2018

CURatio: Genome-wide phylogenomic analysis method using ratios of total branch lengths

Kang, Qiwen; Moore, Neil; Schardl, Christopher L.;
Yoshida, Ruriko

IEEE

Kang, Qiwen, et al. "CURatio: Genome-wide phylogenomic analysis method using ratios of total branch lengths." IEEE/ACM transactions on computational biology and bioinformatics (2018
http://hdl.handle.net/10945/64577

CURatio: Genome-wide phylogenomic analysis method using ratios of total branch lengths

Qiwen Kang, Neil Moore, Christopher L. Schardl, and Ruriko Yoshida,

Abstract—Evolutionary hypotheses provide important underpinnings of biological and medical sciences, and comprehensive, genome-wide understanding of evolutionary relationships among organisms are needed to test and refine such hypotheses. Theory and empirical evidence clearly indicate that phylogenies (trees) of different genes (loci) should not display precisely matching topologies. The main reason for such phylogenetic incongruence is reticulated evolutionary history of most species due to meiotic sexual recombination in eukaryotes, or horizontal transfers of genetic material in prokaryotes. Nevertheless, many genes should display topologically related phylogenies, and should group into one or more (for genetic hybrids) clusters in poly-dimensional "tree space". Unusual evolutionary histories or effects of selection may result in "outlier" genes with phylogenies that fall outside the main distribution(s) of trees in tree space. We present a new phylogenomic method, `CURatio`, which uses ratios of total branch lengths in gene trees to help identify phylogenetic outliers in a given set of ortholog groups from multiple genomes. An advantage of `CURatio` over other methods is that genes absent from and/or duplicated in some genomes can be included in the analysis. We conducted a simulation study under the coalescent model, and showed that, given sufficient species depth and topological difference, these ratios are significantly higher for the "outlier" gene phylogenies. Also, we applied `CURatio` to a set of annotated genomes of the fungal family, Clavicipitaceae, and identified alkaloid biosynthesis genes as outliers, probably due to a history of duplication and loss. The source code is available at https://github.com/QiwenKang/CURatio, and the empirical data set on Clavicipitaceae and simulated data set are available at Mendeley https://data.mendeley.com/datasets/mrxts7wjrr/1.

Index Terms—Evolutionary models, Gene trees, Likelihood functions, Outliers, Phylogenomics, Species trees

1 Introduction

In recent decades the field of phylogenetics has found applications in the analysis of genomic scale data (phylogenomics). In particular, it has been applied to analyze the relationships between species and populations, genome evolution, and the evolutionary processes of speciation and molecular evolution. However, today, we can generate genomic data so cheaply and quickly that we encounter a new problem: the sheer volume of genomic data and the lack of analytical tools for working with such quantities of data.

It is well-known that incomplete lineage sorting leads to differences in phylogenetic tree topologies among gene trees [1], [2], [3], [4]. Therefore, a key issue in systematic biology is to reconstruct the evolutionary history of populations and species from numerous gene trees with varying levels of discordance [5], [6].

Even though there has been much work in discordant phylogenetic relationships [1], [7], [8], [9], it is only recently that researchers have shifted away from single gene or concatenated gene estimates of phylogeny towards these multiple gene (multilocus) approaches, e.g., [10], [11], [12], [13], [14]. For example, researchers have begun to consider the effect of genetic drift in producing patterns of incomplete lineage sorting and gene tree/species tree discordance, largely using coalescent theory [15], [16], [17], [18], [19], [20], [21]. Other research has addressed the reconstruction of species trees from the distribution of estimated gene trees [22], [23], [24], [25], [26], [27], [28], [29], [30].

It is well-known that several processes can reduce the correlation among gene trees, including negative or balancing selection [31], meiotic sexual recombination in eukaryotes [32], and horizontal transfers of genetic material especially in prokaryotes [33]. Such processes can strongly influence phylogenetic/species tree reconstruction from the distribution of gene trees [6], [32], [34].

In this paper we propose a method to detect outlier genes from the distribution of gene trees from multiple whole-genome analysis. Here, we focus on the problem of *discordance* among gene trees, and the distribution of gene trees as a whole. We view "typical" gene trees as samples from some distribution f (e.g., a coalescent model) that generates gene trees as independent samples. We also suppose that there may be "atypical" outlier gene trees that in effect are sampled from some other distribution f' very different from f. We are interested in estimating the distribution f for typical gene trees, and also identifying outlier gene trees that were probably not generated by f. Trees identified as outliers can be inspected for biologically interesting properties or evolutionary histories. Also, identifying and removing outliers that violate model assumptions can improve inferences made from collections of gene trees.

Here we propose the `CURatio` method based on ratios of total branch lengths in unconstrained and constrained

- *R. Yoshida is with the Department of Operations Research, Naval Postgraduate School.*
  E-mail: ryoshida@nps.edu
- *Q. Kang is with the Department of Statistics, University of Kentucky.*
- *C.L. Schardl is with the Department of Plant Pathology, University of Kentucky.*
- *N. Moore is with the Department of Computer Science, University of Kentucky.*

Manuscript revised March 2, 2018.

gene trees. The CURatio method is a statistical test used to compare the goodness of fit of two models: the null model and an alternative model. In this paper, the null model is the evolutionary model constrained to a fixed (e.g., species) tree topology (the "constraint tree") and the alternative is the evolutionary model unconstrained to any fixed tree topology. If a gene tree follows the constraint tree, the ratio between these models should be close to one. If it does not, this ratio should be significantly greater than one. Here we demonstrate the method on simulated data sets, as well as an empirical set from 12 genomes in the fungal family Clavicipitaceae.

2 Methods

2.1 Test statistics

For each gene tree, we consider the following hypotheses:

$$H_0 : \text{A gene tree with the data } \mathcal{D} \text{ is congruent to the given tree topology } \tau.$$
$$H_1 : \text{A gene tree with the data } \mathcal{D} \text{ is not congruent to the given tree topology } \tau.$$

In this paper we are testing these hypotheses using the *ratio between the total branch lengths in the constrained and unconstrained trees*.

Under the maximum likelihood estimation (MLE), branch lengths in a tree are the expected number of mutations per site in certain time period. This means that the total branch length of a tree under the MLE is the expectation of the total number of mutations per site over the certain time period.

Our objective is to test how a gene tree fits a given species tree topology. If the tree topology τ is not the "best" tree topology for the observed dataset and for a given evolutionary model, then the expected number of mutations per site would increase to fit the data to the given tree topology τ. Thus the total branch length would increase if τ is not well-fitted to the given observed data under the given evolutionary model.

