Cause of gene tree discord? Distinguishing incomplete lineage sorting and lateral gene transfer in phylogenetics

Huateng Huang Corresp., [1] , **Jeet Sukumaran** [1] , **Stephen A Smith** [1] , **L.Lacey Knowles** [1]

[1] Ecology and Evolutionary Biology, University of Michigan - Ann Arbor, United States

Corresponding Author: Huateng Huang
Email address: huatengh@umich.edu

Despite recent efforts that have produced data sets with hundreds and thousands of gene regions to resolve regions of the tree of life, recalcitrant nodes persist and disagreement among genes as well as disagreement between individual gene trees and species trees are common. There are a number of evolutionary processes that contribute to these conflicts between gene trees and species trees, including deep coalescence (lineage sorting), horizontal gene transfer or hybridization, etc. While for some of these processes, we have very powerful and sophisticated models that uses the conflict in the gene trees as information that contributes materially to correctly inferring the species tree, such as the multispecies coalescent (MSC). However, usage of these models require a priori recognition of relevant processes, which is often unknown for empirical dataset. Here we propose a new perspective to not only identify the cause of discord among gene trees, but also use it to classify loci by the underlying cause of discord to identify subsets of loci for analysis with the goal of improving phylogenetic accuracy. This approach differs fundamentally from all other criteria used for making decisions about which loci to include in a phylogenetic analysis. In particular, the choice of loci in this framework is based on identifying those that reflect descent from a common ancestor (as opposed to other processes), and thereby can minimize problems with model misspecification. We present preliminary results that demonstrate the potential of this framework in distinguishing the lateral gene transfer (LGT) from incomplete lineage sorting (ILS) process, as implemented in a new software package CLASSIPHY, while also highlighting areas for further development and testing. We discussed why such methods (i) are critical to improving phylogenetic accuracy with the increased complexity of genomic/transcriptomic datasets, and that (ii) characterizing patterns of discordance and the contribution of different processes to this discordance is itself of interest for generating hypotheses about the role of lateral gene transfer, gene duplication, and incomplete lineage sorting during the divergence of different taxa.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3489v1 | CC BY 4.0 Open Access | rec: 27 Dec 2017, publ: 27 Dec 2017

Cause of gene tree discord? CLASSIPHY, a program for distinguishing incomplete lineage sorting and lateral gene transfer in phylogenetics

Huateng Huang[1], Jeet Sukumaran[1], Stephen A. Smith[1], L. Lacey Knowles[1]

[1]*Department of Ecology and Evolutionary Biology, 1109 Geddes Ave., Museum of Zoology, University of Michigan, Ann Arbor, MI 48109-1079*

E-mail: huangh@umich.edu (H.H.); jeetsu@umich.edu (J.S.); eebsmith@umich.edu (S.A.S.) knowlesl@umich.edu (L.L.K.)

Corresponding author: Huateng Huang; *huatengh@umich.edu*

Abstract

Despite recent efforts that have produced data sets with hundreds and thousands of gene regions to resolve regions of the tree of life, recalcitrant nodes persist and disagreement among genes as well as disagreement between individual gene trees and species trees are common. There are a number of evolutionary processes that contribute to these conflicts between gene trees and species trees, including deep coalescence (lineage sorting), horizontal gene transfer or hybridization, etc. While for some of these processes, we have very powerful and sophisticated models that uses the conflict in the gene trees as information that contributes materially to correctly inferring the species tree, such as the multispecies coalescent (MSC). However, usage of these models require *a priori* recognition of relevant processes, which is often unknown for empirical dataset. Here we propose a new perspective to not only identify the cause of discord among gene trees, but also use it to classify loci by the underlying cause of discord to identify subsets of loci for analysis with the goal of improving phylogenetic accuracy. This approach differs fundamentally from all other criteria used for making decisions about which loci to include in a phylogenetic analysis. In particular, the choice of loci in this framework is based on identifying those that reflect descent from a common ancestor (as opposed to other processes), and thereby can minimize problems with model misspecification. We present preliminary results that demonstrate the potential of this framework, as implemented in a new software package CLASSIPHY, while also highlighting areas for further development and testing. In addition, we present an argument why such methods (i) are critical to improving phylogenetic accuracy with the increased complexity of genomic/transcriptomic datasets, and that (ii) characterizing patterns of discordance and the contribution of different processes to this discordance is itself of interest for generating hypotheses about the role of lateral gene transfer, gene duplication, and incomplete lineage sorting during the divergence of different taxa.

Introduction

Recent advances in sequencing technology have encouraged massive data collection efforts aimed at resolving regions of the tree of life that have eluded confident reconstruction (Rokas et al., 2003). However, the resulting phylogenomic datasets present a major challenge as phylogenetic methods for estimating the species tree while accommodating the inherent complexity of these large datasets do not exist and are not computationally feasible (Jeffroy et al., 2006). The discord among individual genes is clear with these genome-scale datasets when the phylogenetic relationships among species are examined in detail (Smith et al. 2015), which in some cases every gene in the dataset has a unique tree (e.g., Song et al., 2012).

Ignoring gene-tree discord can lead to incorrect species-tree inferences (e.g., Nosenko et al., 2013; Sharma et al., 2014; Smith et al., 2015). For example, phylogenetic estimates from concatenated datasets that ignore gene tree discord arising from incomplete lineage sorting (ILS) can be statistically inconsistent (Kubatko and Degnan, 2007). Coalescent theory makes it possible to effectively model ILS and construct a species tree conditioned on a distribution of gene trees in empirical data (Ane et al., 2007; Knowles, 2009; Knowles and Kubatko, 2011; Knowles et al., 2012; Liu et al., 2009; Mirarab et al., 2014). However, ILS may not be the primary contributor to patterns of gene tree discord in phylogenomics (e.g., Arcila et al., 2017). There are many other factors related to evolutionary history (e.g., lateral gene transfer [LGT], hybridization [H], gene duplication and loss [DL]; Maddison (1997)) and molecular evolution (e.g., noise/lack of signal in the sequences, and nonstationarity in base composition) that can contribute to gene tree discord. Yet, we lack a method that estimates phylogenetic relationships considering the many processes that contribute to gene tree discord (but see Boussau et al., 2013). As a consequence, empirical studies have difficulty in judging whether their chosen phylogenetic methods adequately model the sources of discord in the data, and what effect this model mis-specification might have on the accuracy of the phylogenetic estimates. For example, several studies have observed that slight changes to dataset assembly and/or phylogenetic reconstruction methods often generate different species trees (Betancur et al., 2014; Jarvis et al., 2014; Wickett et al., 2014; Xi et al., 2014).

