IMPACT OF RATES OF GENE DUPLICATION AND DOMAIN SHUFFLING ON SPECIES TREE INFERENCE WITH GENE TREE PARSIMONY

by

Tao Shi

A Thesis Submitted to the Faculty of the

DEPARTMENT OF ECOLOGY AND EVOLUTIONARY BIOLOGY

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

2013

STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Tao Shi

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

_____ August 2, 2013

Michael J. Sanderson Date

Professor of Ecology and Evolutionary Biology

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank everyone who has helped me with this dissertation. I thank my thesis advisor Michael J. Sanderson for regular meetings with me giving professional advice on the dissertation. I thank my committee members Michael Worobey and Frans Tax for giving valuable suggestions on my research. I thank Darren Boss who has always been very helpful with computation problems and setting up new programs in the lab server. I also want to thank lab member Derrick Zwickl and Barbara Dobrin for the discussion on my research during lab meetings.

I would also like to thank two professors I worked with very happily. I would really thank Peter Reinthal for the important opportunity of working as research assistant in the Natural History Museum, University of Arizona. I also thank Bruce Walsh who gave me valuable experience as teaching assistant in his wonderful genetics class.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES-*continued*

LIST OF TABLES

ABSTRACT

Genome sequencing technologies are providing huge quantities of data for phylogenetic inference. However, most phylogenomic studies exclude gene families, because many have a complicated history of gene duplication/loss and structural change by domain shuffling, especially in deep phylogenies. Gene tree parsimony (GTP) methods, which seek the species tree that minimizes the cost of gene duplication, have been successfully applied to gene families with frequent duplication history. Their utility and performance in the context of gene families with complex histories of gene duplication and domain reshuffling remains unclear. In this study, we analyzed 4389 gene families from six angiosperm genomes encompassing a wide range of duplication rates, and a broad diversity of domain architecture. Overall species tree inference accuracy increased monotonically with the inclusion of more gene trees, and high accuracy was achieved with 50-100 gene trees. The rate of gene duplication strongly influences species tree inference accuracy, with the highest accuracy at either very low or very high rates of duplication and lowest accuracy centered around one duplication per branch in the unrooted species tree. This is the opposite of the relationship between substitution rates on tree construction accuracy, in which intermediate rates have highest accuracy. Accuracy is generally higher in gene families with high domain architecture diversity but has high variance in families with relatively low domain architecture diversity. The latter is probably due to the high variation of gene duplication number for those gene families. We close with some discussion of potential impacts of domain evolution on phylogenomic reconstruction protocols in general, including its effect on alignment.

INTRODUCTION

Many phylogenetic studies now use sequence data from hundreds, thousands or even more genes to infer species tree using a variety of methods for assembling the sometimes conflicting information from different parts of the genome (Rokas et al., 2003; McMahon and Sanderson, 2006; de Queiroz and Gatesy 2007; Burleigh et al., 2006; Buerki et al., 2011). Gene trees, the phylogenies of individual genes, need not agree with each other or the species tree in which they are contained (Maddison 1997) for many reasons, including problems in phylogeny inference methods (e.g., misspecification of the model) and more direct biological factors such as incomplete lineage sorting, horizontal gene transfer, or gene duplication/loss (Wolf et al., 2002; Bapteste et al., 2004; Sanderson and McMahon, 2007; Ness et al., 2011; Steel et al., 2013; Salichos and Rokas, 2013). Many phylogenomic studies used supermatrix methods, which concatenate sequences from multiple loci into one large alignment. Typically they discarded all genes except select single copy orthologous genes using a variety of informatics protocols (de Queiroz and Gatesy 2007; Baker et al. 2009; Salichos and Rokas, 2013; etc.). Especially in eukaryotic genomes, including those with long histories of whole genome duplication(s) like angiosperms, much information is discarded by excluding any collection of genes that has paralogs in it. Moreover, this approach, though illuminating in many cases, has obviously not eliminated incongruence, with much remaining due to incomplete lineage sorting (Cranston et al.,2009; Yoder et al., 2013; Salichos and Rokas, 2013) and probable errors in orthology assessment. In this paper we explore the utility of including gene families explicitly in phylogenomic inference, focusing particular attention on a facet of gene family evolution that has not yet been well explored: structural diversity of domain architecture. Gene family and

domain architecture diversity can both record a history of evolutionary events leaving a signal about species relationships.

Gene Duplication and Species Tree Inference

Gene duplication has played a key role in the evolution of eukaryotes (Innan et al, 2010). In plants especially, many gene duplicates have been retained from ancient whole genome duplications, and some may have provided the basis for important innovations in seed plant and flowering plant evolution(Jiao et al, 2011). Many gene families have frequent turnover driven by gene duplication and loss, as in the large scale gene duplication/loss events for the bric-a-brac/tramtrack/broad complex ubiquitin-ligase (BTB) gene families after the split of rice and *Arabidopsis* (Gingerich et al., 2007). Frequent gene duplications and losses make it difficult to identify orthologous single copy loci for conventional phylogenetic analysis. "Gene Tree Parsimony" (GTP) (Goodman et al., 1979; Guigó et al., 1996; Maddison, 1997; Page and Charleston, 1997; Slowinski et al., 1997; Slowinski and Page, 1999; Sanderson and McMahon, 2007) exploits the occurrence of duplication events themselves to build trees, by searching for the species tree that minimizes the number of duplications across the set of gene trees (Goodman et al., 1979; Maddison, 1997; Guigó et al., 1996; Page and Charleston, 1997; Slowinski et al., 1997; Slowinski and Page, 1999; Sanderson and McMahon, 2007). GTP has proven useful in a limited but varied number of taxonomic groups including *Drosophila* (Cotton and Page, 2004), Elapidae (Serpentes) (Slowinski et al., 1997), vertebrates (Page, 2000; Cotton and Page, 2002), whales (McGowen et al., 2008), sharks (Martin and Burg, 2002), plants (Sanderson and McMahon, 2007; Burleigh et al., 2010), aquatic plant (Ness et al., 2011), and has also been used to root the whole eukaryotic tree (Katz et al. 2012). Although empirical experience with this (and

like minded) methods is growing, its performance has not been thoroughly evaluated, and a key biological feature of gene family diversity has not been explored in this context—the domain architecture of protein coding genes.

Domain Architecture

Here we define some terms. *Gene* refers to the exon and intron nucleotide sequences of a protein coding gene. A *gene family* is the set of phylogenetically related genes (we do not use *protein family*, here, but it is essentially the same as the gene family). A *protein* is the mature gene product, comprising a sequence of amino acids with structure and function. *Domains* are the "building blocks" of proteins, usually having specific 3D molecular structures or conformations and conserved functions within proteins; most of them can also fold independently (Anfinsen et al. 1961; Jackson 1998). A *domain family* is a set of phylogenetically related domains. *Domain architecture* (DA), also called domain organization, is the linear arrangement of domain(s) in an individual gene (or protein). For example, in Figure 1A, a specific gene family has genes from any of four domain families—A, B, C and D, and those four domains are arrayed in five unique domain architectures in five individual genes—ABCD, ABDC, AAD, C and CD. *Domain architecture diversity* of a gene family is then the number of unique domain architectures in the gene family.

More than half of proteins in eukaryotic organisms are multiple domain proteins (Wang and Caetano-Anolles, 2006). Many can co-exist with a large variety of other domains in different genes (Cohen-Gihon et al., 2011). Kinase domains which conduct phosphorylation, a typical example, are frequently associated with other domains such as Leucine Rich Repeats, LysM, S-locus and others. Collectively, 19 different domain families have representatives in the large

receptor-like kinase (RLK) gene family creating a very large number of distinct domain architecture in *Arabidopsis* and rice, with high functional diversity (Shiu et al. 2004). Another family with high domain architecture diversity is the F-box gene family, whose members mediate ubiquitination. A study in *Arabidopsis*, poplar and rice indicated that F-box subfamilies with frequent duplications were associated with domain architecture diversification, while those with few duplication also has little or no domain architecture variation (Xu et al., 2009).

Domain Architecture and Phylogenetic Signal

The idea that domain architecture changes may contain phylogenetic information is supported by a few studies that have used domain presence-absence data to build trees broadly consistent with current understanding of the tree of life (Yang et al., 2005; Wang et al, 2006; Fukami-Kobayashi et al., 2007). Nonetheless it is unknown whether increased domain architecture diversity in a single gene family might improve or degrade phylogenetic inference using that family. Phylogenetic analysis depends on the accuracy of alignment (e.g., Wong et al, 2008), but sequence alignment programs can easily be misled by changes in domain architecture (Phuong et al., 2006), just as they can by large indels in an alignment of a single domain (Figure 1.B). On the positive side, gene families with diverse DAs might tend to have long sequence length owing to the presence of multiple domains and to include many genes , which might provide more phylogenetic information by species tree inference strategies such as GTP.

The goal of our study is to explore the impact of gene duplication rate and domain architecture diversity on the accuracy of species tree inferences based on gene and/or domain family trees. Species tree construction even using the relatively fast GTP algorithm is still NP complete (Górecki et al., 2012), so we restrict attention to that algorithm, but anticipate that

broad conclusions may carry over to likelihood based inference methods (McCormack et al., 2009). The basic approach will be to assume a "known" species tree for six angiosperms for which complete genome data are available, and evaluate species trees built with GTP using a variety of subsets of the data, in order to gauge the accuracy entailed by using domain families and gene families with various properties. Our current study will provide valuable information not only for understanding species tree—gene tree relationships at a genome wide scale, but also for future design of more sophisticated alignment strategies and improvement/upgrade of species tree inference algorithms suitable for genome-wide data such as GTP .

MATERIALS AND METHODS

Species and Species Tree

We selected six angiosperm species with whole genome sequences separated by enough divergence to have accumulated diversity in their gene families and domain architecture, and also to avoid short edge lengths and incomplete lineage sorting, which is a confounding source of gene tree incongruence. These included four eudicots (*Medicago truncatula*, *Glycine max*, *Fragaria vesca*, *Arabidopsis thaliana*) and two monocots (*Oryza sativa ssp*. *japonica* and *Sorghum bicolor*)(Figure 2) (Hilu et al., 2003; Jansen et al, 2005; APG III [Angiosperm Phylogeny Group III]. 2009; Soltis et al., 2011; etc.).

Genome And Gene Family Data

Gene families were obtained from the Plant Plaza 2.5 database, which was constructed by all-against-all BLASTP (cutoff E_value=1e-05) and graph-based clustering method (Markov clustering implemented in Tribe-MCL, MCL_I=2) to cluster homologous protein sequences (Enright et al., 2002; Proost, et. al, 2009). Domain annotations are critical to our study and we therefore used two independent sources: the Pfam (Bateman et al., 2002) and Superfamily (Wilson et al., 2009) databases for the protein domain annotations, which were also downloaded from Plant Plaza 2.5 database (Hunter et al., 2009, Proost, et. al, 2009). From the 39144 gene families in Plant Plaza 2.5, a subset of 4389 gene families having complete six taxon coverage were used in analyses of the relationship between species tree inference accuracy and gene duplication rates. Only a subset of these gene families are annotated with respect to domain architecture: 3178 gene families with Pfam domain annotation (and all six species) and 1494

gene families with Superfamily domain annotation (and six species) were used in the analysis of domain architecture diversity.

Gene Tree And Species Tree Reconstruction

Full length protein-coding DNA sequences (CDS) for each gene family were aligned with TranslatorX, which translated DNA into proteins for alignment via MUSCLE (Edgar 2004) and translated back into DNA alignment with the default parameter setting (Abascal et al., 2010). We inferred the phylogenetic tree for each gene family (gene tree) initially using RAxML (CAT + GTR) (Stamatakis 2006) and Fasttree 2. (Price et al., 2010) (CAT+GTR). However, 14 gene families with very large number of leaves failed to finish in the time we allotted for each run, whereas all gene families finished using Fasttree 2, and we therefore proceeded with the latter, although we did see a slight advantage to RAxML's performance with respect to the likelihood score on those trees that did terminate.

For each gene family, a species tree was estimated via gene tree parsimony (GTP) as implemented in Duptree (Bansal et al., 2007; Wehe et al., 2008). Duptree takes the unrooted gene tree(s) generated by Fastree 2 as input and performs a heuristic search among rooted species tree using a fast local rooted subtree pruning and regrafting (rSPR) algorithm (Bansal et al., 2007) in order to minimize the total number of duplications (Wehe et al., 2008). Because our gene trees were unrooted, all the gene trees were automatically rerooted with option "-r opt" in Duptree by minimizing the total number of duplication (Wehe et al., 2008). A similar approach was taken for construction of domain trees (see below).

For some analyses a species tree was estimated from a collection of gene trees or domain trees In that case, Duptree was used again to optimize the total number of duplications across all trees in the collection.