Therefore, with the given data set, we used R package ape [35] to infer the MLE tree T' under the null hypothesis H_0 by constraining the tree to have topology τ under the given model, and we infer the MLE tree T under the alternative hypothesis H_1 by not constraining the tree topology (i.e., finding the optimal tree topology under the model). We calculate the ratio, [Lambda-prime] as:

$$\Lambda' = \frac{\sum_{e' \in E(T')} l(e')}{\sum_{e \in E(T)} l(e)},$$

where $l(e)$ is the length of edge e in T and $l(e')$ is the length of edge e' in T'. $E(T)$ defines the set of edges on the MLE tree T under the alternative (without constraint on the tree topology) and $E(T')$ is the set of edges on the MLE tree T' under the null hypothesis (with constraint on the tree topology).

Note that the ratio $\Lambda' \geq 0$, and, Λ' can be greater than one. Also note that ratios close to one constitute evidence favoring H_0 (congruence with the constraint tree), and higher ratios constitute greater evidence for H_1.

Note that the ratio test statistic Λ' is standardized: i.e., like the Z statistic, it does not depend on the scale. In addition, we compute each Λ' independently from each alignment, and since the Λ' values are standardized, we can compare them even though each gene tree is reconstructed independently from each alignment. This is a significant difference from the Shimodaira-Hasegawa (SH) [36] and approximately unbiased (AU) tests [37]. SH and AU test whether the given trees are congruent to each other by comparing likelihood functions in the same given data set. However, our CURatio test compares test statistics that are independent of scale, therefore lacking the constraints of SH and AU.

The CURatio method operates in the following manner: Given a set of alignments $\{A_1, \ldots, A_g\}$ for g genes on n individuals and a tree topology τ for the constraint tree, we reconstruct the MLE gene trees from each alignment both constrained or unconstrained by τ. Next, we calculate the ratio of total branch length of the constrained and the unconstrained tree. The pseudocode in Algorithm 1 summarizes this process.

```
Algorithm 1: CURatio
  Input: A set of alignments $\{A_1, \ldots, A_g\}$ for $g$ genes
         on $n$ individuals (species) and a tree topology $\tau$
         for the constraint tree.
  Output: A sequence of ratios $(r_1, ..., r_g)$.
    1) For $i = 1, \ldots, g$, do
          a) Reconstruct the MLE gene tree $T_i$ from an
             alignment $A_i$ for $i = 1, \ldots, g$ without any
             constraint.
          b) Reconstruct the MLE gene tree $T'_i$ from an
             alignment $A_i$ for $i = 1, \ldots, g$ with the
             constraint tree topology $\tau$.
          c) Compute the total branch length $b_i$ of $T_i$.
          d) Compute the total branch length $b'_i$ of $T'_i$.
          e) Compute $r_i = b'_i / b_i$.
    2) Return the ratios $(r_1, ..., r_g)$.
```

Once we have all the ratios, we set the significance level as $1 - P$ where P is the 95th percentile (or higher) of the collection $\{r_1, ..., r_g\}$ as the default. Finally, we select the genes with ratios which are greater than P.

The hypothesis test is performed as follows: We compute the test statistics r_i from the observed data (alignments) A_i. Then we estimate the distribution of r_i under the null hypothesis (if we know the asymptotic distribution of r_i then we use it, but this is still an approximation). If A_i yields r_i in the rejection region, for example above the 95th percentile of the estimated distribution, then A_i is considered an outlier. The performance of this test is shown in Figure 1 for varing P from 0 to 1.

2.2 Empirical data set

Genome sequences determined for one isolate each of 12 species in the fungal family Clavicipitaceae were annotated with MAKER version 2.28 [38]. The annotation of *Epichloë festucae* Fl1 (GenBank BioProject PRJNA51625) was manually refined based on cDNA and RNA-seq datasets, and the resulting gene models were included as evidence in the MAKER annotations of the other genomes. The other genomes in this study were from *Aciculosporium take* (PRJNA67241), *Atkinsonella texensis* B6155 (PRJNA274998), *Balansia obtecta* B249 (PRJNA221345), *Claviceps purpurea* 20.1 (PRJNA76493), *Epichloë amarillans* E4668 (PRJNA222148), *Epichloë inebrians* E818 (PRJNA174039), *Epichloë glyceriae* E277 (PRJNA67247), *Epichloë mollis* AL9923 (PRJNA215230), *Epichloë typhina* subsp. *poae* E5819 (PRJNA68441), *Metarhizium robertsii* ARSEF 23 (PRJNA38717) and *Periglandula ipomoeae* P4806 (PRJNA67303).

Gene models for the 12 genomes were subjected to OrthoMCL version 2.0.2 [39] to classify ortholog groups, as described in the OrthoMCL algorithm document (https://docs.google.com/document/d/1RB-SqCjBmcpNq-YbOYdFxotHGuU7RK_wqxqDAMjyP_w/pub). Because OrthoMCL-derived groups may contain paralogs as well as orthologs [39], we used the refiner COCO-CL [40] to improve the inference of ortholog groups. To enhance the reliability of the refinement process and the quality of generated alignments, we used a modified version of COCO-CL described in Protocol S2 of [41].

For each gene, the nucleotide sequence was identified from the start codon to the stop codon, including introns; all such gene sequences for each ortholog group were aligned by MAFFT version 6.864b [42], [43]. Finally, the ortholog groups were filtered to exclude those that had more than one representative from any genome, those that had fewer than five orthologs, and those for which the alignment had fewer than 50% non-gap characters for every gene sequence. The latter condition was imposed to filter out groups that included misannotated genes, although it also removed some ortholog groups that included pseudogenes. In total, 4266 out of 16995 ortholog groups passed the filters.

Phylogenies were determined by maximum likelihood estimation (MLE) implemented in the `R` package `ape` [35] under a Jukes-Cantor model. Those 3408 ortholog groups that had a representative from each of the 12 genomes were analyzed in a batch by CONSENSE in the `PHYLIP` version 3.2 package [44], and a 65% consensus tree was chosen as the constraint tree; this corresponded to a 70% consensus of the trees inferred under a GTR+Gamma model.

3 Results

3.1 Simulations

We conducted simulations to test `CURatio` on gene trees generated under the coalescent process,

$$\textit{Depth} = \textit{Population Size} \times C \quad (1)$$

where *Depth* is the depth of the species tree, *Population Size* is the effective population size (N_e) and C is a parameter, which we varied from 0.6 to 6.0 as in [45], [46].

For each value of C, we generated 2000 species trees with 10 leaves each under the Yule process, and calculated the Robinson-Foulds (RF) distance [47] for each pair of trees using the `R` package `phangorn` [48]. Then, for each RF distance 2, 4, 6, 8, 10, 12 and 14, we randomly selected ten pairs of species trees. For each selected pair we called one species tree “TreeOne” and the other “TreeTwo”.