Here, we argue that an alternative approach to the joint modeling of multiple processes underlying discord is to identify subsets of data with reduced heterogeneity such that the fit of the data to our models is better, and hence, the phylogenetic inference is more accurate. In particular, we ask if it is possible to identify communities of loci with similar properties using methods that are not agnostic with respect to biological processes that generate discord? We are not discounting the recent developments for estimating phylogenetic relationships while explicitly modeling specific sources of discord (e.g., gene duplication and loss, Boussau et al. (2013); hybrid origin of taxa,Meng and Kubatko (2009); networks, Solis-Lemus and Ane (2016) and Than et al. (2008)). Yet, considering that models are unlikely to accommodate all of the heterogeneity and complexity in full genomes and transcriptomes in the near future, and that the inherent heterogeneity of datasets will increase with increased taxon sampling, identifying data partitions that are most likely to reflect descent from a common ancestor (ILS as opposed to LGT and DL, for example) may be a more feasible goal. Furthermore, classifying loci according to different discord-generating processes will also provide us with a better understanding of how each process shaped the tree of life. That is, the processes underlying the discord are interesting research questions in their own right (e.g., what is the distribution of DL across the tree of life, and is it commonly associated with hypothesized ecological transitions?).

While we acknowledge this is a challenging and relatively unexplored area, we also note that the approach is not without precedent. For example, statistical procedures for identifying sets of loci with similar tree properties that might be used for phylogenetic inference, but which are agnostic with respect to the biological processes, have been proposed (e.g., Arcila et al., 2017; de Vienne et al., 2012; Fong et al., 2012; Weyenberg et al., 2014). This contrasts with our approach in which subsets of data for phylogenetic inference are identified with respect to the biological processes generating the discord. Specifically, we apply a machine learning approach, called CLASSIPHY, in which gene tree discord simulated under the actual biological processes that are known to produce discord are used to discriminate or classify genes according to cause of discord.

Given the size of datasets generated today, a full probabilistic approach is often computationally infeasible. As such, we focus on summary statistics as a means of distinguishing among sets of genes based on the processes producing discord, as in other applications (e.g., using joint sample frequency spectrum to infer multiple population history, Gutenkunst et al. (2009); topology-based D-statistic to test for introgression, Eaton and Ree (2013)). By using multiple summary statistics, in addition to being computationally tractable, CLASSIPHY is also flexible, as additional summary statistics being applied for future extensions (e.g., for an expanding into other sources of discord). Here, we present the analysis pipeline, use simulation to illustrate its application to distinguish ILS and LGT, or more specifically, discord that arises from ILS alone versus those with some LGT (i.e., trees with LGT also are subject to ILS as well) discuss factors that might affect the method's accuracy, and suggest future extensions for improvement.

Methods

CLASSIPHY Method

CLASSIPHY is a simulation-trained machine learning method (see Figure. 1 for an overview of the simulation/ analysis pipeline). Hence, the first step is simulation— simulating phylogenies under regimes corresponding to different processes that might contribute to discord. Second, we calculate summary statistics on these simulated gene trees (i.e., the training data), and then apply the Discriminant Analysis of Principal components (DAPC; Jombart et al., 2010) procedure to construct a discriminant analysis function based on extracted principal components. Lastly, application of the discriminant analysis function to the empirical set of gene trees classifies the loci with respect to the different processes that might underlie gene-tree discord, along with the posterior probabilities of each process. All the code for CLASSIPHY is available in an R package and could be accessed from https://github.com/huatengh/Classiphy.

The first gene tree simulation step can be carried out by any software as along as it can simulate and keep track of the processes of interested. The CLASSIPHY R package provides a wrapper function for SimPhy (Mallo et al., 2016), a fast and versatile program that can simulate multiple sources of gene-tree discord. In this study, we used this program to simulate the two processes—ILS and LGT. In this study, we will test whether the CLASSIPHY analysis framework can identify LGT-induced gene-tree discord from ILS-generated discord. Briefly, SimPhy simulates gene trees in three hierarchical steps: i) a species tree is simulated under a speciation/extinction model (or can be given), ii) locus trees evolve in the species tree with locus-specific LGT events, and iii) gene trees are simulated with lineage sorting process inside the locus tree (Figure 1). Hence, comparing between species tree, locus tree and gene tree, the true contribution of the two processes to the gene-tree species-tree discord is known.

The choice of summary statistics in the second step is important. The key, as with any approach that relies upon summaries of genetic data, is that they could capture some differences in the patterns generated by the processes/models being studied. Both LGT and ILS can lead to gene-tree species-tree discordance. However, LGT can generate gene-tree topologies that are more distant from the species tree and other gene trees. The distribution of gene-tree species-tree discord would also differ, because LGT does not depend on the species-tree shape as ILS (i.e., the probability of ILS is higher for internodes with short time interval between speciation events). We developed a set of summary statistics to capture these differences based on discordance among gene trees and the distribution of discordance on species tree (see Huang et al. 2017 for descriptions of the summary statistics applied here), as well as included some traditional gene-tree species-tree topological distances (e.g., Robinson–Foulds distance, Robinson and Foulds (1981). The current version of CLASSIPHY R package contains four sets of summary statistics based on tree topology. Note that the list of summary statistics can be easily expanded or adjusted by user for classification of LGT or other discord-generating processes.

It is important to note that the summary statistics are not used directly in the discriminant analysis, but rather the principal components (PCs) extracted from the summary statistics are used. Hence, these summary statistics can be correlated, and some might be relatively uninformative for certain divergent histories without biasing the results. It is the machine learning algorithm (i.e., DAPC in this case) that finds the combination of these summary statistics that can identify LGT-affected loci among gene trees with ILS-caused discord. To avoid the PCs being impacted by different scales of statistics, all summary statistics were scaled by their ranges (i.e., maximum minus the minimum). Because too many PCs will result in overfitting to the training data, whereas too few will result in lack of power (Jombart et al., 2010), we select the number of PCs in the DAPC analysis using a heuristic optimization criterion. Specifically, we first construct an array of discriminant functions using different number of PCs, and re-classify the simulated training dataset using these functions. The optimal number of PCs is the one that maximizes the percent of correct re-classification.

As a simulation-trained method, assessing CLASSIPHY's performance is straightforward. Specifically, we can keep some simulated gene trees as testing data and examine how accurately these trees are classified. It would provide information on whether the chosen summary statistics have enough power to differentiate the underlying processes. Furthermore, comparing the summary statistics between simulated and empirical data gives an indication of whether the simulations are conducted in the right parameter space (e.g., having comparable levels of gene-tree discord).

Simulation Study

We use simulation to illustrate the utility and examine the performance of the CLASSIPHY approach. Specifically, we simulated 1000 species trees with 100 taxa under a birth-death process (birth rate equal to twice of the death rate) at a fixed depth of 50*N* generation, where *N* is the effective population size. Here, we only considered the case of one individual sequenced per species, the usual sampling configuration for phylogenomic studies.