Species Tree Accuracy and Duplication Rates

For each of the 4389 gene families with six taxon coverage, the number of duplications of each gene family were obtained by two ways—one by reconciling the gene tree to the known species tree via Notung 2.6 (Chen et al., 2000); the other by directly using Duptree's reconciliation based on its estimate of the species tree, which might not always match the known tree (Bansal et al., 2007; Wehe et al., 2008). We define a relative duplication rate of gene family as the inferred number of duplications per branch in the unrooted species tree: number of duplications/(2*n*-3)", where $n = 6$. We sorted the 4389 gene families by inferred relative duplication rate, then grouped them into bins of two sizes for display (so there were a total of 22 or 44 bins).The accuracy of species tree inference of each bin was calculated in two ways: first, by the percentage of the gene families producing species trees that completely agree with the known species tree; second, by how close the average distance of species tree topologies is to the known species tree. The distance of species tree topology of each gene family to the known species tree topology was measured by TOPD implementing NODAL method in terms of the root-mean-squared distance (RMSD) of the two topology matrices (Puigbo et al, 2007). Relationship of the accuracy and duplication rate were plotted in Gnumeric 1.12.1 and fitted with Moving Average as trend line with option (span 5 and averaged abscissa). We also summarized gene losses for every gene family using the reconciliation algorithm in Notung 2.6 (Chen et al., 2000). The relationship between duplication and loss for each gene family is plotted in Figure 6.

Domain Architecture and Other Sequence Features

For each gene with domain annotation available, its domain architecture (DA) was obtained by parsing the text description using a PERL script (for example an *Arabidopsis thaliana* sequence AT1G71830 with a Pfam DA of "PF08263-PF00560-PF00560-PF00560-PF00069"). For each gene family, the numbers of non-redundant DAs were counted using the two databases, Pfam and Superfamily, separately. DNA sequences for each domain were extracted from protein-coding DNA sequences (CDS) based on the Pfam and Superfamily annotations by custom PERL scripts. For each gene family, domain sequences annotated by the same domain ID were considered to be part of a "domain family", for example domain sequences with "PF00069" and "PF00560" are defined as two domain families, and the numbers of domain families for each gene family were counted for Pfam and Superfamily annotation systems separately. Other gene family features including the number of leaves (genes), the average gene sequence length, total number of nucleotides, total number of domain sequences, the characters (sites) in the sequence alignment were summarized via custom scripts. Basic statistics were conducted and summarized with Gnumeric 1.12.1(Table 1).

Species Tree Inference Accuracy and Domain Architecture Diversity

The effect of domain architecture diversity on species tree accuracy was assessed similarly to how it was performed for duplication rates above, except the number of families with domain annotations from Pfam or Superfamily available is different, and we used 32 bins in binning trees with different amounts of domain architecture diversity.

Domain Tree and Concatenated Domain Tree Reconstruction

Using the subsequence for each domain extracted based on annotation, phylogenies for each domain family ("domain tree") were reconstructed using the same protocol as that described for gene tree reconstruction. Domain trees based on Pfam and Superfamily were generated separately. In a final treatment sequences from each domain in each gene were also concatenated following the domain architecture order, excluding any non-domain sequence, and concatenated and aligned to build phylogenetic tree ("concatenated domain tree"). For each gene family, the concatenated domain tree was established with the same alignment and tree reconstruction method described in “gene tree” reconstruction.

Species Tree Inference Accuracy from Gene Trees, Domain Trees And Concatenated Domain Trees

First, we binned gene families into three groups according to ascending order of DA diversity levels (No. of DAs in the gene family)—"low DA", "Mid DA" and "high DA". For each DA level, we assembled different sized collections of gene trees by randomly sampling the pool of gene families with replacement (1, 3, 5, 10, 20, 50 and 100 gene trees). At each sample size, we constructed 200 replicate collections of gene trees. For each of these collections, species trees were inferred from "gene trees", "domain trees" and "concatenated domain trees" using Duptree. Accuracy was assayed by the percentage of the 200 replicates in which an estimated species trees agrees with the known species tree.

RESULTS

Species Tree Inference Accuracy and Gene Duplication Rate

The number of duplications per gene family (determined by reconciliation against the known species trees) is highly skewed across the 4389 gene families with all six species, with a mean of 21.2 and a median of 7 duplications (Figure 3). Gene families with lowest relative duplication rate and highest duplication rates have relative higher accuracy (Figure 4, Figure 5) or shorter average topology distance to the known species tree (Figure 6), while those gene families with relative duplication rate close to one duplication per branch in the unrooted species tree have the lowest accuracy (Figure 4, Figure 5) or the longest average topology distance (Figure 6). In general, the relationship between accuracy and relative duplication rate reveals a "V" shape if accuracy is represented by percentage of gene families agreeing with the known species tree (Figure 4, Figure 5) or inverted "V" shape if accuracy is measured by average tree topology distance to the known species tree (Figure 6). Similarly, the "V" or "inverted V" patterns of accuracy and duplication rate relationship were found in analyses based on duplications from Notung 2.6 under 22 bins (Figure 7, Figure 8) and analyses based on duplications from Duptree under 44 bins (Figure 9, Figure 10, Figure 11) or 22 bins (Figure 12, Figure 13).

Also we found a highly strong positive correlation between the number of duplications and the number of losses generated by Notung 2.6. (R^2=0.763845, $p<0.01$)(Figure 14), but gene losses after gene duplications severely confound orthology/paralogy relationship creating conflicting phylogenetic signals among duplicates as previous studies suggested. However, the accuracy starts to increase when a gene family contains enough duplicates as GTP itself is a

reconciling process that leverages the conflicting signal with enough data (gene copies), for example gene family *HOM000277* (Figure 15).

Gene Duplication and Domain Architecture Diversity

Domain architecture (DA) variation is common across our sample of gene families. In the 3178 gene families having a Pfam domain annotation, 62.9% gene families have more than one DA. These include 1608 gene families with more than one domain family and 391 with only one domain family. Of the other 37.1% of gene families from Pfam with only one DA, there are 64 gene families with more than one domain family and 1115 gene families with only domain family (Figure 16). Results are qualitatively similar for the Superfamily annotated gene families (Figure 17). Although most gene families have only one or two domain architecture, there is a long tail toward increasing diversity of domain architecture across gene families (Figure 18; Figure 19). The regression of the number of duplications per family on domain architecture per family is significant for both Pfam ($R^2$is 0.508457, and p-value<0.01; Figure 20) and Superfamily ($R^2$is 0.357371, and p-value<0.01; Figure 21), indicating that structural divergence in domain architecture is correlated with gene duplication.

Species Tree Inference and Domain Architecture Diversity

The accuracy of species tree inference is positively correlated with domain architecture diversity (log10 transformed) for those families with relative high domain architecture diversity indicated by a smoother “moving average” trend line, but for those with relatively low domain architecture diversity the correlation is not as apparent as those big spikes in the “moving average” trend line (Figure 22, Figure 23). Similar patterns were found when we plotted the

average tree topology distance against the domain architecture diversity (log10 transformed)(Figure 24, Figure 25). The weak correlation with accuracy for those bins with low domain architecture diversity is probably due to higher variance in the number of duplications in those gene families (Figure 26, Figure 27), such that the duplication rate of a gene family could be a much stronger driving force affecting species tree inference accuracy.

Comparison of Species Tree Inference Accuracy among Gene Tree, Domain Tree, and Concatenated Domain Tree Methods

When we compared species tree inference accuracy from gene tree, domain tree(s) and concatenated domain tree under increasing sample sizes, we saw three patterns (Figure 28, Figure 29). One obvious trend is that the accuracy increases as the sample size (No. gene families) increases for all three methods (Figure 28, Figure 29). The second general trend is that species tree inference accuracy increases as domain architecture diversity gets higher given the same sample size for all three methods (Figure 28, Figure 29). The third and perhaps most interesting trend is that inference using gene trees generally has a higher accuracy than the either domain trees or concatenated domain trees, and the accuracy for latter methods is very similar (Figure 28, Figure 29). These effects are dependent on the domain architecture diversity. For example, in the Pfam based analysis under "mid DA" level, the accuracy for gene tree is 58%, for the domain tree is 52.5%, for the concatenate domain tree is 47.5%, but at "high DA" level, the differences among gene tree, domain tree and concatenated domain tree were not obvious (Figure 28, Figure 29), probably due to high duplication rates for those gene families. Thus, using domain sequence only does not help to increase species tree inference accuracy, but combining more gene family data together (by increasing sample size of gene trees) does. This

suggests that reliance on only domain region sequences may be excluding informative sequences from nondomain regions. Alternatively, domain sequences are shorter than gene sequences, and thus the domain tree or concatenated domain tree may simply tend to have higher error variance.

Relationship among Domain Architecture Diversity, Sequence Length and No. Characters in Alignment

The positive relationship between species tree accuracy and domain architecture diversity may be driven in part by the correlation between the number of duplications and domain architecture diversity. Other sequence features may also drive this pattern. Regression tests of domain architecture diversity on other features indicated that domain architecture diversity has no relationship with average sequence length (R^2=0.03526, $p<0.01$ for Pfam based analysis, R^2=0.06441,$p<0.01$ for Superfamily based analysis), but a strong relationship with the number of characters(sites) in the alignment (R^2=0.39296, $p<0.01$ for Pfam based analysis; R^2=0.44180, $p<0.01$ for Superfamily based analysis) (Table 1). Domain architecture diversity was strongly correlated with the number of duplications (R^2=0.50846, $p<0.01$ for Pfam based analysis; R^2=0.35737, $p<0.01$ for Superfamily based analysis), and the number of leaves in the gene tree (R^2=0.51775, $p<0.01$ for Pfam based analysis; R^2=0.34749, $p<0.01$ for Superfamily based analysis). These results imply that higher domain architecture diversity in a gene family is not always associated with longer sequence length, but it does correlate with longer alignments. Much of this increased length is due to the introduction of long gaps corresponding to nonhomologous domains. Gaps in general decrease alignment accuracy and should degrade phylogenetic signal (Dwivedi and Gadagkar, 2009), but longer alignments on average contain more sites and thus possibly more phylogenetic signal. The average SH-like local supports from

the Shimodaira-Hasegawa test, a quick estimate for the confidence of each split in the tree, showed a weak negative correlation with number of characters in alignments (Pfam dataset: R^2=0.00730, p<0.01; Superfamily dataset: R^2=0.00953, p<0.01) (Figure 30, Figure 31) but a weak positive correlation with sequence length (Pfam dataset: R^2=0.05982, p<0.01; Superfamily dataset: R^2=0.06095, p<0.01) (Figure 32, Figure 33). This indicated longer sequence length indeed in general contains more phylogenetic signal (better reliability of tree topology) but a greater number of characters(sites) in alignment which possibly are inflated by indels might not. However, more work is needed to tease these conflicting factors apart.

DISCUSSION

Accuracy of Species Tree Inference in Relation to Rates of Gene Duplication

A key result of this paper is the "inverted" curve of the accuracy of GTP reconstruction versus rate of gene duplication. Unlike the pattern observed for substitution rates of single copy genes, in which the best accuracy is achieved at an intermediate rate of substitution (Yang, 1998; Bininda-Emonds and Sanderson 2001), the best species tree inference is made at low or high rates of duplication, with intermediate rates being worst. Our sample of gene trees is very large, encompassing the entire genomes of six divergent angiosperms, a clade at least 140 MYR old, so these results are likely to be generalizable to many other taxa. When rates of duplication are very low or nonexistent, the gene tree closely matches the species tree that would be obtained from use of a single copy orthologous locus. Any conflict between such "nearly" single copy gene trees is resolved by gene tree parsimony by minimizing an optimality score that is not extremely different from what would be done by algorithms trying to minimize lineage sorting events in species tree inference (Bansal et al., 2010) where the genes really were single copy. In fact there is a relationship between the two optimality criteria (Bansal et al., 2010). Thus it is not surprising that GTP performs best in the limit of decreasing rates of gene duplication. Unlike the situation with substitution rate and accuracy, in which no substitutions mean no signal, no duplications do entail the signal retained by the individual single copy gene trees. At the other extreme, gene families with high duplication rates evidently provide a large sample size of evolutionary events for GTP to use in inference, and accuracy even in large complex gene families can be quite good. It is the middle ground of intermediate rates of duplication where species tree inference suffers most, in the vicinity of one duplication per branch of the tree. A key unresolved question is how

this pattern of accuracy will scale to larger species trees or to a set of taxa (clade), such as a genus, whose most recent common ancestor was relative young such 20 million years.
In particular, does it become increasingly more difficult to find gene trees on the high-rate side of the dip in accuracy as species tree size increases? If so, users of GTP may well have to filter gene trees down to those with fewer duplications.