From each species tree, we generated 1000 gene trees with 10 leaves under the coalescent model using the software `Mesquite` [49], with the fixed “Population Size” equal to 10,000 and the depth of the species tree determined by the parameter C (Equation 1). For each pair of species trees, we called the set of gene trees generated from TreeOne “GeneOne”, and the set generated from TreeTwo “GeneTwo”.

We then simulated DNA alignments based on these gene trees using `PAML` [50] under the Jukes-Cantor (JC) model, which is a special case of the GTR model with equal mutation rates $\frac{\mu}{4}$, where μ is the overall substitution rate.

Algorithm 2: Simulating Data Sets Process

```
for each C (from 0.6 to 6.0) do
    generate 2000 species trees randomly and calculate
      pairwise RF distance;
    for each RF distance (2, 4, 6, 8, 10, 12, 14) do
        randomly pick 10 pairs of species trees;
        for each pair of species trees($S_1$,$S_2$) do
            generate 1000 gene trees $G_1$ from $S_1$;
            generate 1000 gene trees $G_2$ from $S_2$;
            generate 1000 alignments $A_1$ from each tree
              in $G_1$;
            generate 1000 alignments $A_2$ from each tree
              in $G_2$;
        end
    end
end
```

The first simulation produced ROC curves for comparing `CURatio` with `KDETREES` [51]. `KDETREES` is a non-parametric method to estimate the distribution of trees and identify potential outlier gene trees which are probably not generated by this distribution; `CURatio`, on the other hand, is a parametric method. Note that `CURatio` does not fit a chi-squared distribution because it is not a traditional likelihood ratio test. Instead, potential outlier genes can be identified by those giving a value of r in a high percentile of the distribution of r values of all the genes in the genome for which phylogenies were determined. To draw ROC curve we vary the value r from 0 to 1 and we plot the true positive (as x-axis) and the false positive (as y-axis). We used the set of alignments GeneOne and their corresponding trees as the non-outlier data set, and we used the set of alignments GeneTwo and their corresponding trees as the outlier data set. The constraint tree was the species tree corresponding to GeneOne. The process is summarized in Algorithm 3.

We randomly selected a data set for each C value from our simulations regardless of RF distance. As shown in Figure 1, `CURatio` performed as well or better than `KDETREES` for C values up to 2. `KDETREES` performed better than `CURatio` at $C = 4$. For $C = 6$ the ROC curves for both methods passed close to the (0,1) point.

Our second simulation procedure is outlined in Algorithm 4. For each pair of species trees and the associated gene trees, we applied `CURatio` (Algorithm 1) four times, to obtain four sets of ratios: once on the set of alignments GeneOne against the corresponding species tree TreeOne; once on GeneOne against the other species tree; and likewise for the GeneTwo alignments. Then we used

Algorithm 3: Summary of the simulation comparing `CURatio` and `KDETREES`. For our simulation, $m = 100$, $n = 1$

```
Input: A set of alignments $\{A_1, \ldots, A_g\}$ for $g$ genes
    and their corresponding trees as the non-outlier
    data set. A set of alignments $\{B_1, \ldots, B_r\}$ for $r$
    genes and their corresponding trees as the
    outlier data set. A species tree, $S$,
    corresponding to the non-outlier trees,
Output: Average number of true and false outlier
    identifications for each method
for each C (from 1.0 to 6.0) do
    Randomly sample $m$ alignments and their
      corresponding trees from the non-outlier data set;
    Randomly sample $n$ alignments and their
      corresponding trees from the outlier data set;
    Detect outliers with both CURatio and
      KDETREES;
    Tally true and false outlier identifications for both
      methods;
end
```

`R` to calculate Tukey's five number summary (minimum, lower-hinge, median, upper-hinge, maximum) of each of the four sets of ratios. We were particularly interested in the trend of the medians of GeneOne with TreeTwo, and GeneTwo with TreeOne, with increasing C and different RF distances between the species trees. Significant differences were apparent at RF = 4 and high C values; at RF = 6 or higher, significant differences were also apparent for C values of 2 or less (Figure 2).

Algorithm 4: LOESS Plot

```
Input: Two sets of alignments, $A^1$ and $A^2$, and their
    corresponding species trees, $S_1$ and $S_2$.
Output: The trend of medians
for each RF distance (2, 4, 6, 8, 10, 12, 14) do
    for each combination of sets of alignments and species
      tree, $(A^1,S_1)(A^1,S_2)(A^2,S_1)(A^2,S_2)$ do
        Apply Algorithm 1;
        Calculate the medians.
    end
    Apply "LOESS" from R to fit a smooth curve.
end
```

For visualization, we applied "LOESS" from `R` on these medians, fitting a smooth curve through the points in Figure 2, where we can observe that both of the two ratios are greater than one. Nevertheless, when the species tree used to simulate the gene tree was also the corresponding constraint tree, the ratios were closer to one than when the other species tree with substantially different topology was the constraint tree.

When using the species tree as the corresponding constraint tree, larger values of C resulted in ratios approaching one. This was as expected because, as C gets larger, the species tree becomes taller and narrower relative to population size, so gene trees tend to follow the species tree topology more closely. Also as expected, such behavior was not apparent when the gene trees differed from the species trees, particularly at RF distances of six or greater.