For each species tree, a rate of LGT was randomly sampled from a uniform distribution (i.e., 1e-9 to 5e-9 LGT events per generation) and 2,000 locus trees were generated. The varying LGT rate means that the portion of LGT-affected trees varies across species trees, which correspond to the fact that we usually do not know the percent of LGT-affected genes in empirical datasets. Gene trees with ILS were then simulated, where the probability of ILS differed across locus trees as a function of the branch lengths. Our analyses are based on the simulated gene trees (as opposed to estimated gene trees from simulated nucleotide datasets). As such, our results do not address the issue of lack of phylogenetic information for gene-tree estimation (see our discussion). However, by analyzing gene genealogies directly, we can focus specifically on the challenges with classification of loci by process without confounding influence from mutational variance (see Huang et al. 2010; Lanier et al. 2014). Depending on the "donor" and "receiver" lineage, LGT events may or may not cause a locus tree to differ topologically from its species tree. Therefore, only LGT events that alter tree topology were considered, and hereafter, the affected loci are referred to as LGT loci or being in the “LGT regime”. The rest of loci are referred to as “ILS loci” or being in the “ILS regime”, unless explained otherwise. In total, we simulated two million gene trees (1,000 x 2,000), and for each gene tree, we calculated an array of summary statistics based on its topology, which were constituted of 25 summary statistics when applying CLASSIPHY’s default setting on our simulated data. Since majority of the gene trees are in ILS regime for the conditions examined here, we equalized the number of loci by randomly dropping out ILS loci from the training dataset. As there were 1,000 species trees, DAPC was run 1,000 times, each time with a different species tree (along with its 2000 gene trees) as the testing data and the rest trees as the training data.

Performance Assessment

We characterized the classification ability of CLASSIPHY by investigating whether the posterior probability of LGT is a good predictor for LGT’s presence. This is evaluated by plotting the receiver operating characteristic (ROC) curve for each species tree, and calculating the area under curve (AUC) using the *pROC* R package (Robin et al., 2011). AUC is a statistic that ranges from 1 to 0.5, for perfect to zero discrimination ability, respectively. We also calculate the percentage of correct classification under two criteria: (i) the default of cutoff of greater than 0.5 as the simulation only has two regimes (i.e., LGT vs. ILS), And (ii) a cutoff that maximizes the Youden's index (i.e., sensitivity plus specificity of the classifier; Youden (1950)). We report the average AUC and proportion of correct identification across species trees.

In addition to performance evaluation, we also used the simulated data to investigate possible factors affecting the performance. This included an examination of the variation among gene trees per species tree. We calculated two RF distances, RF distance between species tree and locus tree, and that between locus tree and gene tree, which represent the true contribution of LGT and ILS to gene-tree discord, and check whether these RF distances are correlated with the posterior probabilities of LGT and ILS. We also examined the variation among species trees. More specifically, why the discrimination ability of trained DAPC model differs among species trees? Llinear regression was used to test of whether the model’s AUC correlates with the LGT rate (i.e., the percentage of true LGT gene trees) and average amount of LGT/ILS in the gene trees (i.e., average species-to-locus-tree and locus-to-gene-tree RF distance).

Results

Our simulation study shows that the posterior weight is a good predictor for the true discord-generating process (Figure 2), with an average AUC (area under curve) of 0.81 across all species trees. The posterior probability of LGT and ILS are highly correlated with the true contribution of the respective processes to gene-tree discord. Specifically, the topological differences induced by LGT (i.e., the RF distance between locus tree and species tree, D_{SL}) is positively and significantly correlated with the gene tree's posterior probability of LGT (average Pearson correlation coefficient 0.82, Fig.3a and b). That is, LGT events with large effect are more likely to be detected than those only resulting in minor topological changes (Fig. 3b). The amount of ILS present in a gene tree (i.e., the RF distance between gene tree and locus tree, D_{LG}) is negatively and significantly correlated with the posterior probability of ILS (average Pearson correlation coefficient -0.66; Fig. 3c and d). This suggests that gene trees with more ILS would have higher chance of being misidentified as LGT. Yet, as ILS gene trees in general have relatively high posterior probability of the correct regime, only a small proportion was misidentified (Fig 2b and Fig. 3d).

In addition to the variation among gene trees, there is considerable variation in terms of model performance among species trees (different AUC curves in Fig. 2a). As expected, the model's AUC is positively correlated with average D_{SL} (Fig 4a; $p<0.001$), and negatively correlated with average D_{LG} (Fig 4b; $p<0.001$). That is, with higher LGT contribution to the gene-tree discord, CLASSIPHY becomes more efficient in identifying LGT gene trees, while ILS acts as noise that reduces the accuracy. Simple linear regression also shows that the model's AUC is positively correlated with percentage of LGT gene trees in the data (Fig 4c; $p<0.001$). However, this most likely reflects the greater chance of having gene trees with high D_{SL} (and hence, high classification accuracy) as the correlation is no longer significant after controlling for D_{SL} (Fig 4d; $p=0.26$).

Discussion

Here, we describe CLASSIPHY—a simulation-based analysis framework to identify different sources of gene tree discord that has applications for current phylogenomic studies, and explored the potential of CLASSIPHY in distinguishing LGT and ILS using simulated data. Both ILS and LGT are considered important processes underlying gene-tree discordance. In particular, the awareness of ILS has increased dramatically in the last decades as more large multi-locus data were collected and more species-tree methods were developed—these methods now are almost routinely applied in phylogenetic studies (e.g., Edwards et al., 2007; Wickett et al., 2014). For LGT, the interest first came from studying prokaryotes' evolution (Brown, 2003), but more and more LGT evidences in eukaryotes are established (Keeling and Palmer, 2008). Just as ILS, multiple methods have been developed to tree reconstruction when genes have conflicting evolutionary history due to LGT events (Bansal et al., 2013; Sjostrand et al., 2014). Studies have proposed various optimizing criteria, from minimizing the total Robinson-Foulds distance of the supertree (Bansal et al., 2010) to the Subtree Prune-and-Regraft distance (Whidden et al., 2014), and robustness to LGT was compared between tree-building approaches (e.g., supertrees versus supermatrix; Lapierre et al., 2014). However, most of the methods dealing with LGT do not model coalescent process, except that a review paper by Szollosi et al. (2015) discussed a potential model by extending and combining current methods (Szollosi et al., 2015). In this study, we modelled ILS and LGT simultaneously, and it should be noted that we tested CLASSIPHY's performance in a very difficult simulated scenario—high levels ILS. We simulated species trees with 100 taxa at depth of 50N, which corresponds to two lineages per million years on average if assuming a large effective population size of one million (smaller population means even higher diversification rates), and no gene tree is identical to locus tree in our simulated dataset. The consequence is that ILS causes much more topological discord than LGT (see the difference in x-axis' scale between Fig. 4a and 4b), which makes LGT events difficult to detect. Simply ignoring ILS or only looking at single summary statistic (such as RF distance) would mistake a lot of ILS loci as LGT. In this sense, the performance of CLASSIPHY is promising that it identifies almost half of the LGT with only ~5% mis-identified ILS loci. For empirical datasets, high diversification rate is certainly possible in some radiations (e.g., cichlids; Seehausen, 2000), but most of the time, lineage diversification rate is much lower (e.g., 0.078-0.14 lineages per my for majority of the fish lineage; Rabosky et al., 2013). The simulation showed that the model's AUC increases with decreasing ILS discord (Fig. 4b), so better performance of CLASSIPHY can be expected in easier scenarios.

