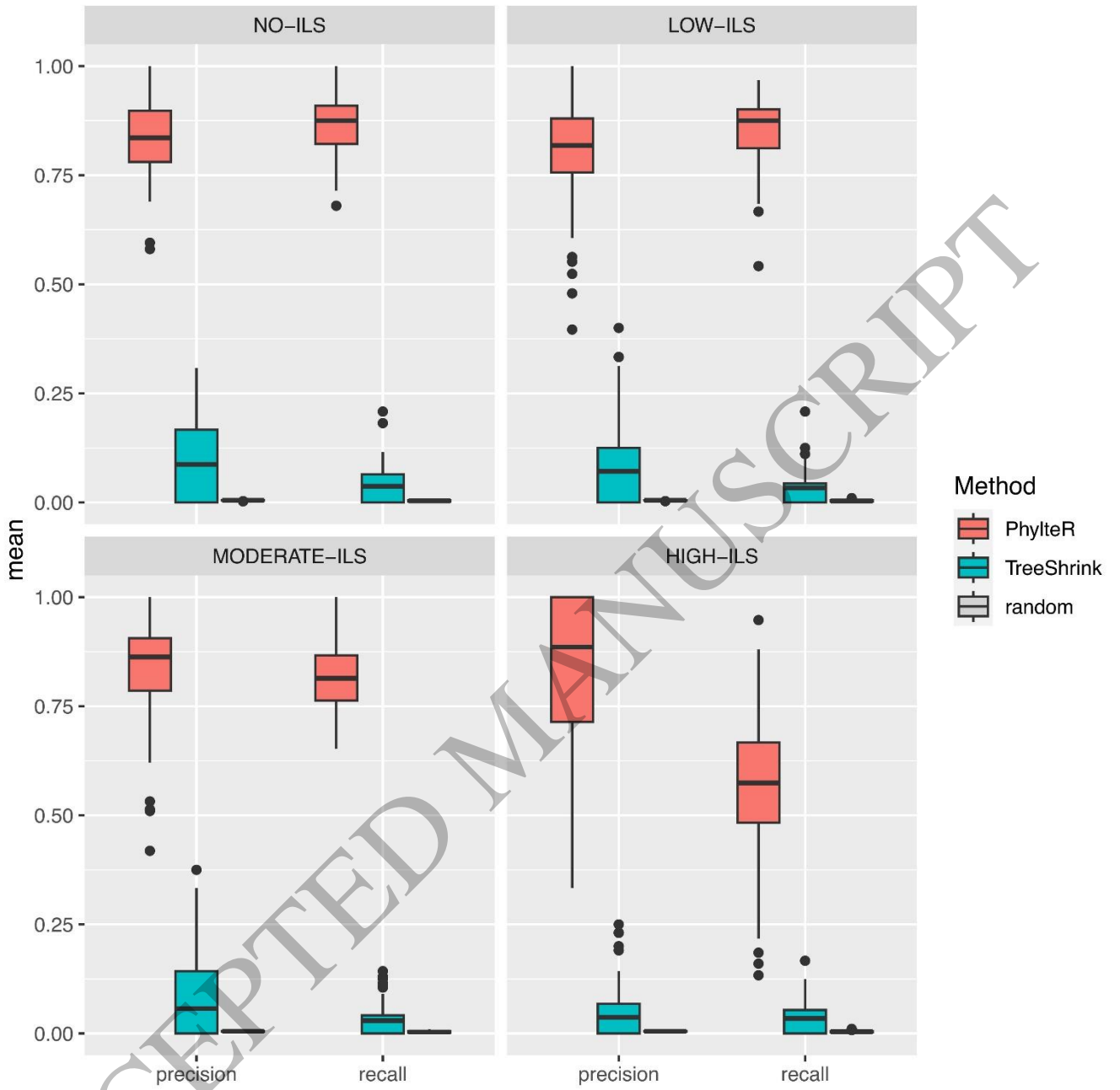


Figure 3
178x178 mm (x DPI)