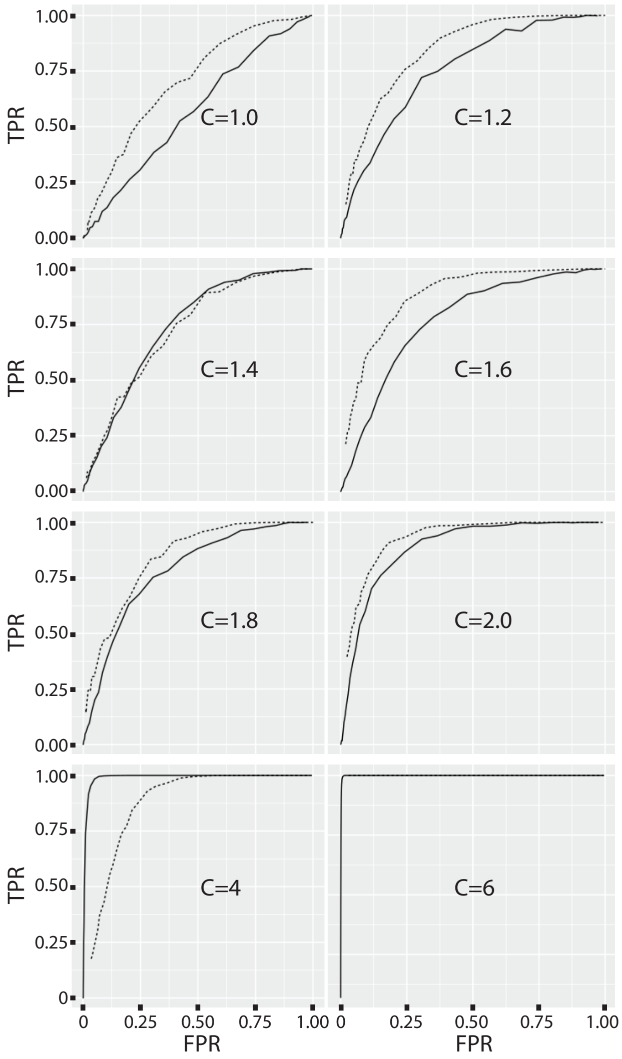


Fig. 1: ROC curves comparing results of `CURatio` (dashed line) and `KDETREES` (solid line) as the C value is changed. TPR stands for true positive rate and FPR stands for false positive rate.

An important feature of `CURatio` is that it is applicable to datasets that include ortholog groups where some taxa lack the gene, as well as ortholog groups with paralogs. For paralogs, in-paralogs arise from gene duplications on terminal branches and should not cause deviation from the constraint tree, whereas out-paralogs arise from gene duplications on internal branches and consequently differ from the constraint tree (Figure 3). To account for paralogs, CURatio modifies the constraint tree as if all paralogs in an alignment are in-paralogs; for any genome with two or more sequences in an alignment the corresponding taxon is represented as the corresponding number of sister taxa in a monophyletic clade. For the examples in Figure 4, where the original constraint tree is ((D,C),(E,A),B);, the paralogs in genomes A and E are treated as in-paralogs to give the constraint tree ((D,C),((E′,E″),(A′,A″)),B);. In the simulation, out-paralogs resulted in ratios significantly greater than one

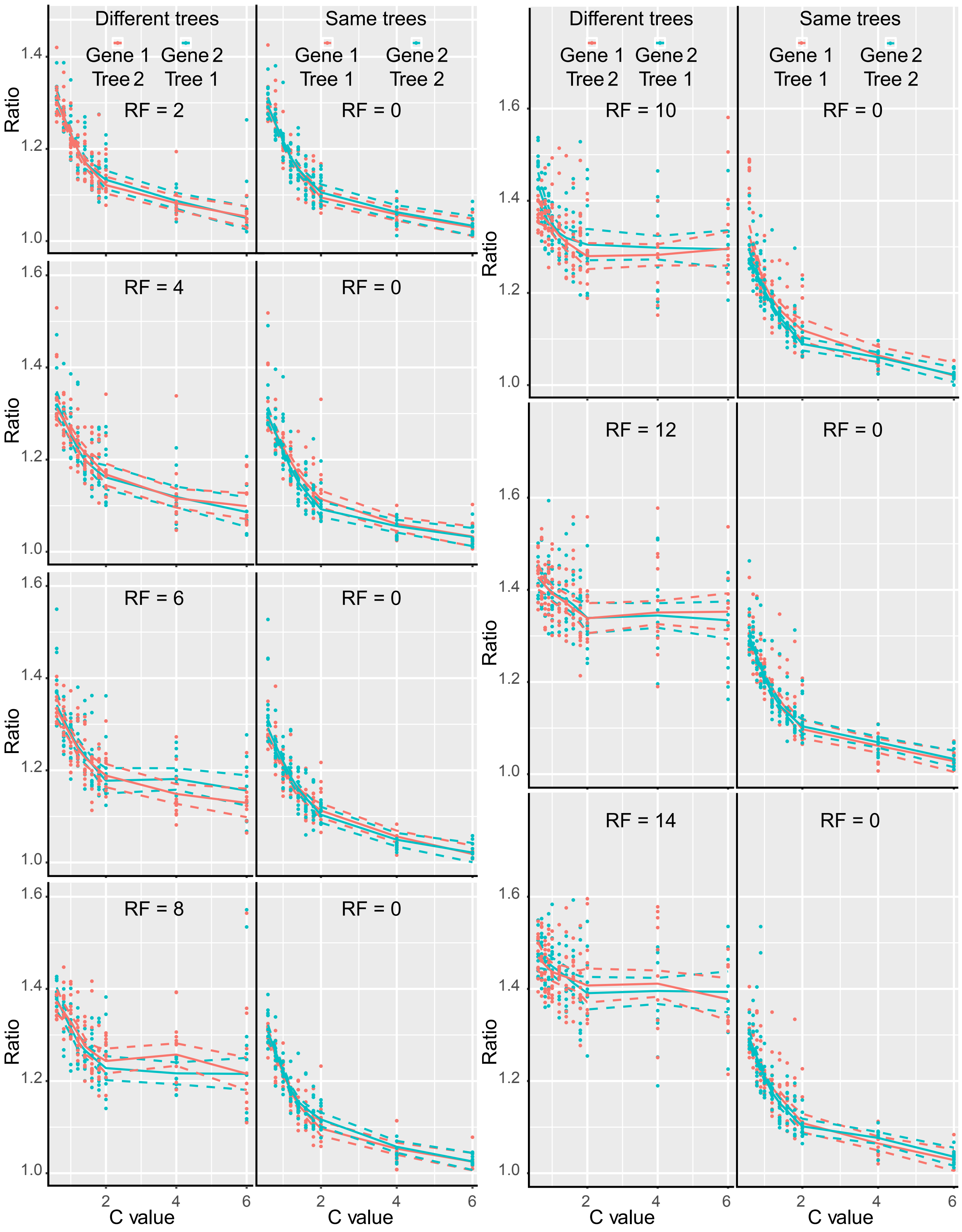


Fig. 2: LOESS on medians of four sets of ratios fitting a smooth curve through the points. Each set contains 10 points for each C value. The area between the two dashed lines is the 95% confidence interval. When using a constraint tree with a different topology from the species tree ("Different trees" columns), the ratio tends to be greater than one. Significant differences were apparent at RF = 4 for high values of C; at RF $\geq$ 6, significant differences were also apparent for smaller C values ($C \leq 2$)

(Figure 4).