The appeal of a species tree construction method like GTP over more conventional supermatrix methods is that orthology detection and paralog removal can be skipped. This accomplishes two things: it brings to bear a much larger fraction of the genome on species tree inference, and it sidesteps one of the most problematic steps in phylogenomic analysis (Chen et al., 2007; Salichos and Rokas, 2011). Orthology detection has been reviewed in many places (Kuzniar et al., 2008; Altenhoff and Dessimoz, 2009; Dessimoz et al., 2012), but we note a few relevant issues here. Ortholog relationships can be obscured for methodological and biological reasons. Sim ulations suggest matters become worse at higher duplication and loss rates (Dalquen et al. 2013), which likely reflects noise and model misspecification. Salichos and Rokas (2011) found that all methods tend to produce many more false positive results in the context of whole genome duplications. Loss of true orthology relationships is especially severe then because of reciprocal loss of paralogs in descendant lineages. In three polyploid yeast genomes about 20% of loci suffered reciprocal loss (Scannell et al., 2006). Also genomic analysis on the closely related taxa, *Tetraodon* and zebrafish, indicated that reciprocal loss of genes after whole genome duplication is common and could be highly associated with reproductive isolation and the speciation process (Semon and Wolfe, 2007).

Ironically, some of the best orthology detection tools use gene tree reconciliation——but they require a species tree as input. If species tree construction is the goal, one is left with other

methods such as reciprocal best blast hits (Li et al., 2003). The strong performance of GTP in species tree inference supports the notion of a more integrative approach to inference with genome scale data sets across deep divergences in the tree of life. For example, this might entail an iterative procedure starting with all gene families and using GTP (or a more nuanced model-based alternative) to estimate an approximate species tree; use this to identify single copy orthologs to be used in concatenated analyses; then return those resulting trees to the pool of "gene" trees for species tree inference. Important variations on this idea might account for linkage and synteny if that level of genome scale data are available.

Domain Architecture Diversity and New Alignment Approaches

Previous studies indicated that domain and domain architecture changes are not only common but also frequent during plant evolution: domain gain rate is about 6.64/Ma, domain loss rate is about 6.11/Ma on average, (gene) fusion rate is about 4.59/Ma and (gene) fission rate is about 1.98/Ma (Kersting et al., 2012). As an important feature for gene diversity, domain architecture has not yet been paid much attention in molecular phylogeny. A main contribution of our current analysis is characterization of the relationship between duplication events, domain architecture diversity, and species tree inference accuracy. High duplication rates generate gene families with abundant signal for accurate species tree inference overall, but it also brings about high levels of domain architecture diversity, which can pose significant problems for sequence alignment (and by extension, phylogenetic inference). For example, if a pair of genes in the same gene family have domain architectures, “AB” and “BA”, many alignment programs used by default in phylogenomics pipelines will have trouble aligning the homologous domains. Alignment not only can affect tree inference, it can affect the results of other down-stream

evolutionary analyses such as positive selection detection and branch length measurement (Kumar and Filipski , 2007; Lunter et al., 2007; Wong et al, 2008; Jordan and Goldman, 2012; Blackburne and Whelan, 2013). Future studies on domain architecture evolution models can provide valuable information for designing new multiple sequence alignment algorithms to deal with gene families with complex domain architecture. We would not be surprised if the phylogenetic analysis of such families will raise as many issues regarding data set assembly and analysis as the fundamental question of orthology and paralogy has raised over the years. Indeed, consideration of domain evolution leads to a fascinating conceptual hierarchy of domain trees within gene trees within species trees——the intrinsic complexity of which will have to be confronted algorithmically as more complete genomes and transcriptomes become available.

APPENDIX A: TABLES AND FIGURES

Table 1. Correlation of determination (R^2) for number of domain architectures and other gene family feature. Note that all correlation tests with p-value<0.

		Pfam annotation system	Superfamily annotation system
Correlation of determination (R^2) for No. Domain Architectures(DA) and other features for gene families	No. Gene families	3178	1494
	No.Domain families	0.758039	0.669565
	No. Leaves	0.51775	0.347495
	No. Duplications	0.508457	0.357371
	No. DA/No. Leaves	0.008716	0.026524
	No. Domains	0.425372	0.167357
	No. Characters	0.392957	0.441798
	No. Leaves × No. Nucl.	0.664012	0.387604
	No. Nucl	0.675441	0.39015
	No. Domains/No. Leaves	0.130507	0.003361

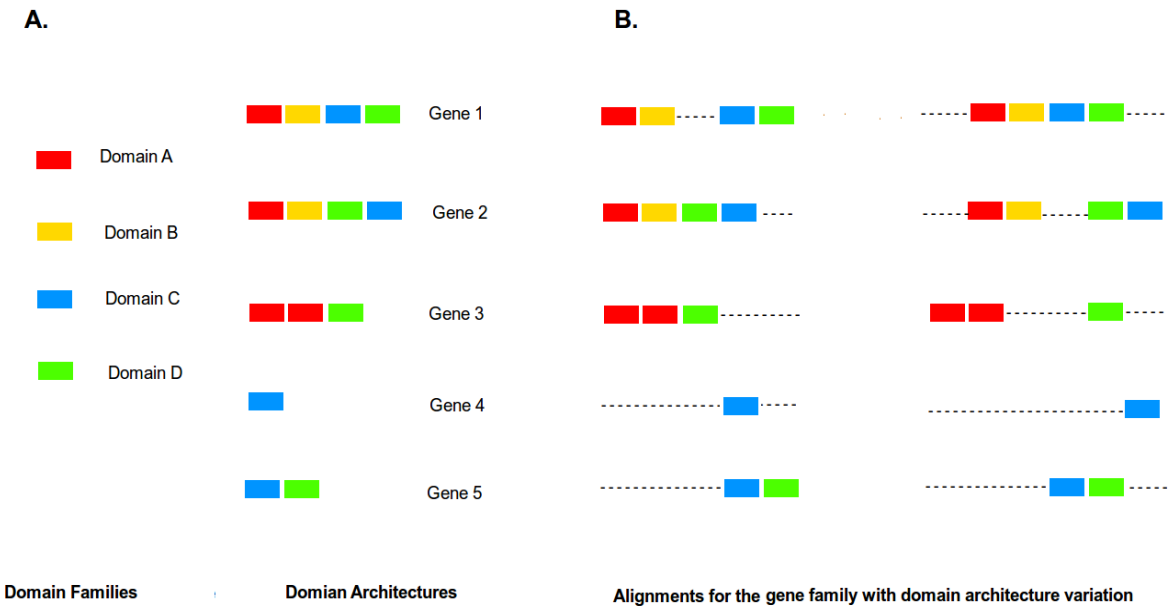


Figure 1. Domain architecture diversity and alignment. A: definition of domain families and domain architecture. B: Ambiguity of homologous sequence alignment produced by domain architecture variations including alignment of nonhomologous domains and failure of alignment of homologous domains.

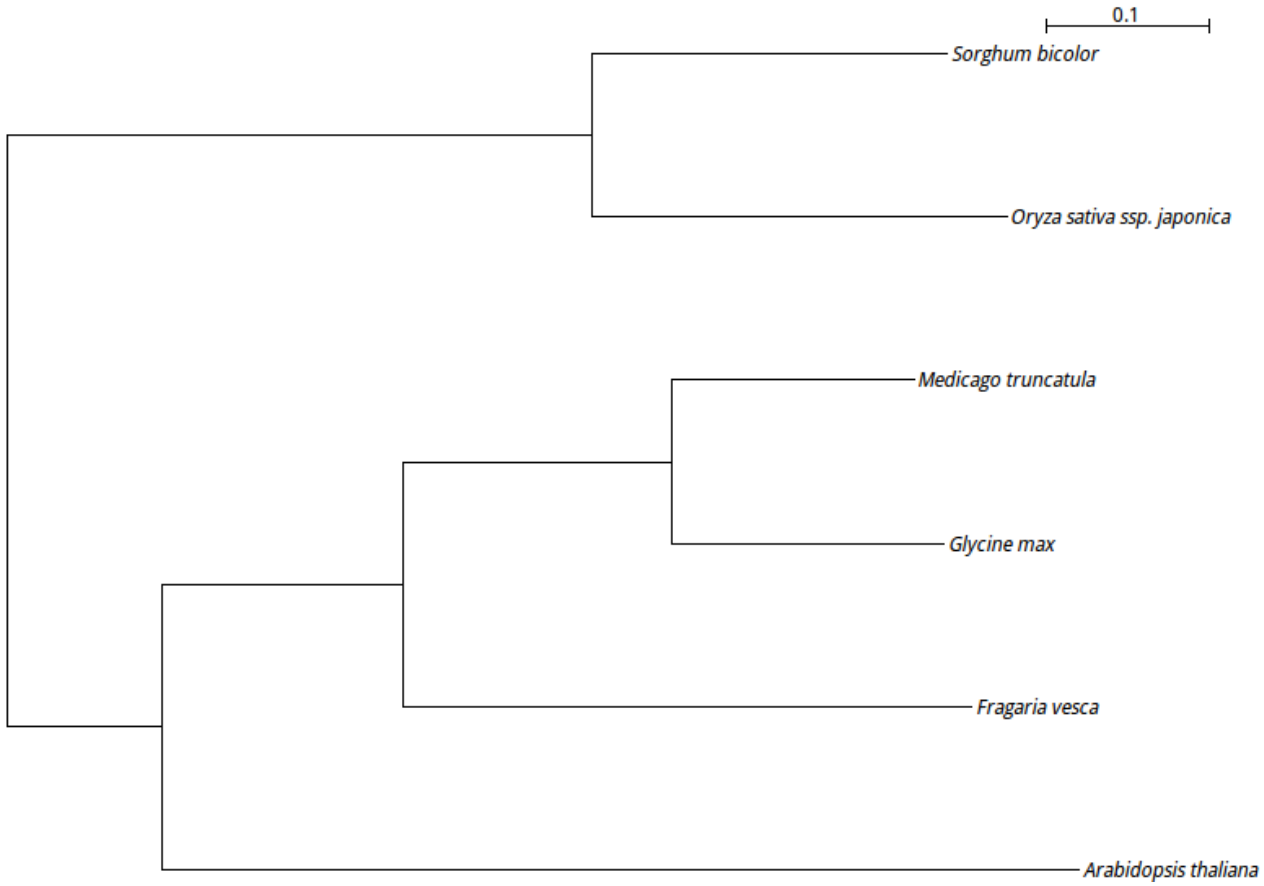


Figure 2. The known species tree of the six taxa in our research. The tree is based on single copy nuclear gene family HOM006375 whose topology is consistent with previous reports.

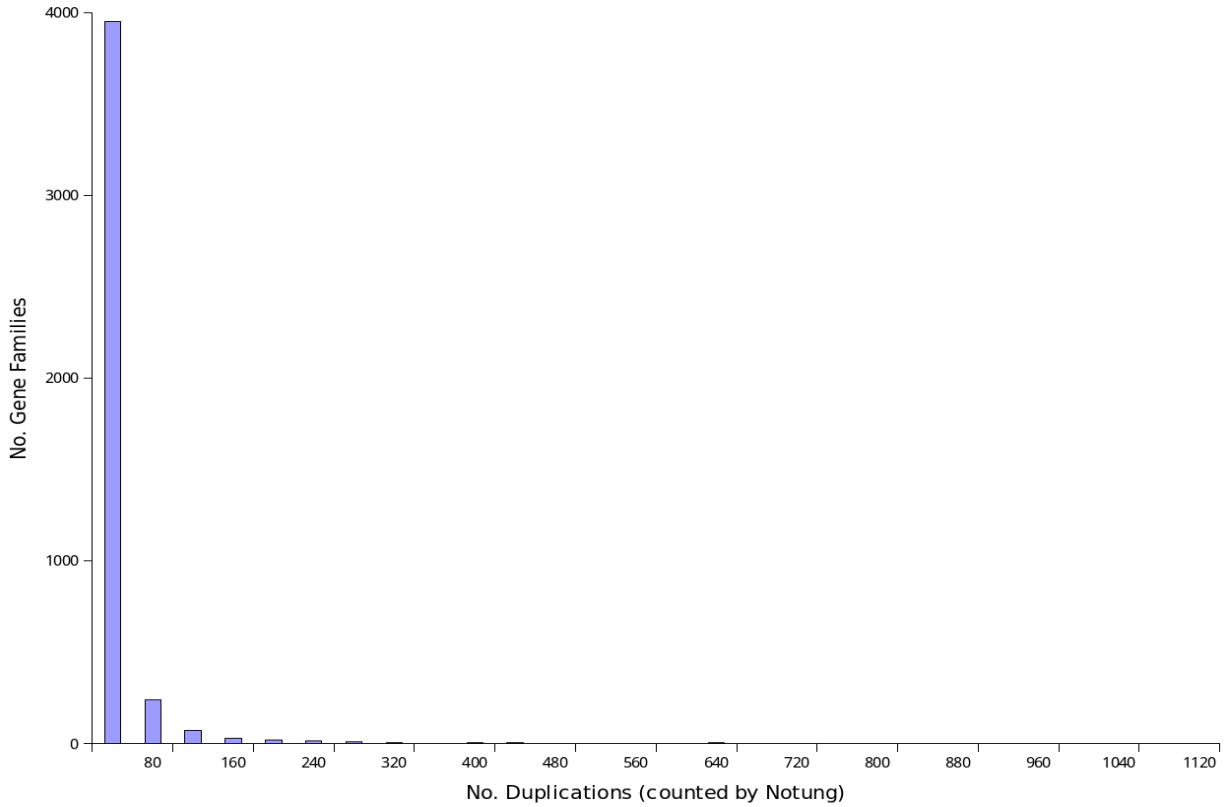


Figure 3. Frequncy of gene duplications in gene families. Gene family frequency distribution according to number of duplications for the 4389 gene families with six taxa coverage. Average number of duplications =21.2 , Median number of duplications =7 and 78 gene families have no duplication. Number of duplications are estimated by Notung 2.6.