The performance of CLASSIPHY can also improve if more information is extracted from the divergent history itself. Imagining if we have the true divergent history at hand (not only the topology but also the time of divergent events in unit of effective population size), identifying processes other than ILS, would be quite straightforward— the probability distribution of gene tree (e.g., COAL; Degnan and Salter, 2005) can be calculated and gene trees that are too unlikely could be identified as outliers. However, in empirical studies, the goal often is to reconstruct an unknown divergent history from a heterogeneous gene-tree set with an unknown proportion of outliers. Here, when testing CLASSIPHY's performance, we set up the simulation to reflect these "unknowns": only two pieces of information are shared between testing and training data— the tree depth and birth-death model of the species tree. As a result, species trees in the training dataset differ vastly in terms of the amount of ILS (Fig. 4b), and the rate of LGT (ranging from affecting 6% of the trees to 52%; Fig. 4a). With these settings, the trained DAPC model is applicable to divergent histories from a large parameter space. Yet, the divergent history itself clearly has an impact on the model performance (Fig. 2a and Fig . 4). How to incorporate some information about the divergent history in simulating training data without risking having a model not in the right parameter space warrants further investigation. In this study, we also used gene-tree information "conservatively"— we only used topology-based summary statistics. Branch lengths would be a rich source of information, in particular, helping identify LGT events that have little effect on topology. However, they are more sensitive to mutational variance (Huang et al., 2010). Whether branch-length-based summary statistics (and which statistics) would help improve the model performance need more evaluation with estimated gene trees. One advantage of CLASSIPHY analysis framework is its flexibility—it has simulation, summary statistic calculation, and DAPC modelling as separated parts, so users can easily alter the simulation setting, modify the list of summary statistics, and test how the changes affect the model performance.

The results from CLASSIPHY would have many applications for current phylogenomic studies. As the scope of phylogenomic studies expand in terms of genomic coverage and taxa, many empirical studies suggest that multiple processes are contribute more or less to discord in a heterogeneous way throughout a phylogeny (e.g., Fontaine et al., 2015; Smith et al., 2015; Wickett et al., 2014). For example, in the recent bird genome phylogeny (Jarvis et al., 2014), lack of signal, selection due to life history evolution, and incomplete lineage sorting were all thought to play a role in shaping the phylogeny. Successfully identifying different sources of discord would allow us partition a large dataset into homogenous subsets that can be adequately modeled by existing methods (e.g., Chen et al., 2015; Xi et al., 2014). In this sense, CLASSIPHY analysis framework would be complementary to various data filtering tools that have been proposed to address the negative impact of data heterogeneity on phylogenomic estimates. Although inferred topologies in phylogenomic studies typically have high support values due to the large number of basepairs (which is a problem by itself, see Brown and Thomson (2017)), many of the new resolutions to difficult nodes on the tree of life are not accepted by researchers without reservation—there is a long list of such controversial examples from plants, fungi and animal (Shen et al., 2017). Do these new resolutions represent “whole-genome evidence”? Or reflect biases in data processing steps and tree reconstruction methods (e.g., Dell'Ampio et al., 2014; Fernández-Mazuecos et al., 2017)? Or are driven by strong signals in small number of genes or sites (which could be outliers; e.g., Shen et al., 2017; Xi et al., 2014)? Many data filtering strategies were proposed based on “interrogating” large empirical datasets, from rate of evolution to gene functional categories (e.g., Betancur et al., 2014; Doyle et al., 2015; Klopfstein et al., 2017; Romiguier et al., 2013; Salichos and Rokas, 2013). CLASSIPHY differs from these strategies based on sequence or gene-tree properties in that it employs machine learning to dissects the distribution of discord among the gene trees with respect to potential biological processes that could generate the discord-- do the pattern reflect neutral lineage sorting process (hence, a simple multi-species coalescent model is enough?) Or are there significant deviations (i.e., other processes might contribute to the conflicts among trees)? Researchers can use its result for data filtering, testing the robustness of their tree-estimation methods when mixing in varying proportions of loci affected by other processes, or as an evaluation of the prevalence of different processes to identify what should be integrated into phylogenetic models (i.e., choosing or developing appropriate species-tree estimation methods for the whole dataset).

CLASSIPHY would also help to understand more about discord-generating processes, which are interesting biological phenomena in their own right. For example, although LGT is often considered as a signature characteristic for plant genome evolution and a challenge for phylogenetic estimates (Bock, 2010), we have little information and many basic questions remain. What is the average rate of LGT? How does it vary across time and phylogeny? Does the propensity to transfer differ among different functional categories? Or chromosomal locations? With classifying tools, we can make use of large databases from projects such as the 1KP plant transcriptome project and Bird 10K project to answer these questions (Matasci et al., 2014; Zhang et al., 2015). Moreover, CLASSIPHY not only assigns loci into categories, but also outputs the posterior probability of a locus being affected by a process (Fig. 1), and we showed in simulation that this posterior probability is correlated with the true contribution of discord-generating processes (Fig. 3). Hence, users can use correlations and regressions to answer questions mentioned above (e.g., whether regressing posterior probability of LGT against gene functional categories is significant).

Conclusions

As more and more genomic-scale datasets are collected, the complexity and heterogeneity within the data becomes clear. The gap between the data we collect for phylogenetic analyses (i.e., large-scale transcriptomic and genomic data) and the methods that accommodate the inherent complexity of big data have created a tension where the accuracy of phylogenetic inferences do not necessarily increase with more data (Jeffroy et al., 2006; Philippe et al., 2011). We expect CLASSIPHY, as a tool for understanding the processes generating these complexities and conflicts, to be applicable to many phylogenomic datasets, helping in reconstructing phylogenetic histories and facilitating our understanding of genome evolution.

References

Ane, C., Larget, B., Baum, D.A., Smith, S.D., and Rokas, A. (2007). Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution *24*, 1575-1575.

Arcila, D., Ortí, G., Vari, R., Armbruster, J.W., Stiassny, M.L.J., Ko, K.D., Sabaj, M.H., Lundberg, J., Revell, L.J., and Betancur-R, R. (2017). Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nature Ecology & Evolution *1*, 0020.

Bansal, M.S., Banay, G., Harlow, T.J., Gogarten, J.P., and Shamir, R. (2013). Systematic inference of highways of horizontal gene transfer in prokaryotes. Bioinformatics *29*, 571-579.

Bansal, M.S., Burleigh, J.G., Eulenstein, O., and Fernandez-Baca, D. (2010). Robinson-Foulds supertrees. Algorithms for molecular biology : AMB *5*, 18.

Betancur, R.R., Naylor, G.J., and Orti, G. (2014). Conserved genes, sampling error, and phylogenomic inference. Syst Biol *63*, 257-262.

Bock, R. (2010). The give-and-take of DNA: horizontal gene transfer in plants. Trends Plant Sci *15*, 11-22.

Boussau, B., Szollosi, G.J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. Genome Res *23*, 323-330.

Brown, J.M., and Thomson, R.C. (2017). Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst Biol *66*, 517-530.

Brown, J.R. (2003). Ancient horizontal gene transfer. Nature reviews. Genetics *4*, 121-132.

Chen, M.Y., Liang, D., and Zhang, P. (2015). Selecting Question-Specific Genes to Reduce Incongruence in Phylogenomics: A Case Study of Jawed Vertebrate Backbone Phylogeny. Syst Biol *64*, 1104-1120.

de Vienne, D.M., Ollier, S., and Aguileta, G. (2012). Phylo-MCOA: A Fast and Efficient Method to Detect Outlier Genes and Species in Phylogenomics Using Multiple Co-inertia Analysis. Molecular Biology and Evolution *29*, 1587-1598.