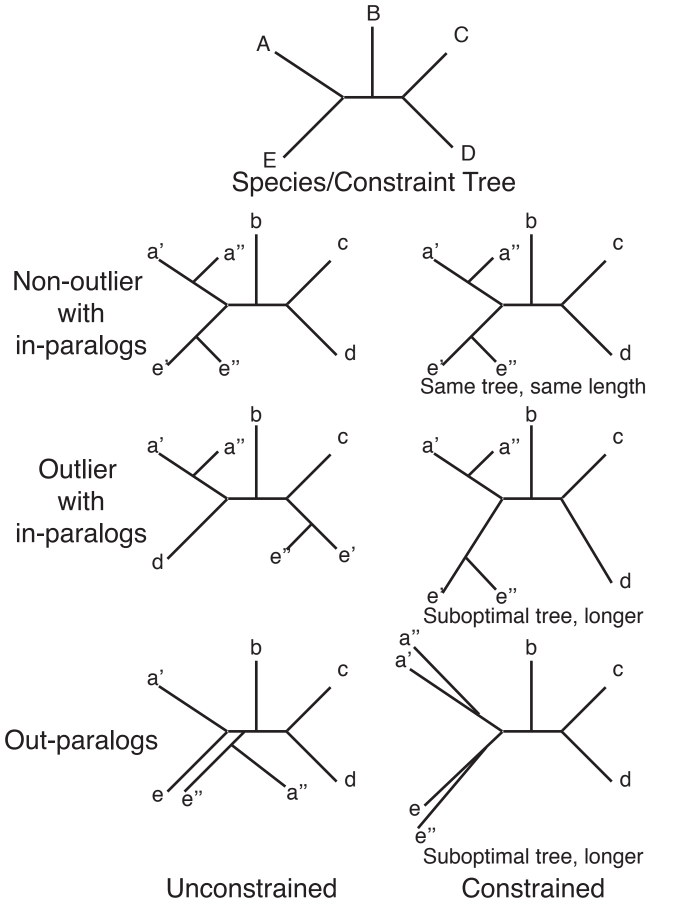


Fig. 3: Examples of constrained and unconstrained tree configurations for ortholog groups with either in-paralogs or out-paralogs. In-paralogs arise from gene duplication on terminal branches, whereas out-paralogs arise from gene duplication in common ancestors of two or more species. For non-outlier trees the ratios of constrained to unconstrained tree lengths should be close to one, whereas for ortholog groups with outlier phylogenies and ortholog groups with out-paralogs the ratios should be greater than one.

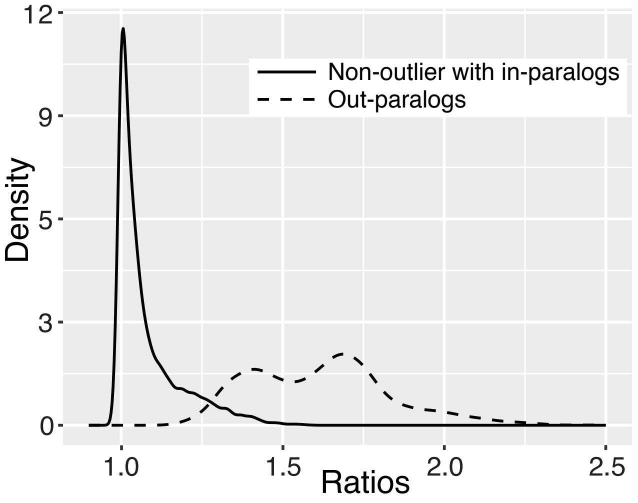


Fig. 4: Density plots of ratios of constrained to unconstrained tree lengths for non-outlier ortholog groups with in-paralogs and ortholog groups with out-paralogs as disgrammed in Figure 3. The p-value of a two-sample t-test was 2.2×10^{-16}, indicating a statistically significant difference between non-outliers with in-paralog and out-paralog groups.

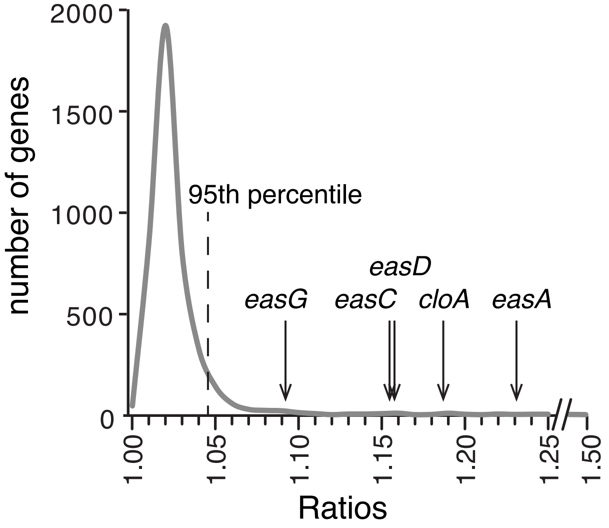


Fig. 5: A histogram of log ratios of constrained tree length to unconstrained tree length based on the empirical data set of 4266 ortholog groups from 12 annotated fungal genomes. The lowest observed ratio was approximately 0.994. The ratios obtained for ergot alkaloid biosynthesis genes are indicated by arrows.

3.2 Analysis of an empirical data set

CURatio was applied to ortholog groups from a set of 12 genomes of fungi in the family Clavicipitaceae; a histogram of ratios of constrained tree length to unconstrained tree length is presented in Figure 5. Although there was a negative trend between the ratios and the numbers of genomes containing orthologs in an ortholog group, the correlation coefficient was -0.433. Thus, there was not a strong general relationship between whether a gene was a core gene (present in all 12 genomes) or accessory gene (present in fewer than all genomes) and the ratio values for its conformity to the species tree.

It has been noted previously that, in the Clavicipitaceae, phylogenies of ergot alkaloid biosynthesis (*EAS*) genes fail to match phylogenies of core housekeeping genes commonly used to infer species relationships [52], [53]. Ortholog groups for five *EAS* genes passed the filters (see Section 2.2) and were included in our analysis. All five *EAS* genes gave ratios exceeding 1.09, and were therefore considered significant outliers as expected (Figure 5). Figure 6 compares ratios for nine core housekeeping genes and a mating type gene (*mtAC*) with those of the five *EAS* genes, *easG*, *easC*, *easD*, *cloA* and *easA*. If, instead of ratios, genes were ordered by RF values, the difference between *EAS* genes and housekeeping genes was much less apparent. With RF $= 5$, *easG* was in the 52nd percentile, and with RF $= 9$, *easC*, *easD*, *cloA* and *easA* were in the 95th percentile. RF values for housekeeping genes ranged from 2 to 9, with *tefA*, *rpbB* and *actG* having $RF = 5$ (52nd percentile), *tubP* RF $= 7$ (80th percentile), and *gapD* RF $= 9$ (95th percentile). In contrast, ratio values for the ten housekeeping genes in Figure 6 ranged from the 4th to the 73rd percentile, whereas ratio values for the *EAS* genes were all in the 99th percentile.