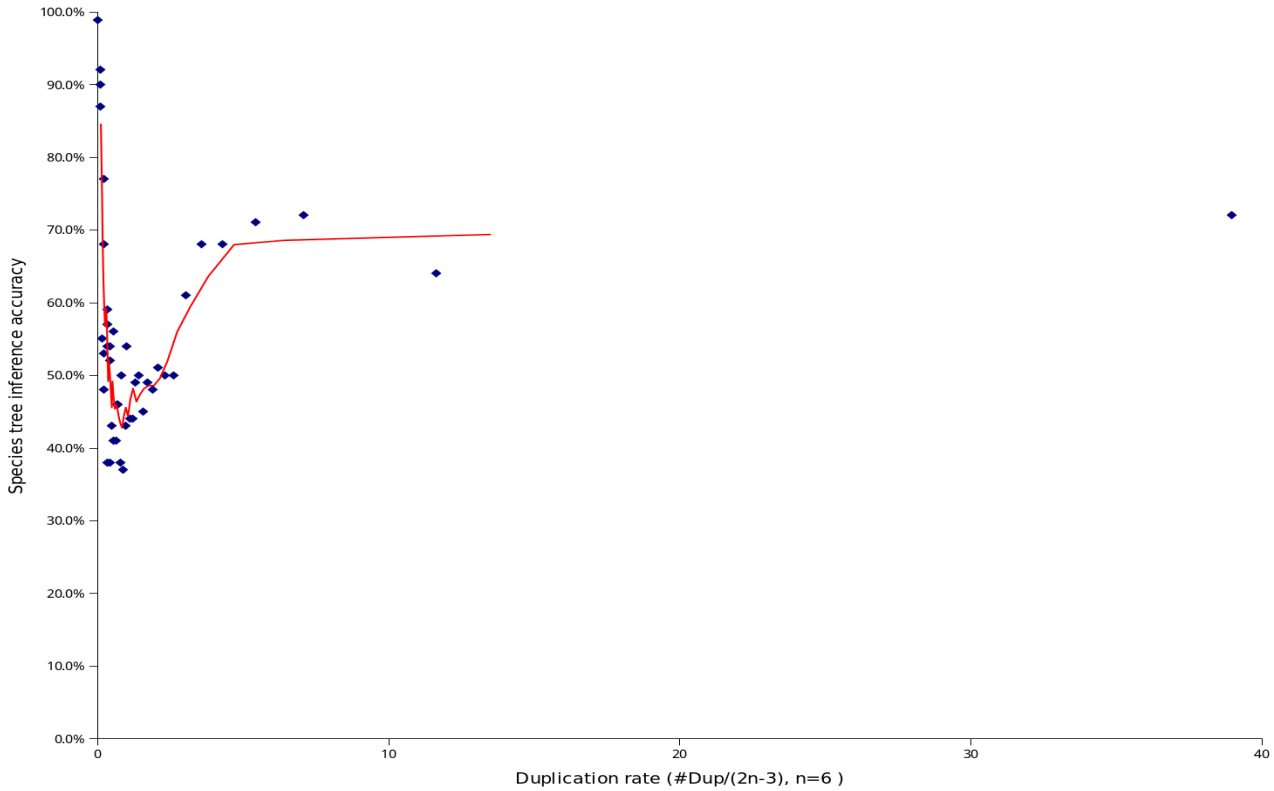


Figure 4. Relationship between species tree inference accuracy and gene duplication rate. The 4389 gene families with six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "44 bins". Each dot represents a bin. Number of duplications are estimated by Notung 2.6.

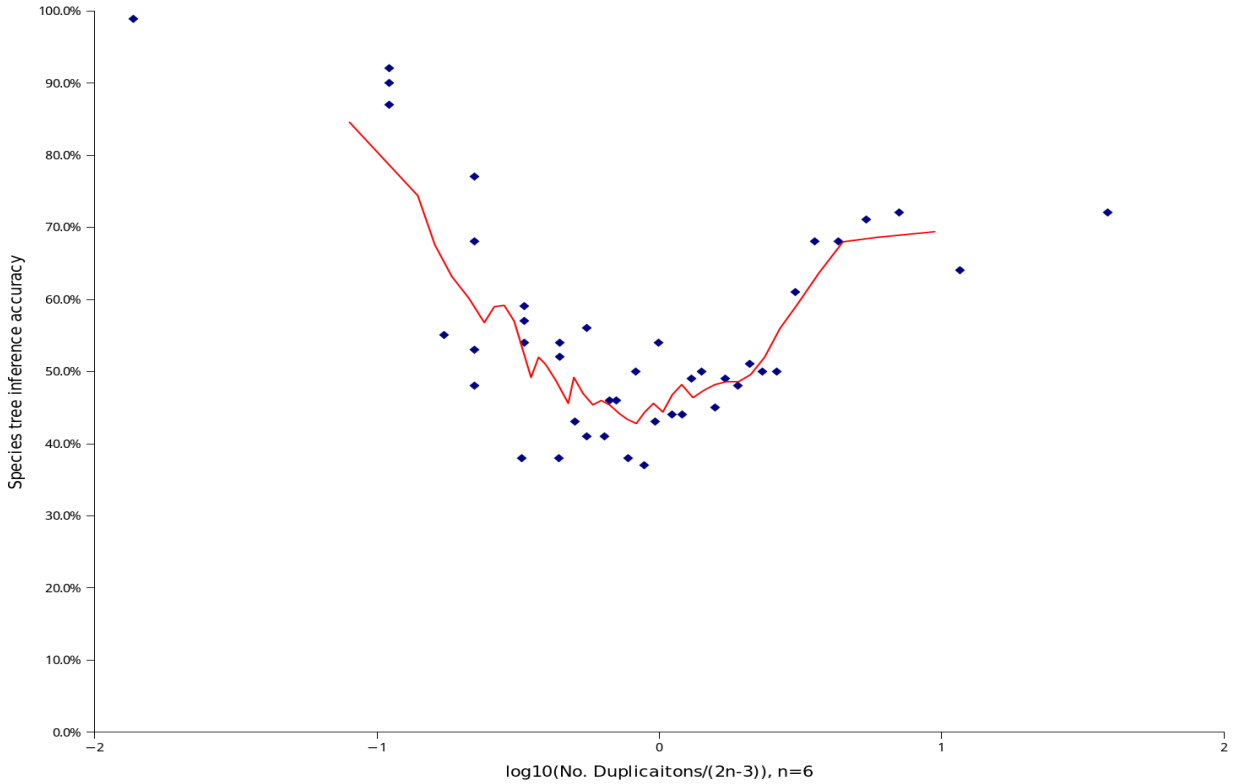


Figure 5. Relationship between species tree inference accuracy and gene duplication rate (log10 transformed). The 4389 gene families with six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "44 bins". Each dot represents a bin. Number of duplications are estimated by Notung 2.6.

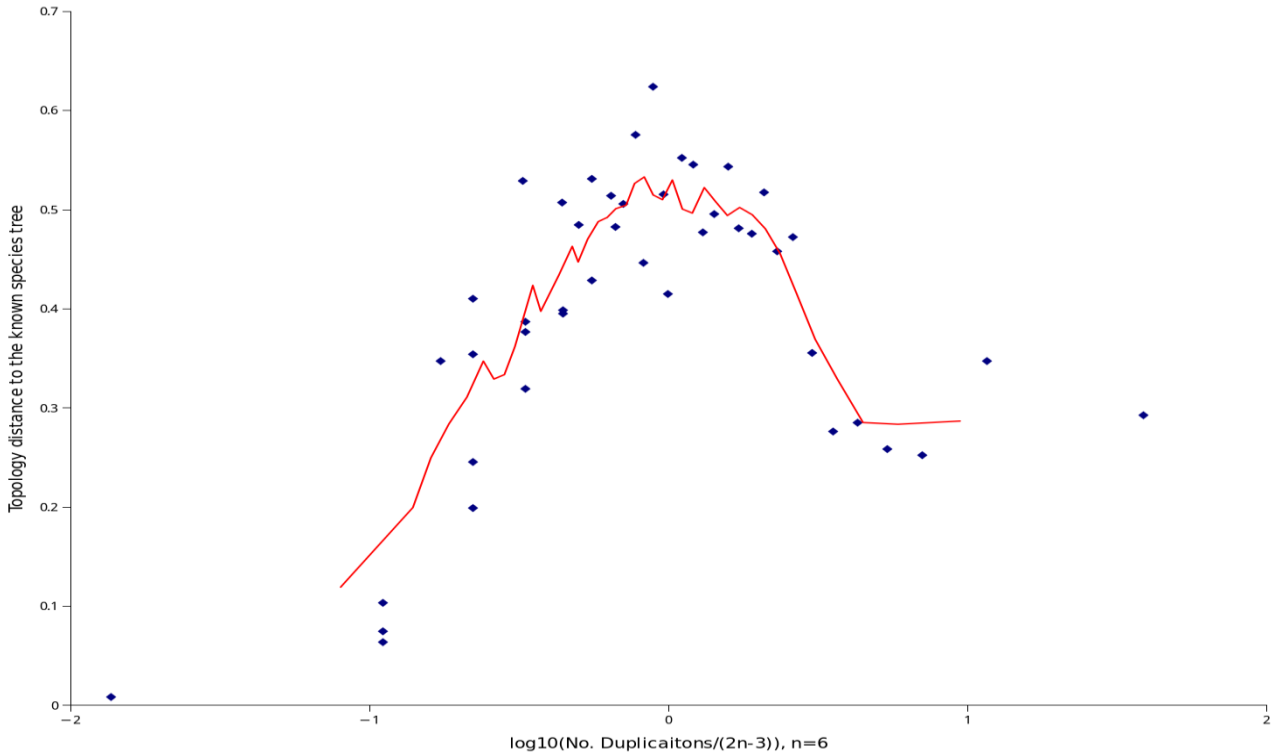


Figure 6. Relationship between average inferred species tree topology distance to the known species tree and gene duplication rate (log10 transformed). The 4389 gene families with six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "44 bins". Each dot represents a bin. Number of duplications are estimated by Notung 2.6.

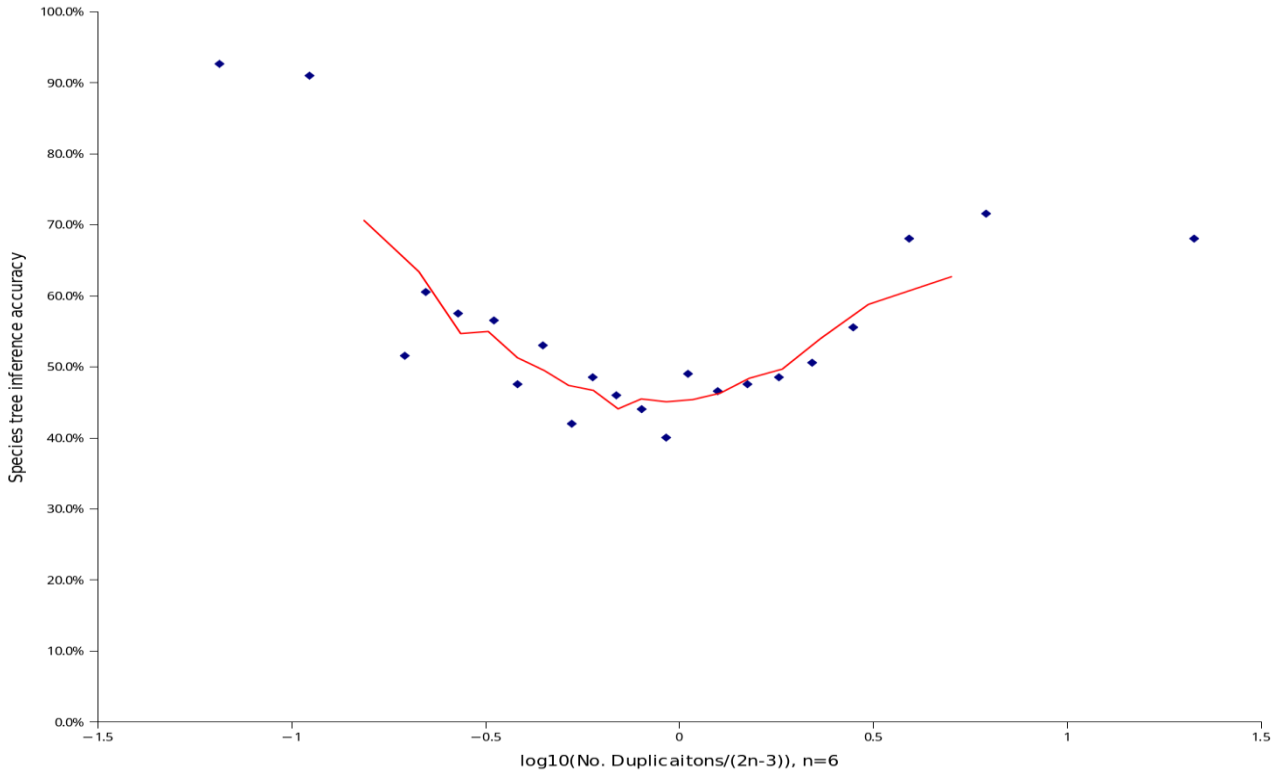


Figure 7. Relationship between species tree inference accuracy and gene duplication rate (log10 transformed). The 4389 gene families with six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "22 bins". Each dot represents a bin. Number of duplications are estimated by Notung 2.6.