2

3

4

5

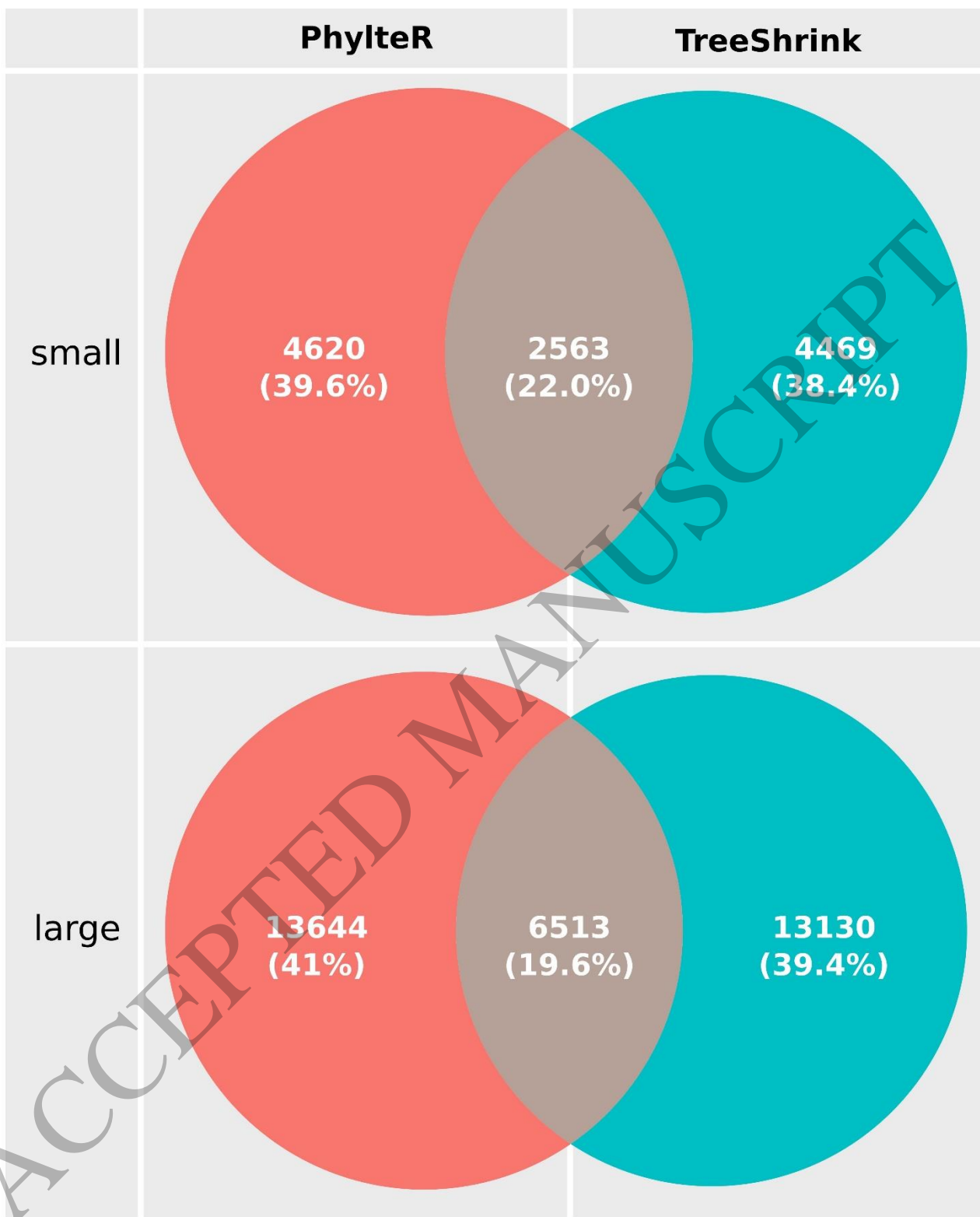


Figure 4
172x211 mm (x DPI)