Degnan, J.H., and Salter, L.A. (2005). Gene tree distributions under the coalescent process. Evolution; international journal of organic evolution *59*, 24-37.

Dell'Ampio, E., Meusemann, K., Szucsich, N.U., Peters, R.S., Meyer, B., Borner, J., Petersen, M., Aberer, A.J., Stamatakis, A., Walzl, M.G., *et al.* (2014). Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. Mol Biol Evol *31*, 239-249.

Doyle, V.P., Young, R.E., Naylor, G.J.P., and Brown, J.M. (2015). Can We Identify Genes with Increased Phylogenetic Reliability? Systematic Biology *64*, 824-837.

Eaton, D.A.R., and Ree, R.H. (2013). Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). Systematic Biology *62*, 689-706.

Edwards, S.V., Liu, L., and Pearl, D.K. (2007). High-resolution species trees without concatenation. P Natl Acad Sci USA *104*, 5936-5941.

Fernández-Mazuecos, M., Mellers, G., Vigalondo, B., Saez, L., Vargas, P., and Glover, B.J. (2017). Resolving Recent Plant Radiations: Power and Robustness of Genotyping-by-Sequencing. Syst Biol *syx062*.

Fong, J.J., Brown, J.M., Fujita, M.K., and Boussau, B. (2012). A Phylogenomic Approach to Vertebrate Phylogeny Supports a Turtle-Archosaur Affinity and a Possible Paraphyletic Lissamphibia. Plos One *7*, e48990.

Fontaine, M.C., Pease, J.B., Steele, A., Waterhouse, R.M., Neafsey, D.E., Sharakhov, I.V., Jiang, X.F., Hall, A.B., Catteruccia, F., Kakani, E., *et al.* (2015). Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science *347*, 1258524.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. Plos Genet *5*, e1000695.

Huang, H.T., He, Q.I., Kubatko, L.S., and Knowles, L.L. (2010). Sources of Error Inherent in Species-Tree Estimation: Impact of Mutational and Coalescent Effects on Accuracy and Implications for Choosing among Different Methods. Systematic Biology *59*, 573-583.

Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y., Faircloth, B.C., Nabholz, B., Howard, J.T., *et al.* (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. Science *346*, 1320-1331.

Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006). Phylogenomics: the beginning of incongruence? Trends Genet *22*, 225-231.

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. Bmc Genet *11*, 94.

Keeling, P.J., and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. Nature reviews. Genetics *9*, 605-618.

Klopfstein, S., Massingham, T., and Goldman, N. (2017). More on the Best Evolutionary Rate for Phylogenetic Analysis. Syst Biol *syx051*.

Knowles, L.L. (2009). Estimating Species Trees: Methods of Phylogenetic Analysis When There Is Incongruence across Genes. Systematic Biology *58*, 463-467.

Knowles, L.L., and Kubatko, L.S. (2011). Estimating species trees: practical and theoretical aspects (John Wiley and Sons).

Knowles, L.L., Lanier, H.C., Klimov, P.B., and He, Q. (2012). Full modeling versus summarizing gene-tree uncertainty: method choice and species-tree accuracy. Mol Phylogenet Evol *65*, 501-509.

Kubatko, L.S., and Degnan, J.H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol *56*, 17-24.

Lapierre, P., Lasek-Nesselquist, E., and Gogarten, J.P. (2014). The impact of HGT on phylogenomic reconstruction methods. Briefings in bioinformatics *15*, 79-90.

Liu, L., Yu, L.L., Pearl, D.K., and Edwards, S.V. (2009). Estimating Species Phylogenies Using Coalescence Times among Sequences. Systematic Biology *58*, 468-477.

Maddison, W.P. (1997). Gene trees in species trees. Systematic Biology *46*, 523-536.

Mallo, D., Martins, L.D., and Posada, D. (2016). SimPhy: Phylogenomic Simulation of Gene, Locus, and Species Trees. Systematic Biology *65*, 334-344.

Matasci, N., Hung, L.H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., *et al.* (2014). Data access for the 1,000 Plants (1KP) project. Gigascience *3*, 17.

Meng, C., and Kubatko, L.S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. Theor Popul Biol *75*, 35-45.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics *30*, i541-548.

Nosenko, T., Schreiber, F., Adamska, M., Adamski, M., Eitel, M., Hammel, J., Maldonado, M., Muller, W.E., Nickel, M., Schierwater, B., *et al.* (2013). Deep metazoan phylogeny: when different genes tell different stories. Mol Phylogenet Evol *67*, 223-233.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T., Manuel, M., Worheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol *9*, e1000602.

Rabosky, D.L., Santini, F., Eastman, J., Smith, S.A., Sidlauskas, B., Chang, J., and Alfaro, M.E. (2013). Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. Nature communications *4*, 1958.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., and Muller, M. (2011). pROC: an open-source package for R and S plus to analyze and compare ROC curves. Bmc Bioinformatics *12*, 77.

Robinson, D.F., and Foulds, L.R. (1981). Comparison of Phylogenetic Trees. Math Biosci *53*, 131-147.

Rokas, A., Williams, B.L., King, N., and Carroll, S.B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature *425*, 798-804.

Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N., and Douzery, E.J.P. (2013). Less Is More in Mammalian Phylogenomics: AT-Rich Genes Minimize Tree Conflicts and Unravel the Root of Placental Mammals. Molecular Biology and Evolution *30*, 2134-2144.

Salichos, L., and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. Nature *497*, 327-331.

Seehausen, O. (2000). Explosive speciation rates and unusual species richness in haplochromine cichlid fishes: Effects of sexual selection. Adv Ecol Res *31*, 237-274.

Sharma, P.P., Kaluziak, S.T., Perez-Porro, A.R., Gonzalez, V.L., Hormiga, G., Wheeler, W.C., and Giribet, G. (2014). Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal. Mol Biol Evol *31*, 2963-2984.

Shen, X.-X., Hittinger, C.T., and Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nature Ecology & Evolution *1*, 0126.

Sjostrand, J., Tofigh, A., Daubin, V., Arvestad, L., Sennblad, B., and Lagergren, J. (2014). A Bayesian method for analyzing lateral gene transfer. Syst Biol *63*, 409-420.

Smith, S.A., Moore, M.J., Brown, J.W., and Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol Biol *15*, 150.

Solis-Lemus, C., and Ane, C. (2016). Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. Plos Genet *12*, e1005896.

Song, S., Liu, L., Edwards, S.V., and Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci U S A *109*, 14942-14947.

Szollosi, G.J., Tannier, E., Daubin, V., and Boussau, B. (2015). The inference of gene trees with species trees. Syst Biol *64*, e42-62.

Than, C., Ruths, D., and Nakhleh, L. (2008). PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. Bmc Bioinformatics *9*, 322.

Weyenberg, G., Huggins, P.M., Schardl, C.L., Howe, D.K., and Yoshida, R. (2014). KDETREES: non-parametric estimation of phylogenetic tree distributions. Bioinformatics *30*, 2280-2287.

Whidden, C., Zeh, N., and Beiko, R.G. (2014). Supertrees Based on the Subtree Prune-and-Regraft Distance. Syst Biol *63*, 566-581.

Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., *et al.* (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A *111*, E4859-4868.