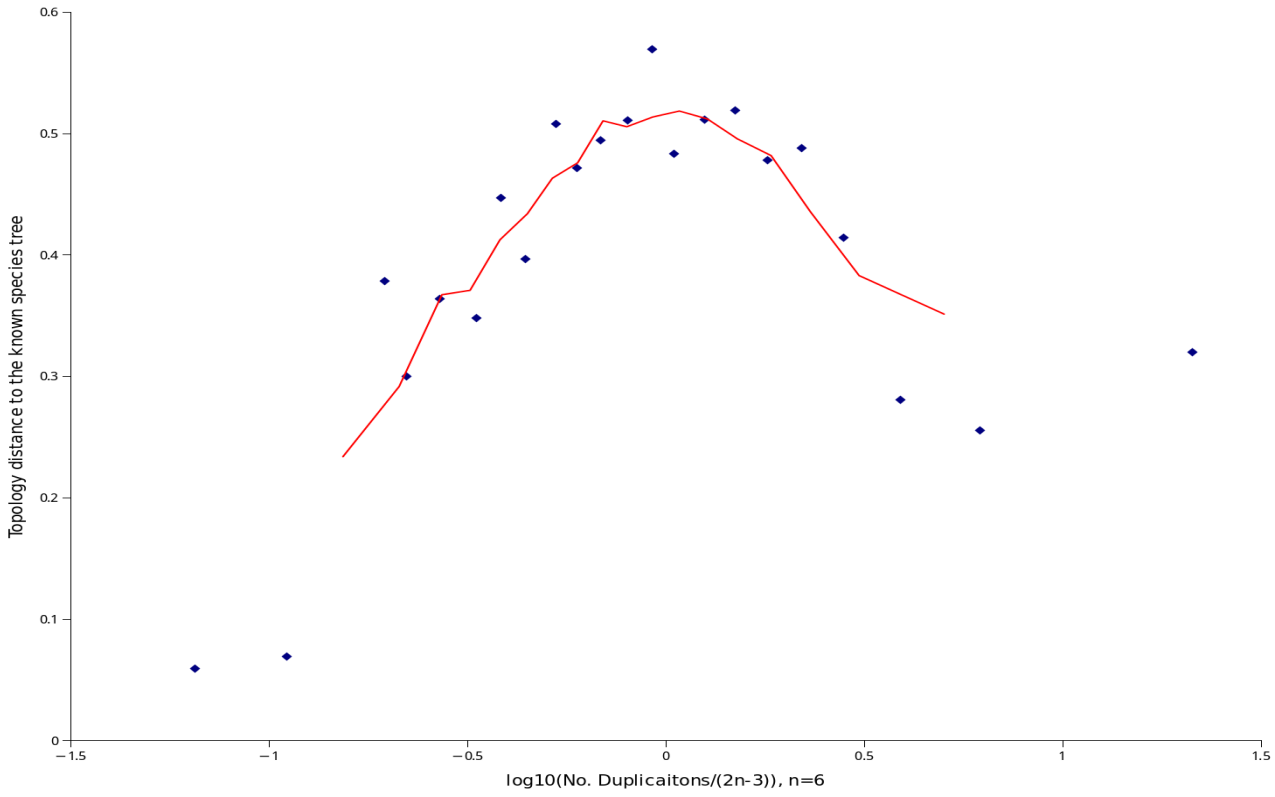


Figure 8. Relationship between average inferred species tree topology distance to the known species tree and gene duplication rate (log10 transformed). The 4389 gene families with six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "22 bins". each dot represent a bin. Number of duplications are estimated by Notung 2.6.

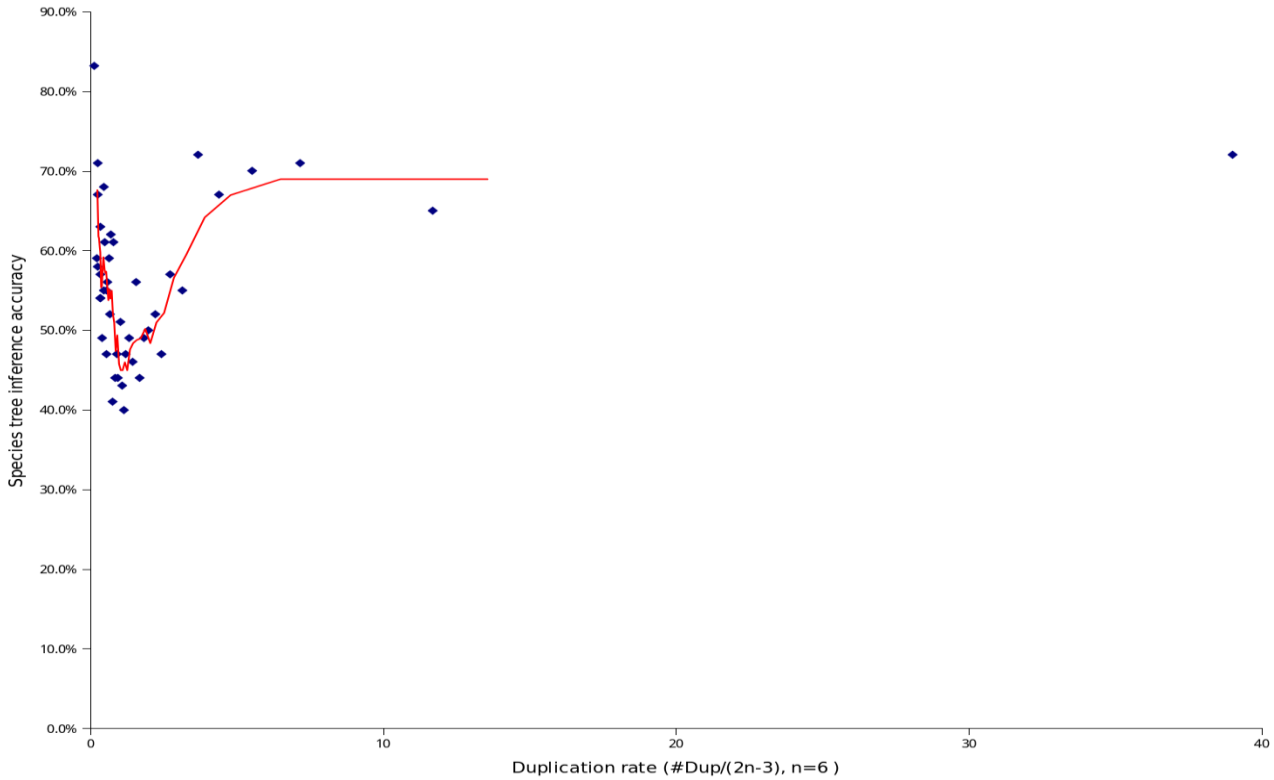


Figure 9. Relationship between species tree inference accuracy and gene duplication rate. The 4389 gene families with six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "44 bins". Each dot represents a bin. Number of duplications are estimated by Duptree.

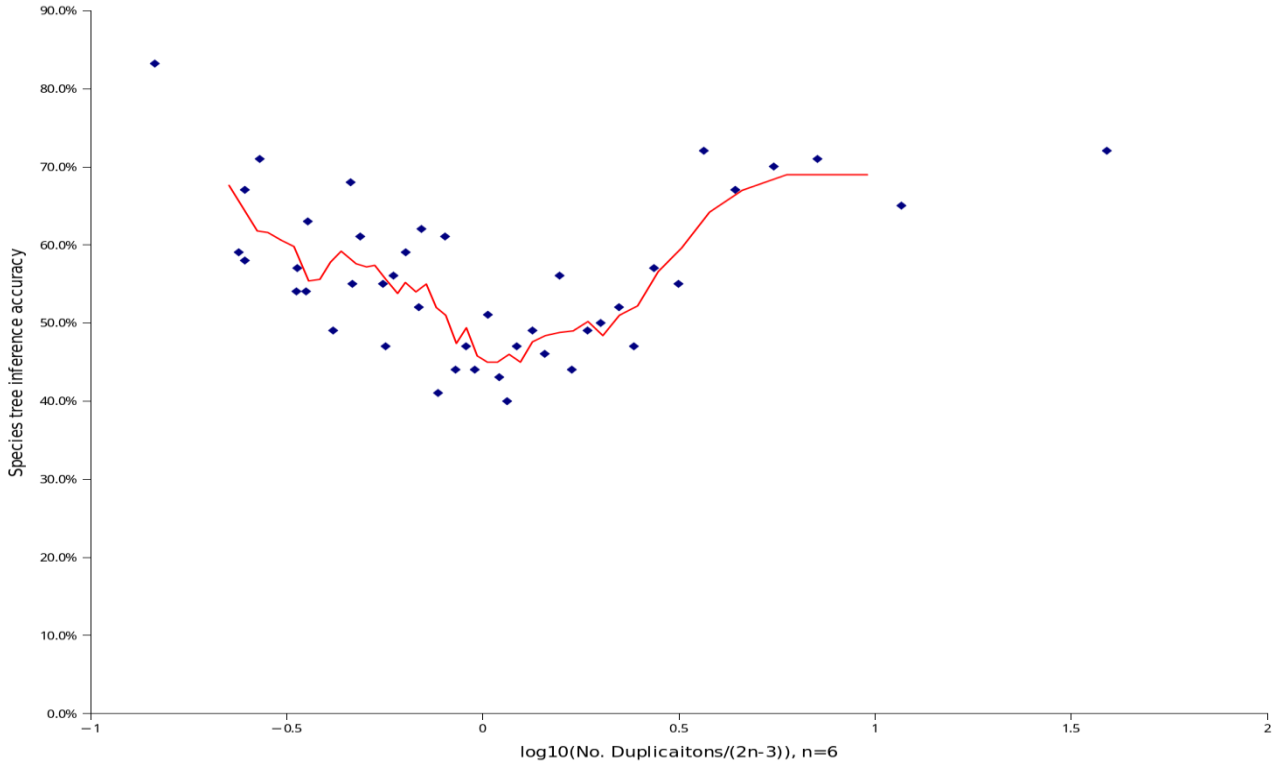


Figure 10. Relationship between species tree inference accuracy and gene duplication rate (log10 transformed). The 4389 gene families with six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "44 bins". Each dot represents a bin. Number of duplications are estimated by Duptree.