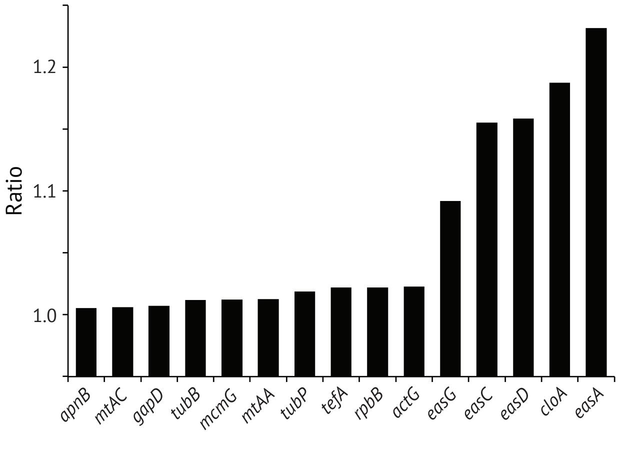


Fig. 6: Ratios of constrained to unconstrained tree lengths for nine core housekeeping genes and a mating type gene (*mtAC*) with those of the five ergot alkaloid biosynthesis genes, *easG*, *easC*, *easD*, *cloA* and *easA*.

4 Discussion

Our objective was to develop a simple statistical approach to identify genes with evolutionary histories that significantly deviate from their corresponding species phylogeny, and particularly an approach that can accommodate genes that are missing or duplicated (paralogs) in some genomes. We have proposed a novel statistical method, `CURatio`, to detect outlying gene trees from a large set of gene trees, for example obtained by whole genome analyses. For each set of orthologous genes we calculate the length of the MLE tree constrained to the postulated species tree, divided by the length of the unconstrained MLE tree to give the ratio statistic. We approximate the distribution of the ratios from the observations, such as the entire set of orthologous gene alignments from the gene annotations generated from multiple whole-genome sequences. In our simulations without outliers, as well as in our analysis of empirical data, the distributions appear close to a gamma function. Therefore, potential outliers can be identified in the upper tail of that distribution, such as the 95th or 99th percentile. An obvious phylogenomic use for this method is to explore relative deviations from the more common phylogenies, such as different ratio percentiles, to address questions such as whether some classes of genes tend to deviate more than others. Importantly, the `CURatio` method can be applied to gene sets in which some genes are lacking in some of the taxa, making it possible to compare such accessory genes with the species tree.

We applied the `CURatio` method to simulated data, with gene trees derived from the coalescent model, based on species trees differing by RF distances of 2 through 14, assuming $N_e = 10,000$ and various C values for population depth $= N_e \times C$. With these parameters, average ratios were significantly different for the same versus different species trees for $C \geq 0.6$ at moderate to high RF distances.

A set of genomes from Clavicipitaceae was chosen for an empirical test of `CURatio` because previous investigations of species and alkaloid gene phylogenies indicated different evolutionary histories [53]. Of the 12 genomes included, *EAS* genes were present in 10 of the genomes. The maximum number of *EAS* genes was 14, and nine *EAS* genes were shared among all 10 genomes. Despite sharing a similar topology, *easG* had a much lower RF ($= 5$) than the other *EAS* genes (RF $= 9$), simply because *easG* was not represented in all of the genomes that contained the other *EAS* genes. Nevertheless, the *EAS* genes all had ratios in the 99th percentile. Furthermore, the 10 housekeeping and mating type genes had a wide range of RF values (2 to 9), but all had ratios very close to 1.00. Given the overlap in RF values, *EAS* genes were not discoverable as outliers based on RF. In fact, RF did not correlate significantly with ratios ($R^2 = 0.0483$). The obvious reason is that RF is a purely topological measure, and some genes that gave high RF differed from the constraint tree only in short branches. Constraining such trees only slightly lengthened them.

For various reasons, only five of the 14 *EAS* genes passed the filter to be included in the analysis (see Section 2.2). Of the excluded genes, three were present in fewer than five of the genomes, *dmaW* was duplicated in *C. purpurea* (in this run we excluded duplicated genes), the closely linked *easF* and *easE* genes were sometimes misannotated as a single gene, and the *lpsA*, *lpsB* and *lpsC* genes were not separated from other nonribosomal peptide synthetase genes by the `OrthoMCL`/`COCO-CL` pipeline. The stringency of the filter was deemed necessary to minimize cases of outliers originating from misannotations or incorrect inferences of orthology, but in future, consideration can be given to refining orthology searches and subsequent filters to capture a greater proportion of shared genes for the `CURatio` test. It seems likely that accessory genes were disproportionately excluded, so more inclusive representation may well affect the observed distribution of ratios. Additionally, although not included in our empirical analysis, the implementation of `CURatio` allows for inclusion of genes for which more than one ortholog (up to a user-set maximum) may occur in a taxon or genome.

Because maximum likelihood estimation is an NP-hard problem, `CURatio` is likewise NP-hard: With the best known exact MLE algorithms, it requires time exponential in the size of the largest input alignment. However, each alignment's MLE trees are calculated independently of the other alignments, so the algorithm has only linear complexity in the number of alignments. Calculating a single ratio from an alignment in our empirical data set with 12 taxa and a total size of 40 kilobytes (approximately 3300 bases in length) required 3.1s. Calculating ratios for the entire empirical data set, with 4266 alignments, required 21 minutes. These timings were performed on a computer with a 3.40GHz Intel Core i7-6700 processor (8 cores), 16 GB of memory, and the Ubuntu 17.10 64-bit operating system.

In simulations in which the constraint tree differed from the species tree by RF distances of 2–14, we observed that tree-length ratios leveled out at 1.1–1.3 for $C = 2$ or greater. It would be of interest to derive an explicit formula for the expected ratio under a given model and number of leaves. Also of interest is the possibility of estimating population depths based on the distributions of ratios from empirical data sets.

In this paper, we employed the JC evolutionary model, a specific case of Markov models, which are popular in the area of molecular evolution. Markov models have the "no memory" feature that the transition probabilities depend only upon the current state, which makes it natural to assume that the nucleotide sites in the DNA sequence evolved independently of each other. However, such an assumption is often inappropriate in co-evolution [54]. We will discuss this situation and develop alternative models in future work.

Acknowledgments

We thank Walter Hollin for laboratory support, Jennifer S. Webb for genome sequencing, Jolanta Jaromczyk for bioinformatic support, and Jerzy W. Jaromczyk for helpful algorithmic discussions. This work was supported by National Institute of Food and Agriculture grant 2012-67013-19384, National Institutes of Health grants R01GM086888 and 2 P20 RR-16481, and National Science Foundation grant EPS-0814194. Genome sequence analysis was conducted in the University of Kentucky Advanced Genetic Technologies Center.