Xi, Z.X., Liu, L., Rest, J.S., and Davis, C.C. (2014). Coalescent versus Concatenation Methods and the Placement of Amborella as Sister to Water Lilies. Systematic Biology *63*, 919-932.

Youden, W.J. (1950). Index for Rating Diagnostic Tests. Biometrics *6*, 172-173.

Zhang, G., Rahbek, C., Graves, G.R., Lei, F., Jarvis, E.D., and Gilbert, M.T. (2015). Genomics: Bird sequencing project takes off. Nature *522*, 34.

Figure Captions:

Figure 1. Overview of the CLASSIPHY analysis pipeline. The analysis can be conceptually divided into two parts—simulation and model training (left half of the figure), and applying the model to empirical gene tree sets (right half of the figure on grey background). Gene trees are simulated in hierarchical steps, in which ILS and other processes of discord are incorporated. Underlined grey text shows some of the parameters used in this study. Summary statistics are then calculated for each simulate gene trees, constituting a large training data matrix, which was used by the DAPC method to build a discriminant function for different discord processes. This function is then applied to the summary statistics calculated from empirical gene trees. It calculates the posterior probability of each discord process (in this study, ILS and LGT) and classify trees into different processes.

Figure 2. CLASSIPHY performance across species trees. A) the ROC (Reviver Operating Characteristic) curves. In general, the closer the curve follows the left and then the top axis (i.e., closer to the upper-left corner), the more accurate is the classification; the closer the curve follows the diagonal dash line, the worse is the model performance. B) Percentage of correct classification for LGT and ILS process with different cutoffs on LGT posterior probability.

Figure 3. Variation of model performance among gene trees. For each species tree, the correlation between species-to-locus RF distance and the posterior probability of LGT was calculated for LGT gene trees. A) shows the frequency distribution of these correlations across species trees, and B) shows an example of such correlation for one of the species tree. For each species tree, the correlation between locus-to-gene RF distance and the posterior probability of ILS was calculated for ILS gene trees. C) shows the frequency distribution of these correlations across species trees, and D) shows an example of such correlation for one of the species tree. RF distances were "jittered" (adding small noise) in C) and D) to show the density of points.

Figure 4. Variation of the model performance (AUC) among species trees. A) Positive correlation between AUC and the average species-to-locus tree RF distance (D_{SL}), each point represents data from one species

tree. B) Negative correlation between AUC and the average locus-to-gene tree RF distance (D_{LG}). C) Positive correlation between AUC and the percentage of LGT trees. D) Correlation between AUC and the percentage of LGT trees after controlling for D_{SL} is not significant ($p > 0.05$).

Figure 1

Figure 1. Overview of the CLASSIPHY analysis pipeline.

The analysis can be conceptually divided into two parts—simulation and model training (left half of the figure), and applying the model to empirical gene tree sets (right half of the figure on grey background). Gene trees are simulated in hierarchical steps, in which ILS and other processes of discord are incorporated. Underlined grey text shows some of the parameters used in this study. Summary statistics are then calculated for each simulate gene trees, constituting a large training data matrix, which was used by the DAPC method to build a discriminant function for different discord processes. This function is then applied to the summary statistics calculated from empirical gene trees. It calculates the posterior probability of each discord process (in this study, ILS and LGT) and classify trees into different processes.

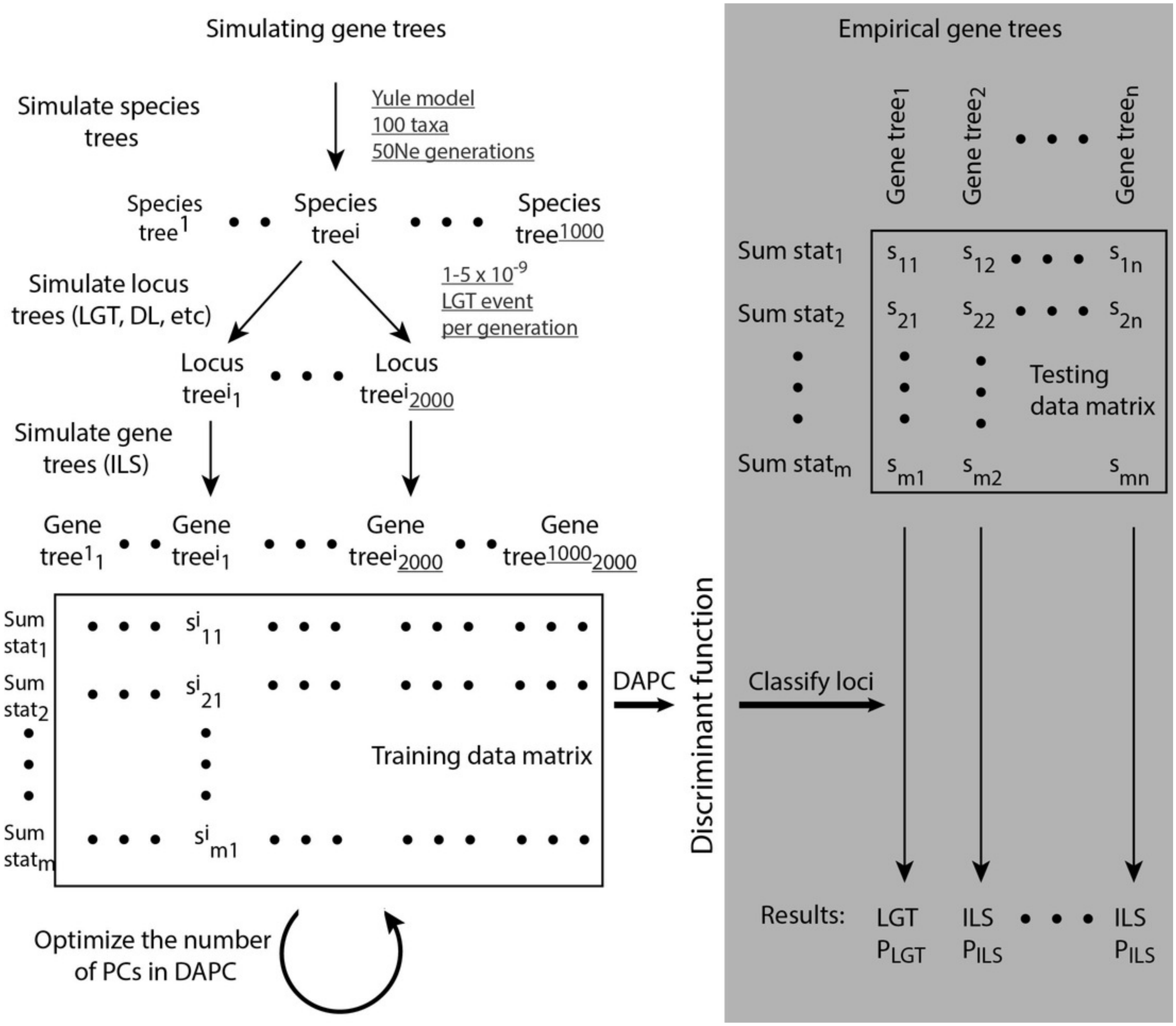
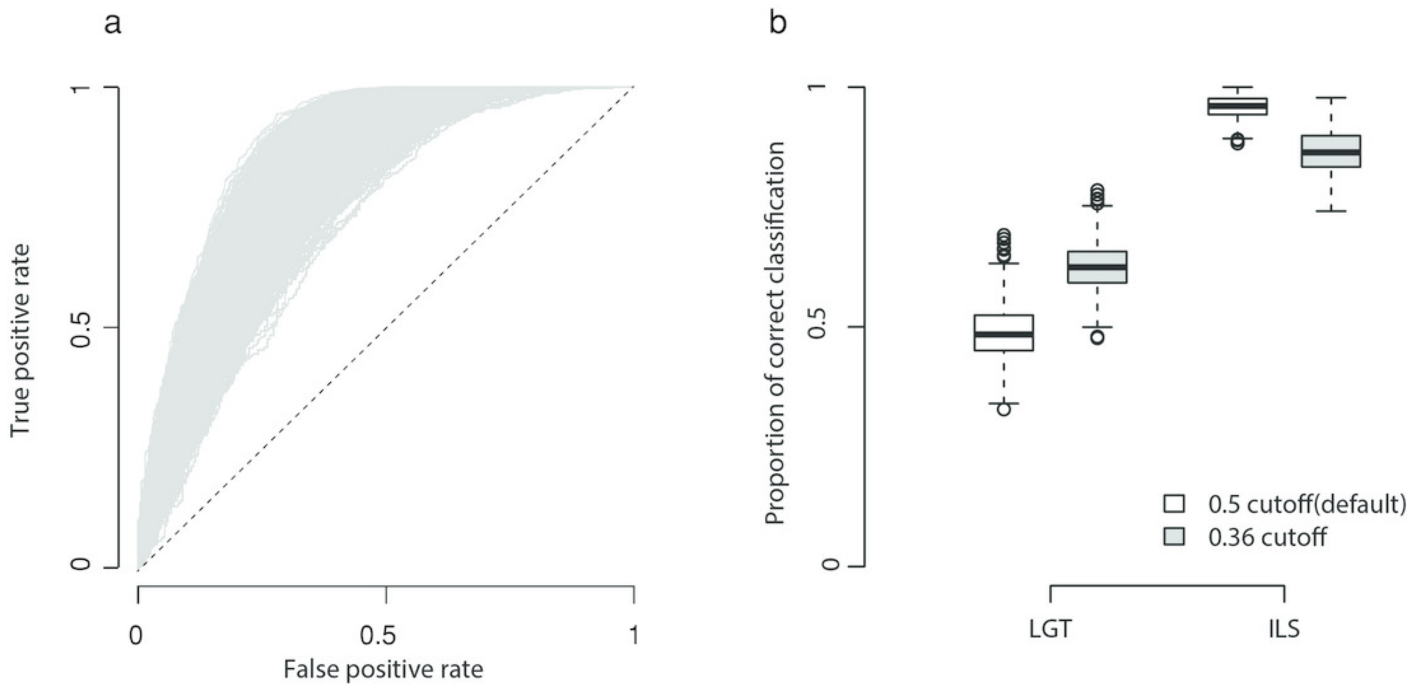