1
2
3
4

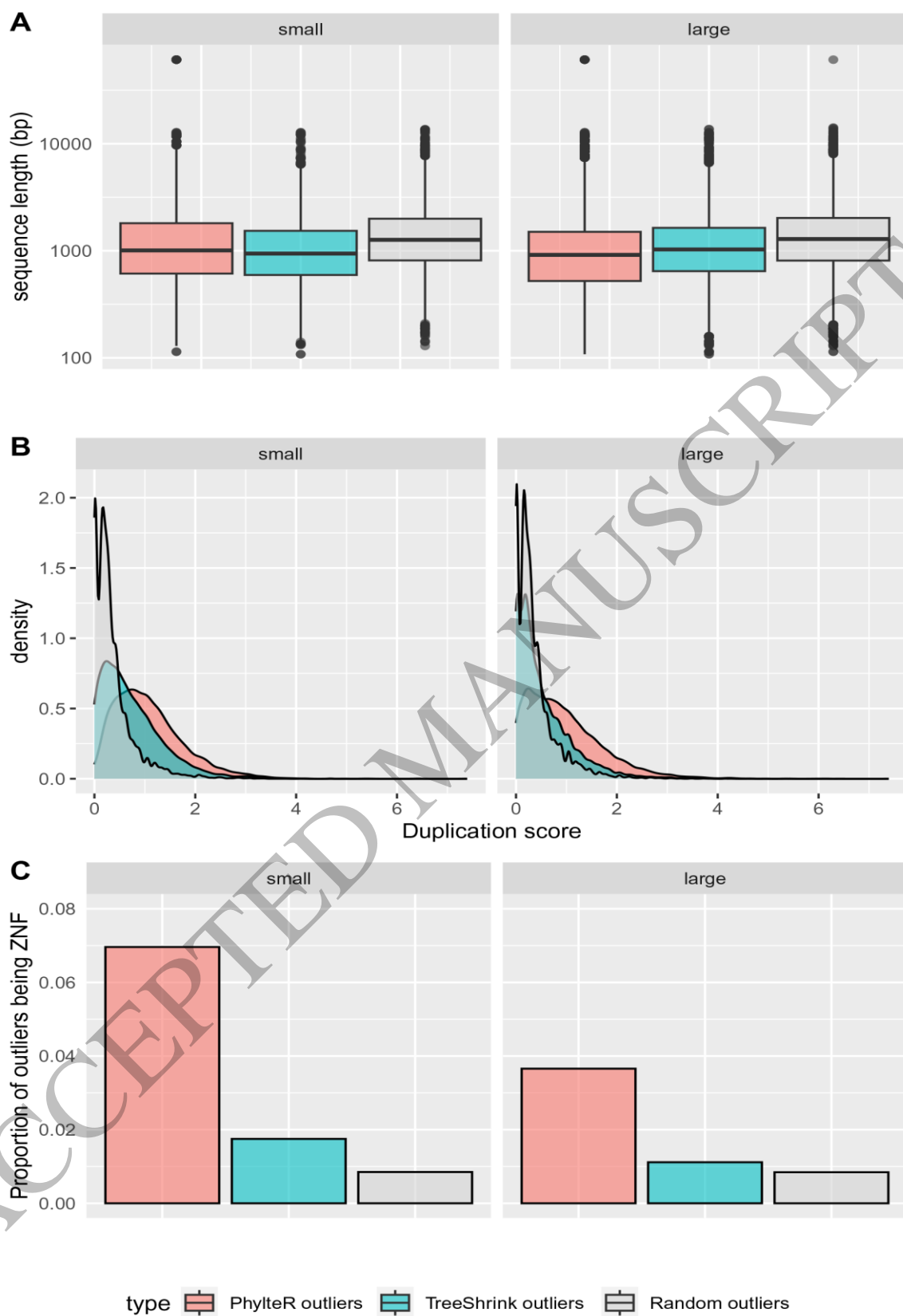


Figure 5
152x254 mm (x DPI)