References

[1] W. P. Maddison, "Gene trees in species trees," *Systematic Biology*, vol. 46, no. 3, pp. 523–536, 1997.

[2] D. H. Huson, T. Klopper, P. J. Lockhart, and M. A. Steel, *Reconstruction of reticulate networks from gene trees*. Research in Computational Molecular Biology, Proceedings, Berlin: Springer-Verlag Berlin, 2005.

[3] D. W. Weisrock, H. B. Shaffer, B. L. Storz, S. R. Storz, S. R. Storz, and S. R. Voss, "Multiple nuclear gene sequences identify phylogenetic species boundaries in the rapidly radiating clade of mexican ambystomatid salamanders," *Molecular Ecology*, vol. 15, pp. 2489–2503, 2006.

[4] J. W. Taylor, D. J. Jacobson, S. Kroken, T. Kasuga, D. M. Geiser, D. S. Hibbett, and M. C. Fisher, "Phylogenetic species recognition and species concepts in fungi," *Fungal Genetics and Biology*, vol. 31, pp. 21 – 32, 2000.

[5] P. Brito and S. Edwards, "Multilocus phylogeography and phylogenetics using sequence-based markers," *Genetica*, vol. 135, pp. 439–455, 2009.

[6] S. Edwards, "Is a new and general theory of molecular systematics emerging?," *Evolution*, vol. 63, pp. 1–19, 2009.

[7] P. Pamilo and M. Nei, "Relationships between gene trees and species trees," *Mol. Biol. Evol.*, vol. 5, pp. 568–583, 1988.

[8] N. Takahata, "Gene genealogy in three related populations: consistency probability between gene and population trees," *Genetics*, vol. 122, pp. 957–966, 1989.

[9] J. Bollback and J. Huelsenbeck, "Parallel genetic evolution within and between bacteriophage species of varying degrees of divergence," *Genetics*, vol. 181, no. 1, pp. 225–234, 2009.

[10] M. Carling and R. Brumfield, "Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in passerina buntings," *Genetics*, vol. 178, pp. 363–377, 2008.

[11] Y. Yu, T. Warnow, and L. Nakhleh, "Algorithms for MDC-based multi-locus phylogeny inference: Beyond rooted binary gene trees on single alleles," *J Comput Biol*, vol. 18, no. 11, pp. 1543–1559, 2011.

[12] R. Betancur, C. Li, T. Munroe, J. Ballesteros, and G. Ortí, "Addressing gene tree discordance and non-stationarity to resolve a multilocus phylogeny of the flatfishes (Teleostei: Pleuronectiformes)," *Systematic Biology*, p. doi:10.1093/sysbio/syt039, 2013.

[13] J. Heled and A. Drummond, "Bayesian inference of species trees from multilocus data," *Molecular Biology and Evolution*, vol. 27, no. 3, pp. 570–580, 2011.

[14] K. Thompson and L. Kubatko, "Using ancestral information to detect and localize quantitative trait loci in genome-wide association studies," *BMC Bioinformatics*, vol. 14, p. 200, 2013.

[15] N. Rosenberg, "The probability of topological concordance of gene trees and species trees," *Theor. Popul. Biol.*, vol. 61, pp. 225–247, 2002.

[16] N. A. Rosenberg, "The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model," *Evolution*, vol. 57, pp. 1465–1477, 2003.

[17] J. H. Degnan and L. A. Salter, "Gene tree distributions under the coalescent process," *Evolution*, vol. 59, pp. 24–37, 2005.

[18] L. Liu, L. Yu, L. Kubatko, D. K. Pearl, and S. V. Edwards, "Coalescent methods for estimating phylogenetic trees," *Molecular Phylogenetics and Evolution*, vol. 53, no. 1, pp. 320–328, 2009.

[19] L. Knowles, "Statistical phylogeography," *Annual Review of Ecology, Evolution, and Systematics*, vol. 40, pp. 593–612, 2009.

[20] Y. Yu, J. D. C. Than, and L. Nakhieh, "Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting," *Systematic Biology*, vol. 60, no. 2, pp. 138–149, 2011.

[21] Y. Tian and L. Kubatko, "Gene tree rooting methods give distributions that mimic the coalescent process," *Molecular Phylogenetics and Evolution*, vol. 70, pp. 63–69, 2014.

[22] W. P. Maddison and L. L. Knowles, "Inferring phylogeny despite incomplete lineage sorting," *Syst. Biol.*, vol. 55, pp. 21–30, 2006.

[23] B. C. Carstens and L. L. Knowles, "Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers," *Syst. Biol.*, vol. 56, pp. 400–411, 2007.

[24] S. V. Edwards, L. Liu, and D. K. Pearl, "High-resolution species trees without concatenation," *Proc. Natl. Acad. Sci.*, vol. 104, pp. 5936–5941, 2007.

[25] E. Mossel and S. Roch, "Incomplete lineage sorting: consistent phylogeny estimation from multiple loci," 2007. arXiv q-bio.PE.

[26] A. RoyChoudhury, J. Felsenstein, and E. A. Thompson, "A two-stage pruning algorithm for likelihood computation for a population tree," *Genetics*, vol. 180, pp. 1095–1105, 2008.

[27] L. Knowles, "Estimating species trees: Methods of phylogenetic analysis when there is incongruence across genes," *Systematic Biology*, vol. 58, no. 5, pp. 463–467, 2009.

[28] Z. Yang and B. Rannala, "Bayesian species delimitation using multilocus sequence data," *PNAS*, vol. 107, no. 20, pp. 9264–9269, 2009.

[29] A. Leaché and B. Rannala, "The accuracy of species tree estimation under simulation: A comparison of methods," *Systematic Biology*, vol. 60, no. 2, pp. 126–137, 2011.

[30] R. Hovmoller, L. L. Knowles, and L. S. Kubatko, "Effects of missing data on species tree estimation under the coalescent," *Molecular Phylogenetics and Evolution*, vol. 69, pp. 1057–1062, 2013.

[31] N. Takahata and M. Nei, "Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci," *Genetics*, vol. 124, pp. 967–978, 1990.

[32] D. Posada and K. Crandall, "The effect of recombination on the accuracy of phylogeny reconstruction," *Journal of Molecular Evolution*, vol. 54, pp. 396–402, 2002.

[33] M. C. Rivera, R. Jain, J. E. Moore, and J. A. Lake, "Genomic evidence for two functionally distinct gene classes," *Proc Natl Acad Sci USA*, vol. 95, no. 11, pp. 6239–6244, 1998.