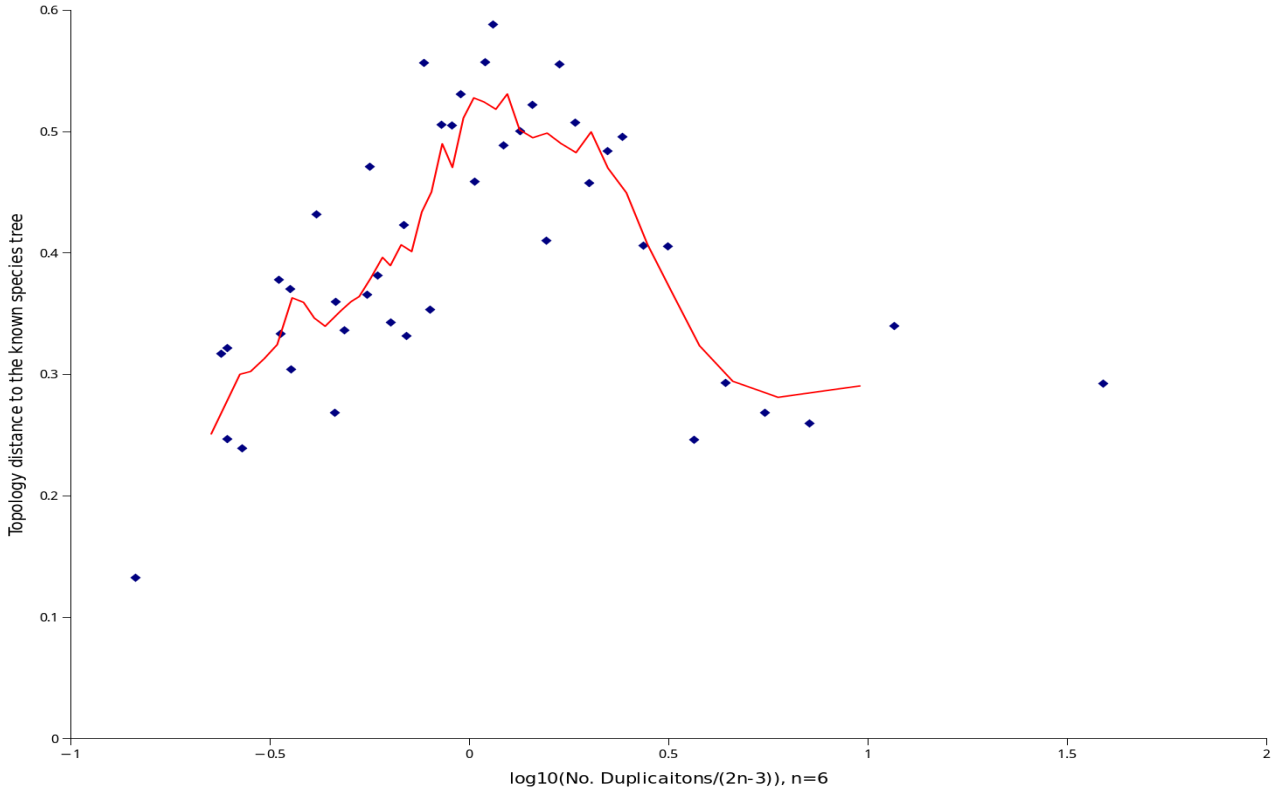


Figure 11. Relationship between average inferred species tree topology distance to the known species tree and gene duplication rate (log10 transformed). The 4389 gene families with six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "44 bins". Each dot represents a bin. Number of duplications are estimated by Duptree.

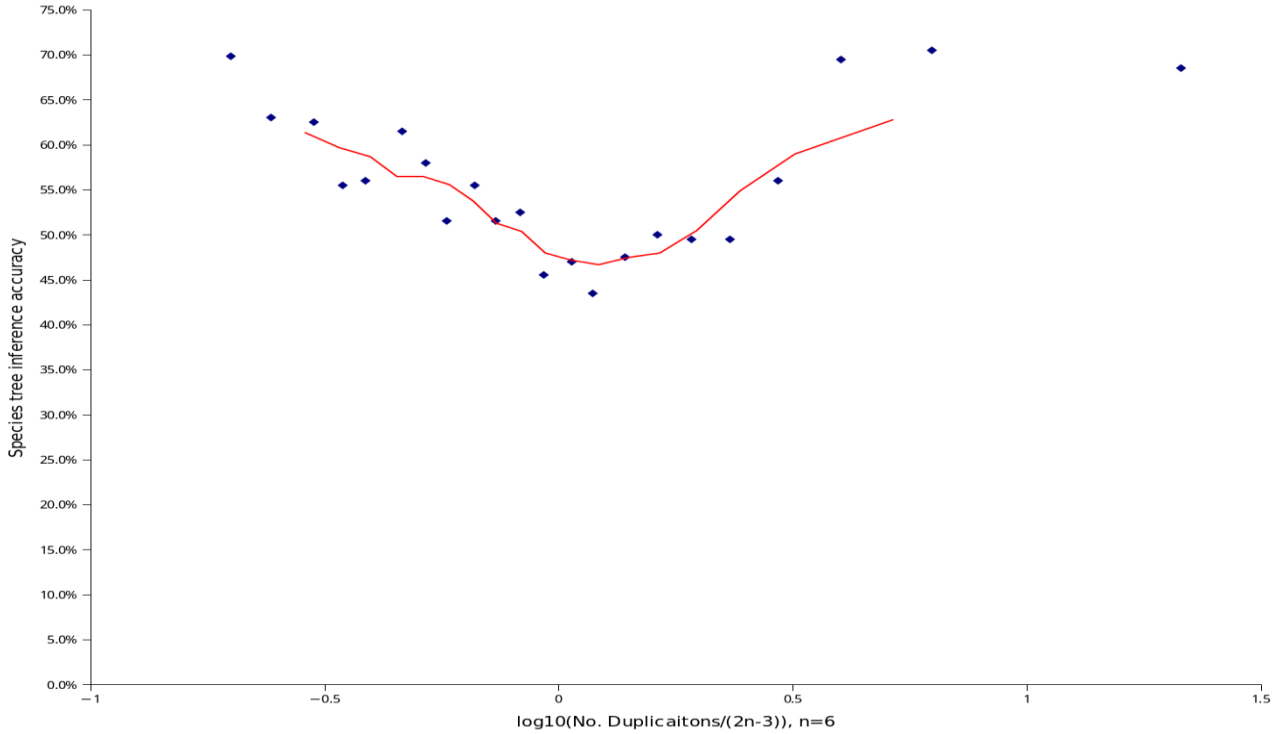


Figure 12. Relationship between species tree inference accuracy and gene duplication rate (log10 transformed). The 4389 gene families with six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "22 bins". Each dot represents a bin. Number of duplications are estimated by Duptree.

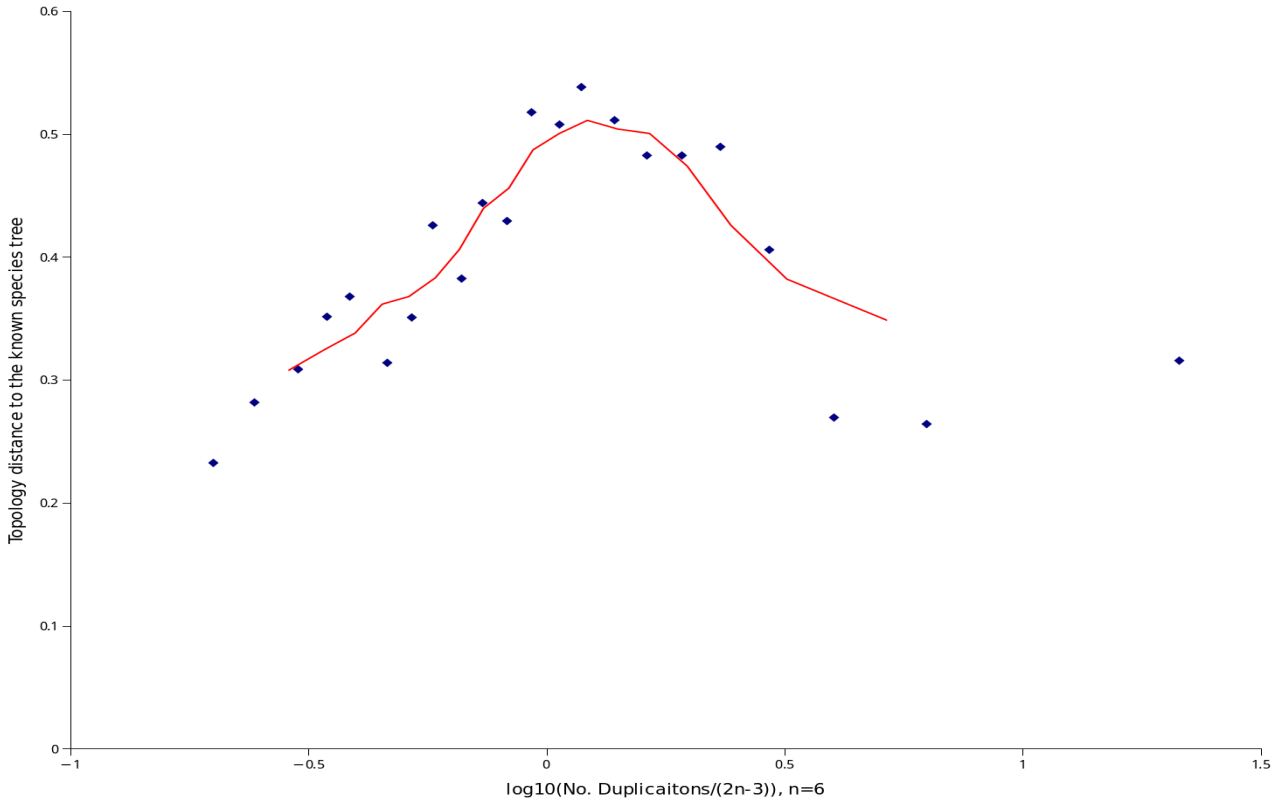


Figure 13. Relationship between average inferred species tree topology distance to the known species tree and gene duplication rate (log10 transformed). The 4389 gene families with six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "22 bins". Each dot represents a bin. Number of duplications are estimated by Duptree.

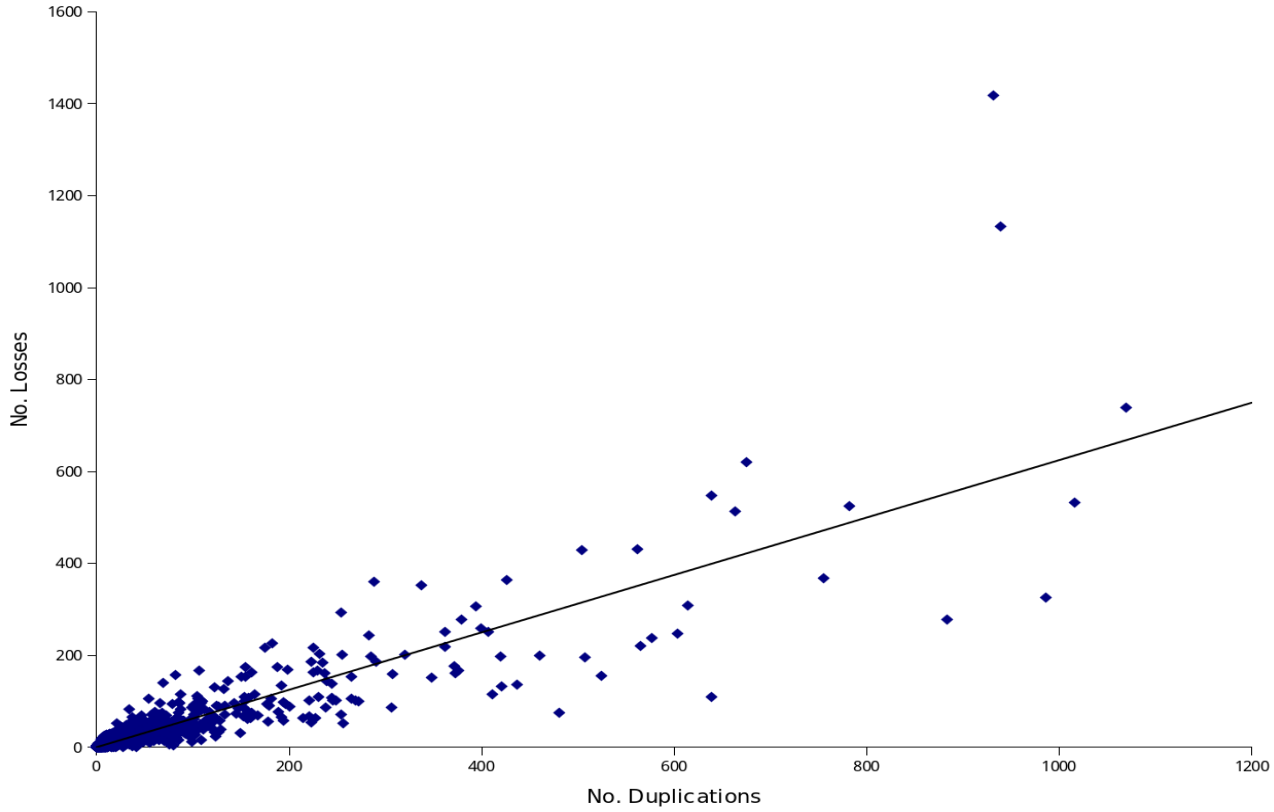


Figure 14. Relationship between number of duplications and number of losses. Each dot represents a gene family and a total of 4389 gene families of six taxa coverage are plotted here. R^2=0.763845, $p<0.01$. Number of duplications and number of losses are estimated by Notung 2.6.