Simulating gene trees
Simulate species trees
Yule model
100 taxa
50Ne generations
Species tree1
Species treei
Species tree1000
Simulate locus trees (LGT, DL, etc)
1-5 x 10-9
LGT event
per generation
Locus treei1
Locus treei2000
Simulate gene trees (ILS)
Gene tree11
Gene treei1
Gene treei2000
Gene tree10002000
Sum stat1
Sum stat2
Sum statm
si11
si21
sim1
Training data matrix
DAPC
Discriminant function
Optimize the number of PCs in DAPC
Empirical gene trees
Gene tree1
Gene tree2
Gene treen
Sum stat1
Sum stat2
Sum statm
s11
s12
s1n
s21
s22
s2n
sm1
sm2
smn
Testing data matrix
Classify loci
Results:
LGT
PLGT
ILS
PILS
ILS
PILS

Figure 2

Figure 2. CLASSIPHY performance across species trees.

A) the ROC (Reviver Operating Characteristic) curves. In general, the closer the curve follows the left and then the top axis (i.e., closer to the upper-left corner), the more accurate is the classification; the closer the curve follows the diagonal dash line, the worse is the model performance. B) Percentage of correct classification for LGT and ILS process with different cutoffs on LGT posterior probability.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3489v1 | CC BY 4.0 Open Access | rec: 27 Dec 2017, publ: 27 Dec 2017

Figure 3

Figure 3. Variation of model performance among gene trees.

For each species tree, the correlation between species-to-locus RF distance and the posterior probability of LGT was calculated for LGT gene trees. A) shows the frequency distribution of these correlations across species trees, and B) shows an example of such correlation for one of the species tree. For each species tree, the correlation between locus-to-gene RF distance and the posterior probability of ILS was calculated for ILS gene trees. C) shows the frequency distribution of these correlations across species trees, and D) shows an example of such correlation for one of the species tree. RF distances were “jittered” (adding small noise) in C) and D) to show the density of points.