[34] A. P. Martin and T. M. Burg, "Perils of paralogy: Using HSP70 genes for inferring organismal phylogenies," *Systematic Biology*, vol. 51, pp. 570–587, 2002.

[35] E. Paradis, J. Claude, and K. Strimmer, "APE: analyses of phylogenetics and evolution in R language," *Bioinformatics*, vol. 20, pp. 289–290, 2004.

[36] H. Shimodaira and M. Hasegawa, "Multiple comparisons of log-likelihoods with applcations to phylogenetic inference," *Mol Biol Evol*, vol. 16, pp. 1114 – 1116, 1999.

[37] S. H., "An approximately unbiased test of phylogenetic tree selection," *Syst Biol*, vol. 51, no. 3, pp. 492–508, 2002.

[38] B. Cantarel, I. Korf, S. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Alvarado, and M. Yandell, “MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes,” *Genome research*, vol. 18, no. 1, pp. 188–196, 2008.

[39] L. Li, C. J. Stoeckert, and D. S. Roos, “OrthoMCL: Identification of ortholog groups for eukaryotic genomes,” *Genome Res*, vol. 13, pp. 2178–2189, 2003.

[40] R. Jothi, E. Zotenko, A. Tasneem, and T. Przytycka, “COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations,” *Bioinformatics*, vol. 22, pp. 779–788, 2006. doi: 10.1093/bioinformatics/btl009.

[41] C. Schardl, C. Young, U. Hesse, S. Amyotte, K. Andreeva, P. Calie, D. Fleetwood, D. Haws, N. Moore, B. Oeser, D. Panaccione, K. Schweri, C. Voisey, M. Farman, J. Jaromczyk, B. Roe, D. O’Sullivan, B. Scott, P. Tudzynski, Z. An, E. Arnaoudova, C. Bullock, N. Charlton, L. Chen, M. Cox, R. Dinkins, S. Florea, A. Glenn, A. Gordon, U. Güldener, D. Harris, W. Hollin, J. Jaromczyk, R. Johnson, A. Khan, E. Leistner, A. Leuchtmann, C. Li, J. Liu, J. Liu, M. Liu, W. Mace, C. Machado, P. Nagabhyru, P. J, J. Schmid, K. Sugawara, U. Steiner, J. Takach, E. Tanaka, J. Webb, E. Wilson, J. Wiseman, R. Yoshida, and Z. Zeng, “Plant-symbiotic fungi as chemical engineers: multi-genome analysis of the Clavicipitaceae reveals dynamics of alkaloid loci,” *PLoS Genet*, vol. 9, no. 2, p. e1003323. doi: 10.1371/journal.pgen.1003323, 2013.

[42] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, “MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform,” *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[43] K. Katoh and H. Toh, “Recent developments in the MAFFT multiple sequence alignment program,” *Briefings in Bioinformatics*, vol. 9, no. 4, pp. 286–298, 2008.

[44] J. Felsenstein, “PHYLIP – Phylogeny inference package (Version 3.2),” *Cladistics*, vol. 5, pp. 164–166, 1989.

[45] D. Haws, P. Huggins, E. M. O’Neill, D. W. Weisrock, and R. Yoshida, “A support vector machine based test for incongruence between sets of trees in tree space,” *BMC Bioinformatics*, vol. 13, no. 210, 2012. doi:10.1186/1471-2105-13-210.

[46] R. Yoshida, K. Fukumizu, and C. Vogiatzis, “Multi loci phylogenetic analysis with gene tree clustering,” *Annals of Operations Research*, pp. https://doi.org/10.1007/s10479–017–2456–9, 2017.

[47] D. F. Robinson and L. R. Foulds, “Comparison of phylogenetic trees,” *Math Biosci*, vol. 53, pp. 131–147, 1981.

[48] K. P. Schliep, “phangorn: phylogenetic analysis in R,” *Bioinformatics*, vol. 27, no. 4, pp. 592–593, 2011.

[49] W. P. Maddison and D. R. Maddison, “Mesquite: a modular system for evolutionary analysis. version 2.75,” 2011.

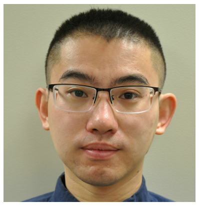
[50] Z. Yang, “PAML: A program package for phylogenetic analysis by maximum likelihood,” *CABIOS*, vol. 15, pp. 555–556, 1997.

[51] G. Weyenberg, P. Huggins, C. Schardl, D. Howe, and R. Yoshida, “KDETREES: non-parametric estimation of phylogenetic tree distributions,” *Bioinformatics*, vol. 30, no. 16, pp. 2280–2287, 2014.

[52] M. Liu, D. G. Panaccione, and C. L. Schardl, “Phylogenetic analyses reveal monophyletic origin of the ergot alkaloid gene dmaW in fungi,” *Evolutionary Bioinformatics*, vol. 5, pp. 15–30, 2009.

[53] C. A. Young, C. L. Schardl, D. G. Panaccione, S. Florea, J. E. Takach, N. D. Charlton, N. Moore, J. S. Webb, and J. Jaromczyk, “Genetics, genomics and evolution of ergot alkaloid diversity,” *Toxins (Basel)*, vol. 7, pp. 1273–1302, 2015.

[54] T. Tuller and E. Mossel, “Co-evolution is incompatible with the markov assumption in phylogenetics,” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 8, no. 6, pp. 1667–1670, 2011.



Qiwen Kang is a PhD candidate student in statistics at the University of Kentucky. He is currently working at Sanders-Brown Center on Aging as a graduate student researcher. His work is centered on phylogenomics, longitudinal data analysis and clinical trials.

Christopher Schardl is a Professor and Chair of the University of Kentucky Department of Plant Pathology. His research interests are on seed-transmitted fungal symbionts (endophytes), and particularly the biochemistry and evolutionary genomics of selective metabolites such as alkaloids, which the endophytes produce as defenses against vertebrate or invertebrate herbivores.

Neil Moore received a PhD in Computer Science from University of Kentucky in 2012. He is an Assistant Professor (Special Title Series) in the University of Kentucky Department of Computer Science, where he teaches in the First-Year Engineering Program. His research interests include phylogenomics, multihierarchical markup of text, and fine-grained access control.

Ruriko Yoshida is an Associate Professor in the Department of Operations Research at Naval Postgraduate School. Her main interests lie in applications of tools in algebra and combinatorics to problems under graphical models; Graphical Gaussian models, Bayesian networks, and discrete exponential families. She is a mathematical statistician with extensive experience in systematic biology. Her research group has a long track record of developing computational tools and statistical methods for systematic biology.