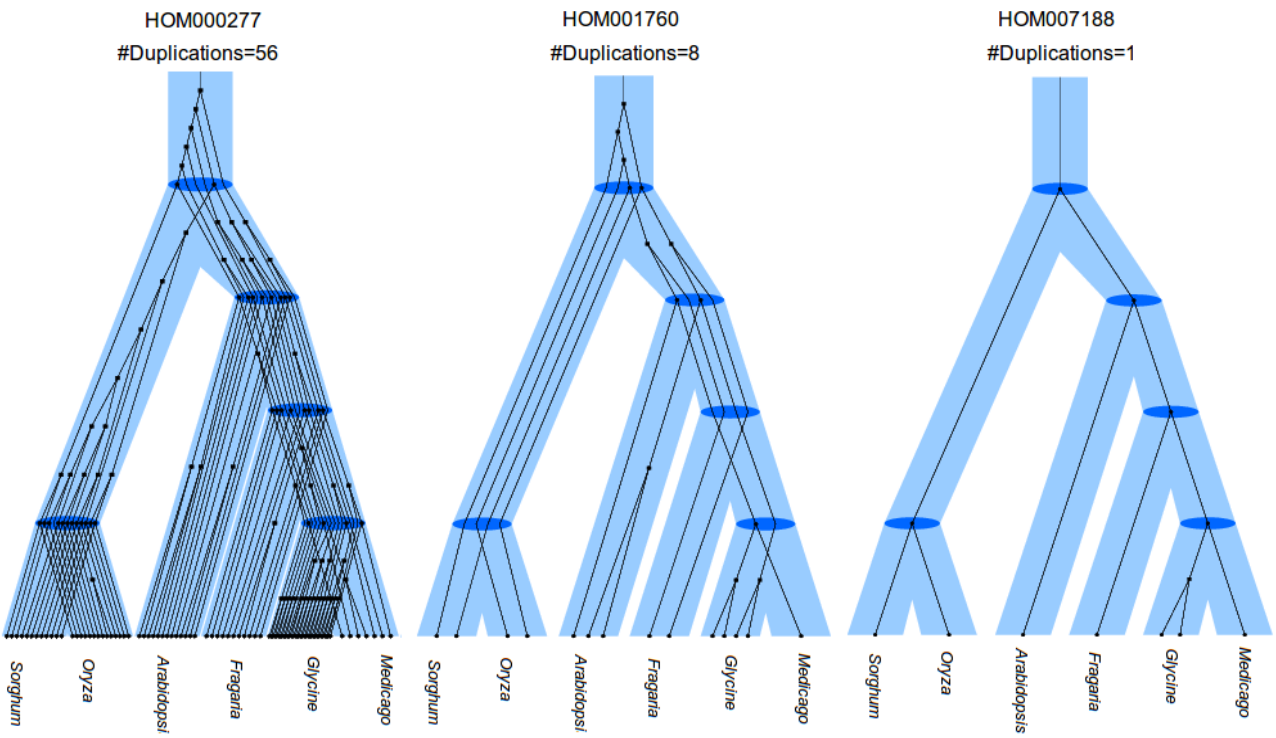
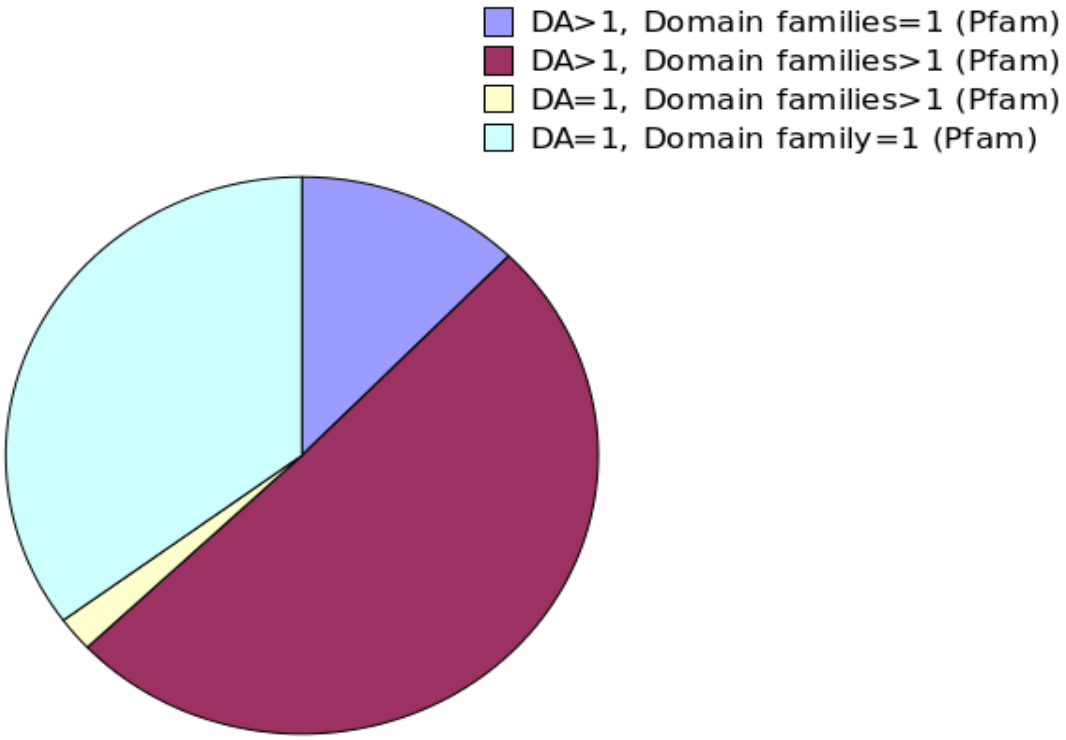


Figure 15. Three examples of gene tree—species tree reconciliations. Three gene families—"HOM00027", "HOM001760", "HOM007188" are of different levels of duplication rates. Gene tree were rooted and reconciled with known species tree with Notung 2.6. and visualized with PrIMETV(Sennblad et al., 2007). Speciation events are inferred as blue circles.



.

Figure 16. Pie chart of gene family distribution according to number of domain architecture (DA) and No. Domain families for the 3178 gene families with Pfam annotation of six taxa coverage.

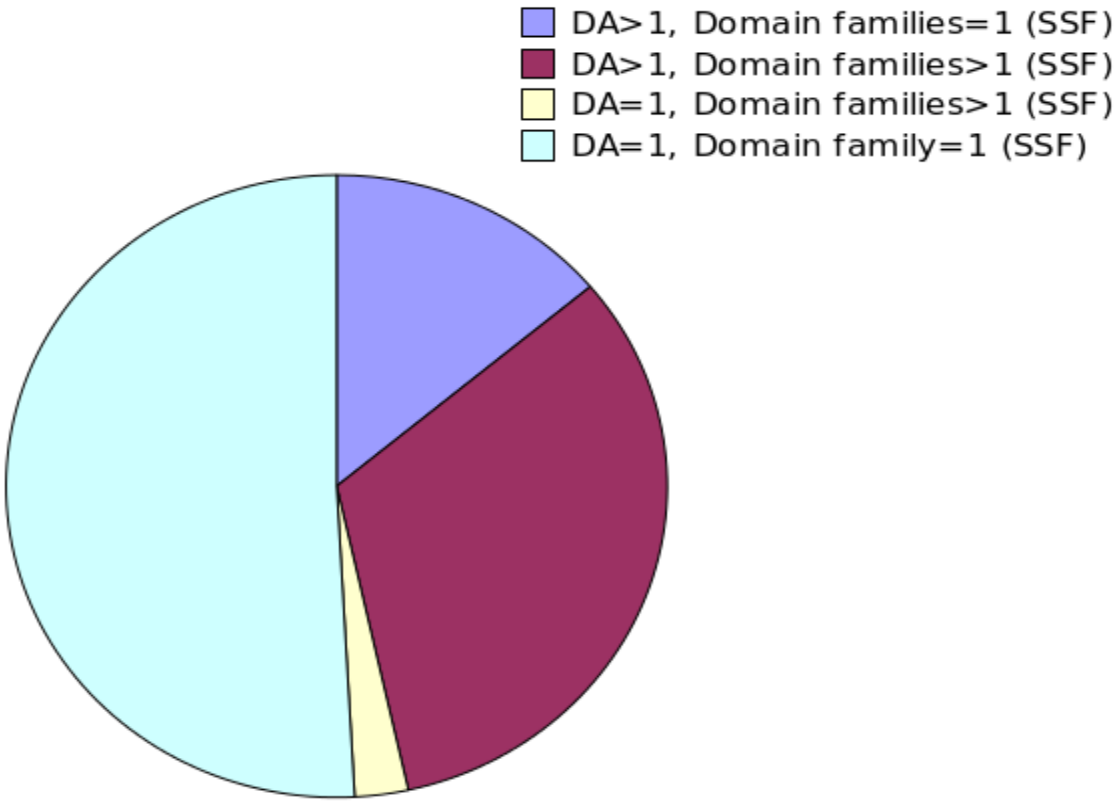


Figure 17. Pie chart of gene family distribution according to number of domain architecture (DA) and number of Domain families for the 1494 gene families with Superfamily annotation of six taxa coverage.

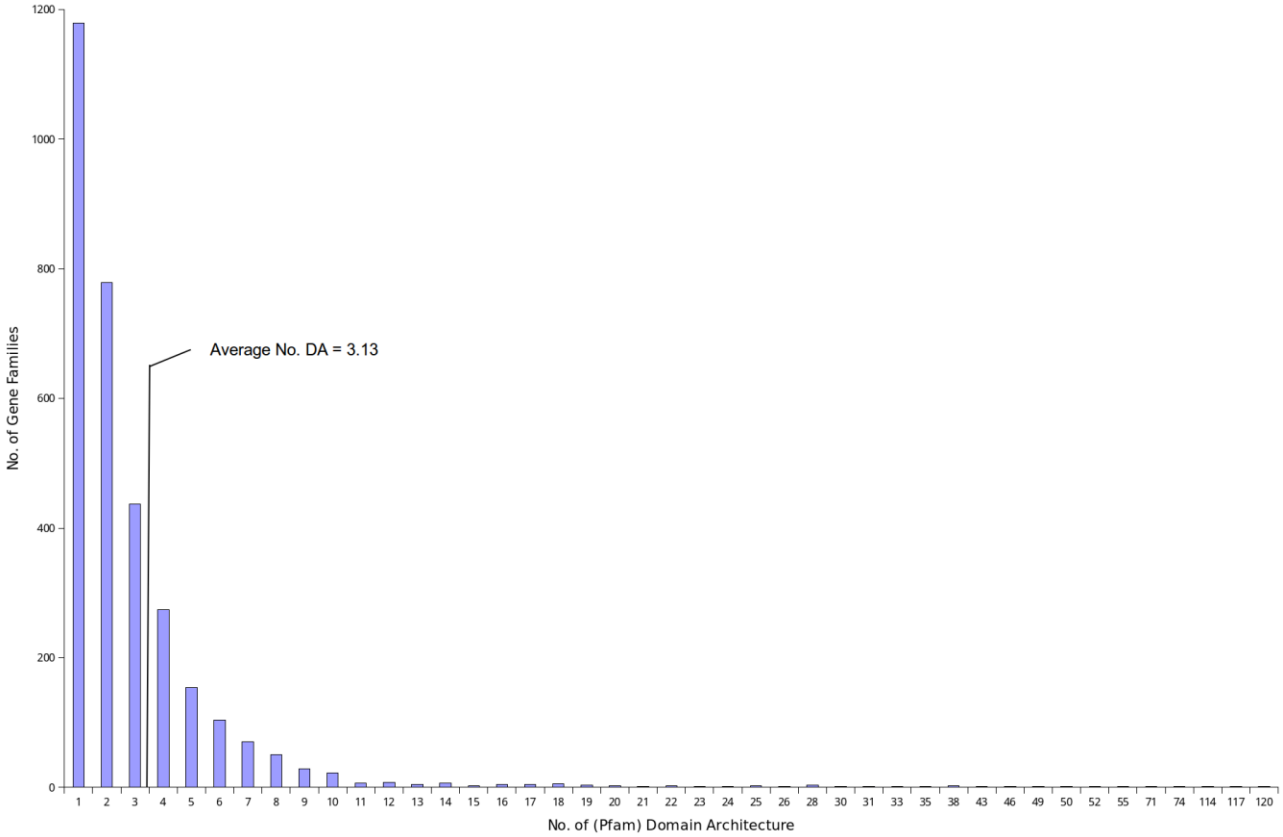


Figure 18. Gene family frequency distribution according to number of domain architecture (DA) for the 3178 gene families with Pfam annotation of six taxa coverage. Mean No. DA =3.13 , Median No. DA =2.