1
2
3
4

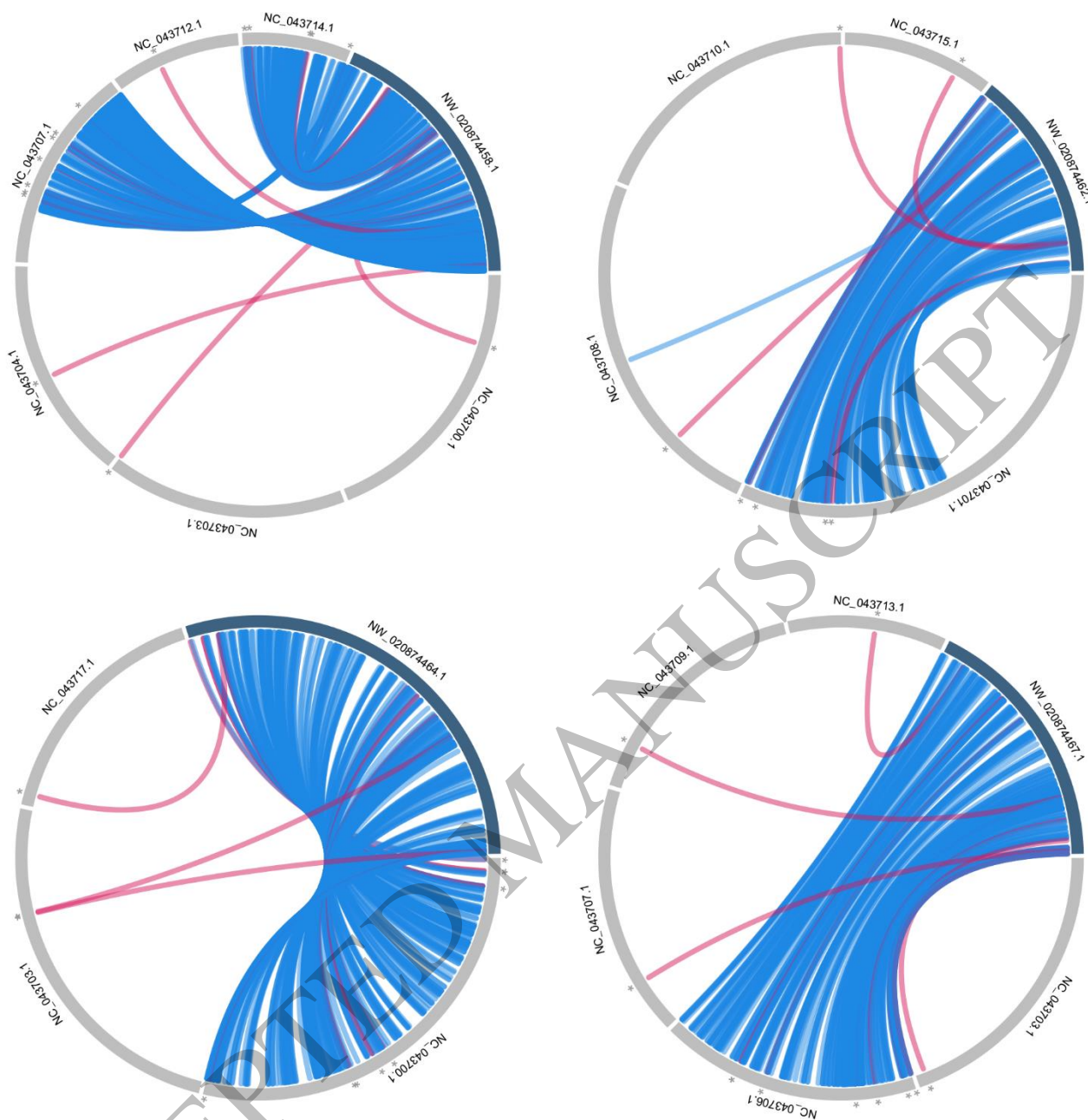


Figure 6
168x169 mm (x DPI)

1
2
3
4



1
2
3

Figure 7
178x178 mm (x DPI)