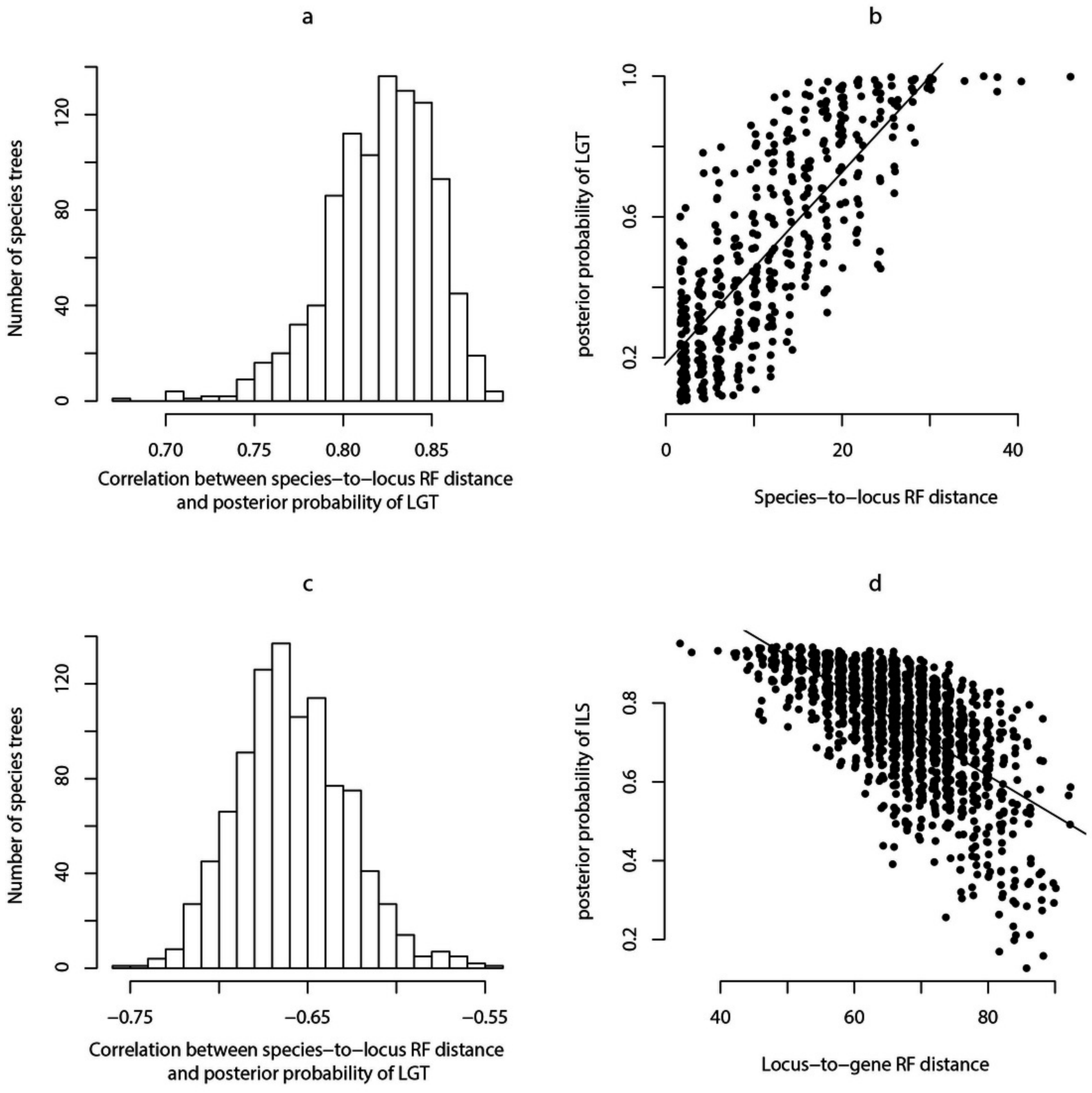
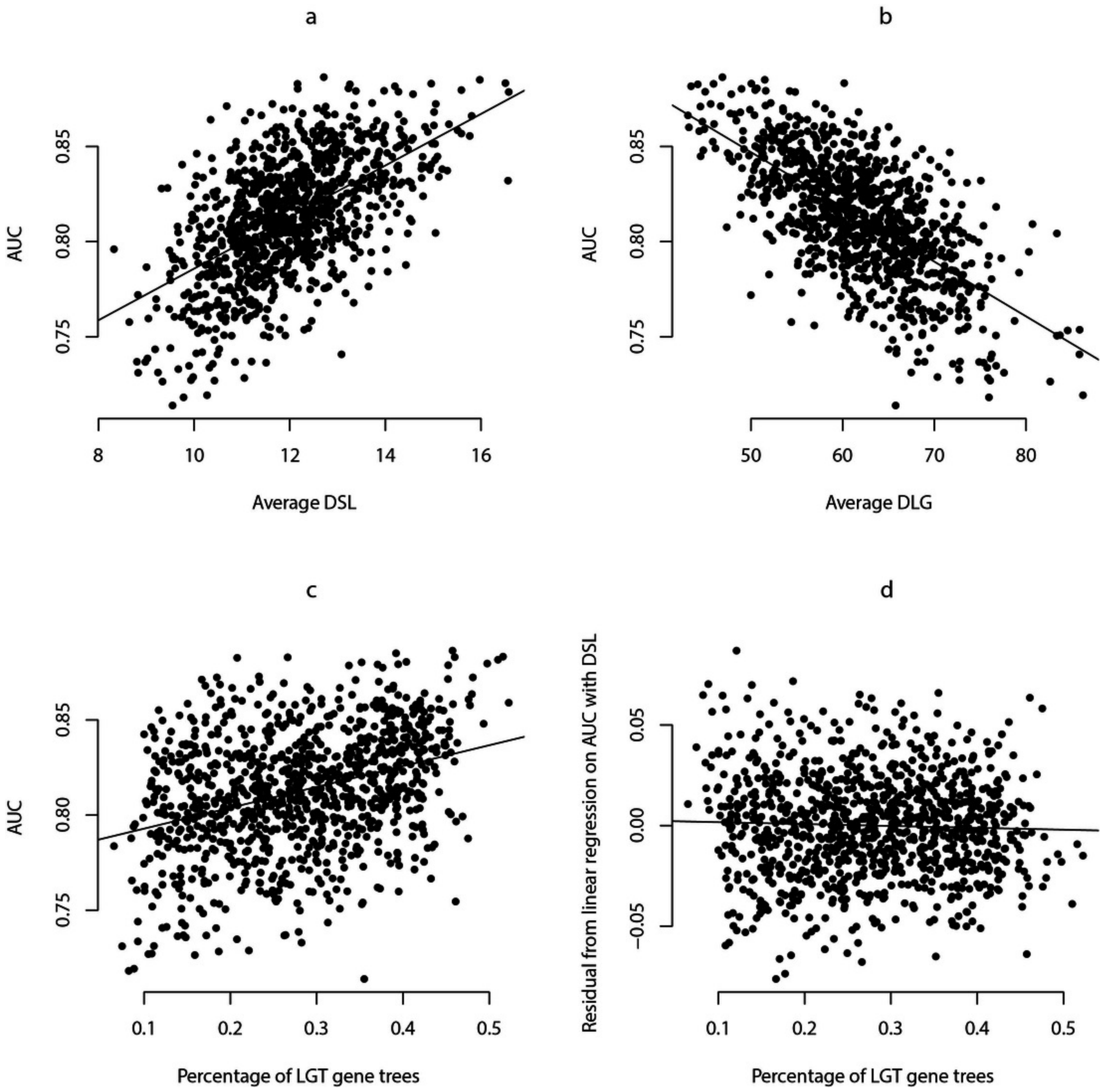

a
Number of species trees
0
40
80
120
0.70
0.75
0.80
0.85
Correlation between species–to–locus RF distance
and posterior probability of LGT
b
posterior probability of LGT
0.2
0.6
1.0
0
20
40
Species–to–locus RF distance
c
Number of species trees
0
40
80
120
−0.75
−0.65
−0.55
Correlation between species–to–locus RF distance
and posterior probability of LGT
d
posterior probability of ILS
0.2
0.4
0.6
0.8
40
60
80
Locus–to–gene RF distance

Figure 4

Figure 4. Variation of the model performance (AUC) among species trees.

A) Positive correlation between AUC and the average species-to-locus tree RF distance (D_{SL}), each point represents data from one species tree. B) Negative correlation between AUC and the average locus-to-gene tree RF distance (D_{LG}). C) Positive correlation between AUC and the percentage of LGT trees. D) Correlation between AUC and the percentage of LGT trees after controlling for D_{SL} is not significant ($p > 0.05$).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3489v1 | CC BY 4.0 Open Access | rec: 27 Dec 2017, publ: 27 Dec 2017


a
AUC
0.75
0.80
0.85
8
10
12
14
16
Average DSL
b
AUC
0.75
0.80
0.85
50
60
70
80
Average DLG
c
AUC
0.75
0.80
0.85
0.1
0.2
0.3
0.4
0.5
Percentage of LGT gene trees
d
Residual from linear regression on AUC with DSL
−0.05
0.00
0.05
0.1
0.2
0.3
0.4
0.5
Percentage of LGT gene trees