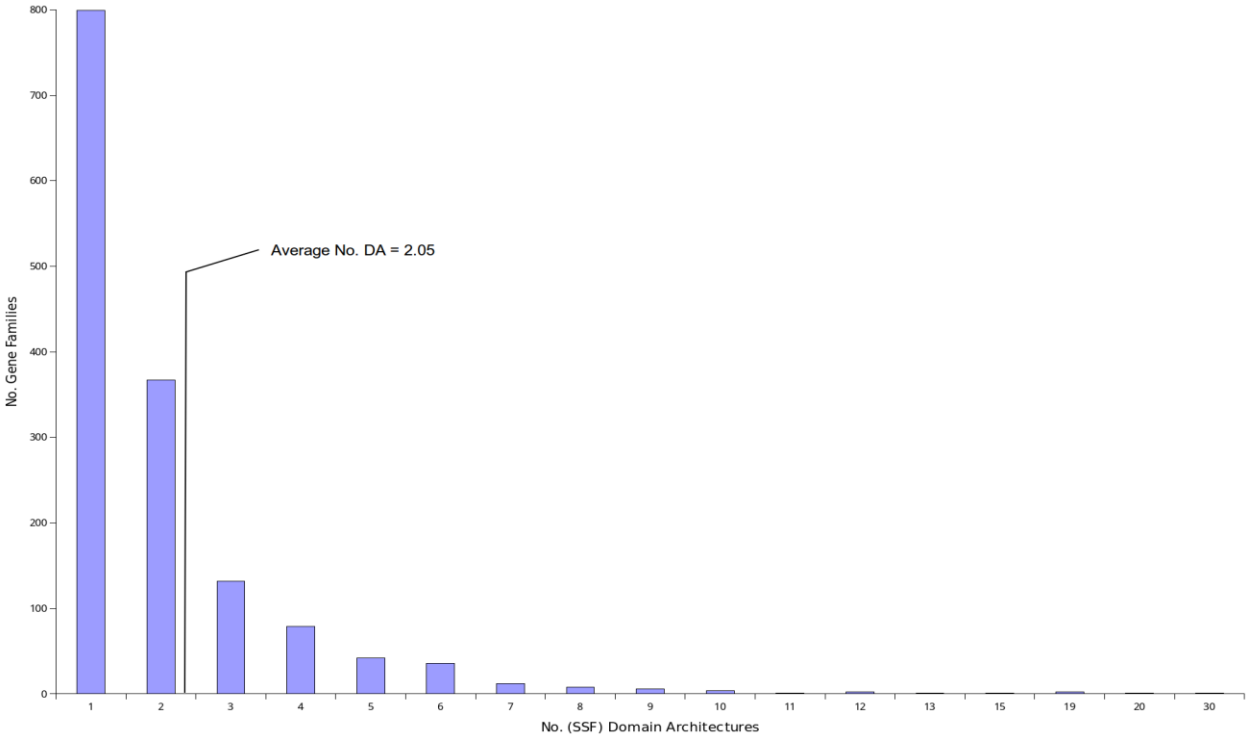


Figure 19. Frequency of domain architecture diversity of gene families. Gene family frequency distribution according to number of domain architecture (DA) for the 1494 gene families with Superfamily annotation of six taxa coverage. Mean number of DA =2.05 , Median number of DA =1.

.

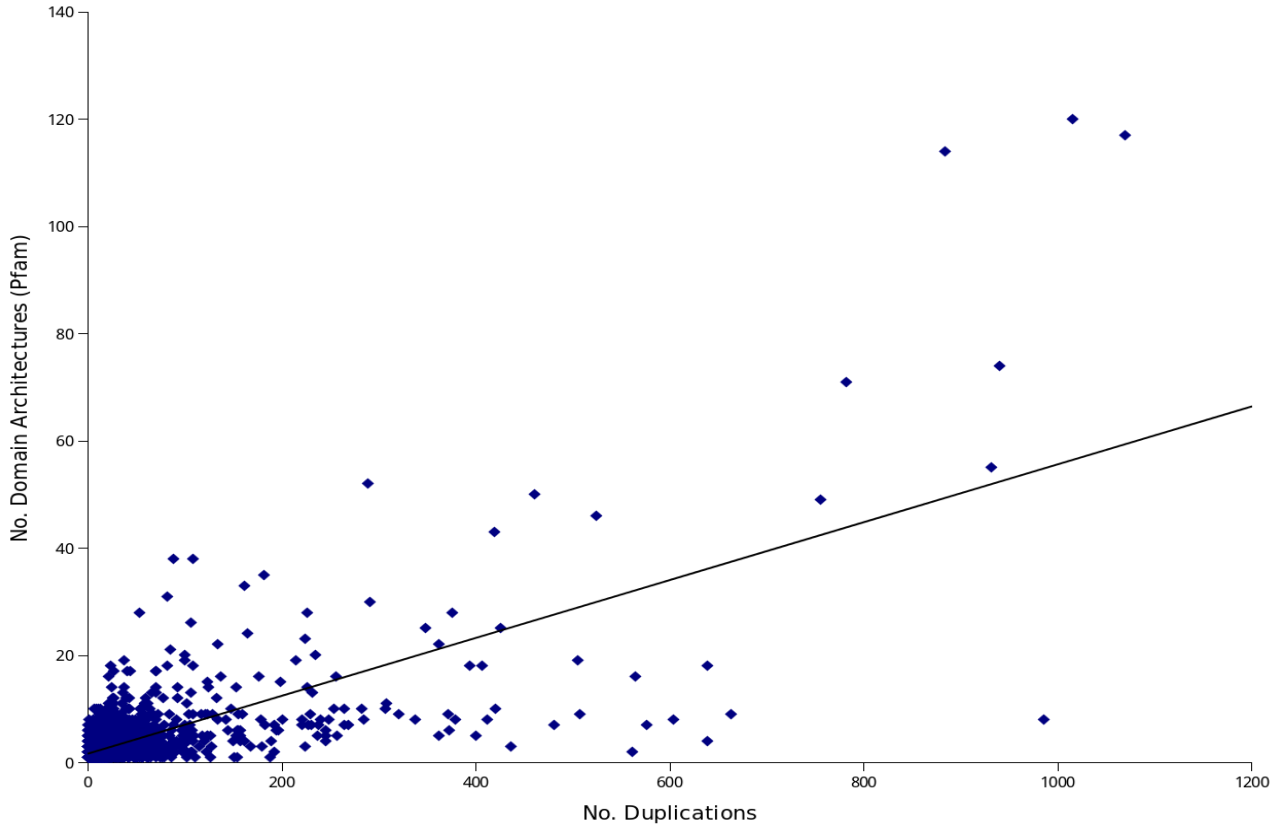


Figure 20. Relationship between number of duplications and number of domain architecture. Each dot represents a gene family and a total of 3178 gene families with Pfam annotation of six taxa coverage are plotted here. R^2=0.508457, $p<0.01$. Number of duplications are estimated by Notung 2.6.

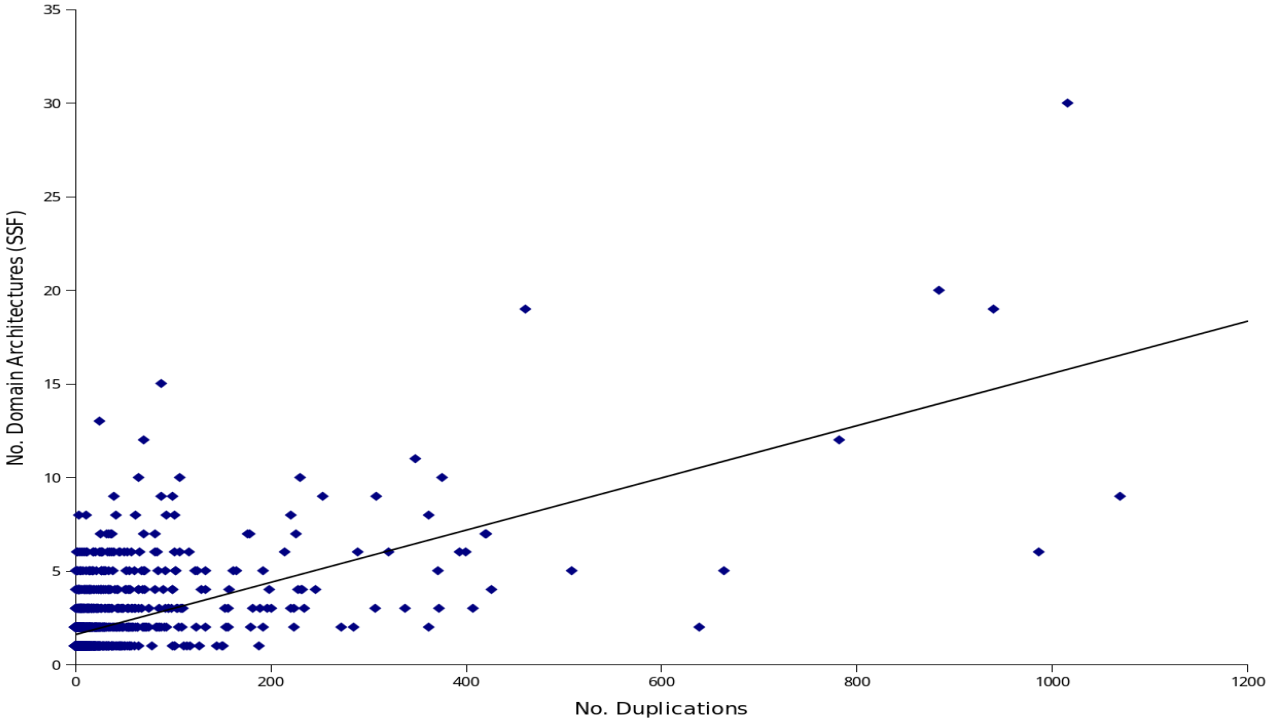


Figure 21. Relationship between number of duplications and number of domain architecture. Each dot represents a gene family and a total of 1494 gene families with Superfamily annotation of six taxa coverage are plotted here. R^2=0.357371, $p<0.01$. Number of duplications are estimated by Notung 2.6.

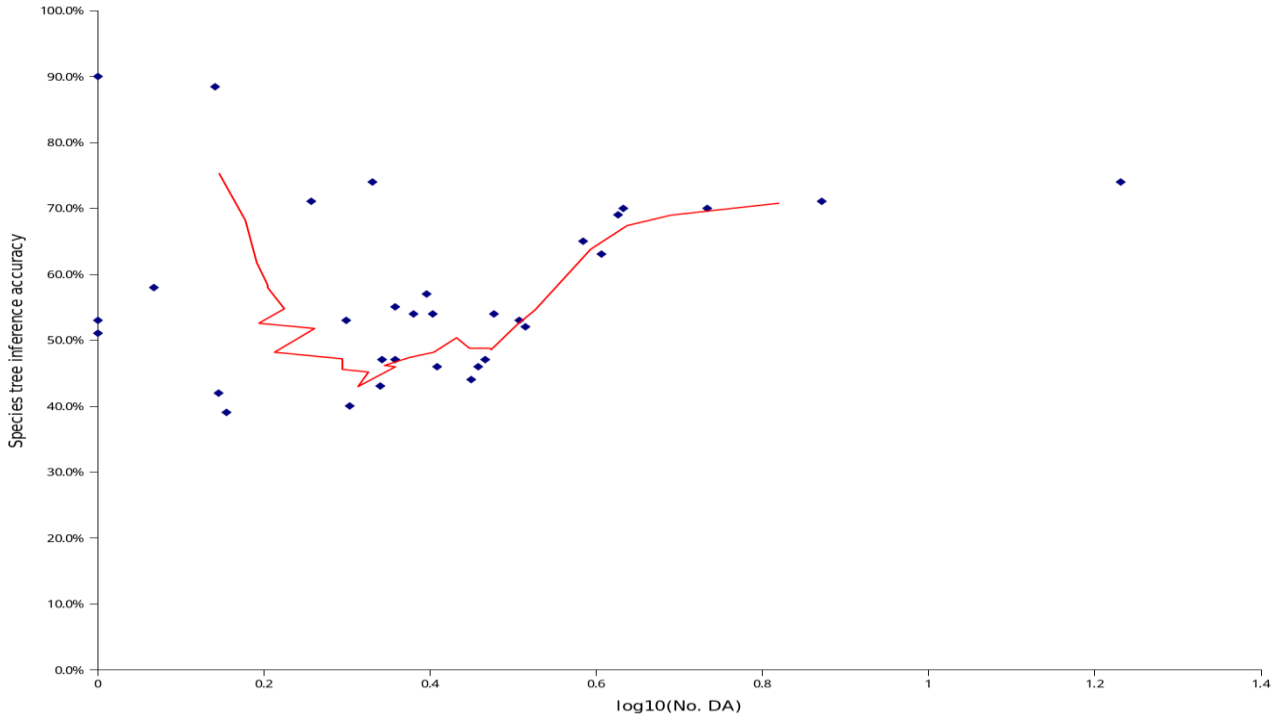


Figure 22. Relationship between species tree inference accuracy and number of domain architecture (log10 transformed). The 3178 gene families with Pfam annotation of six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "32 bins". Each dot represents a bin. Number of duplications are estimated by Notung 2.6. Average number of domain architecture for each bin is used.

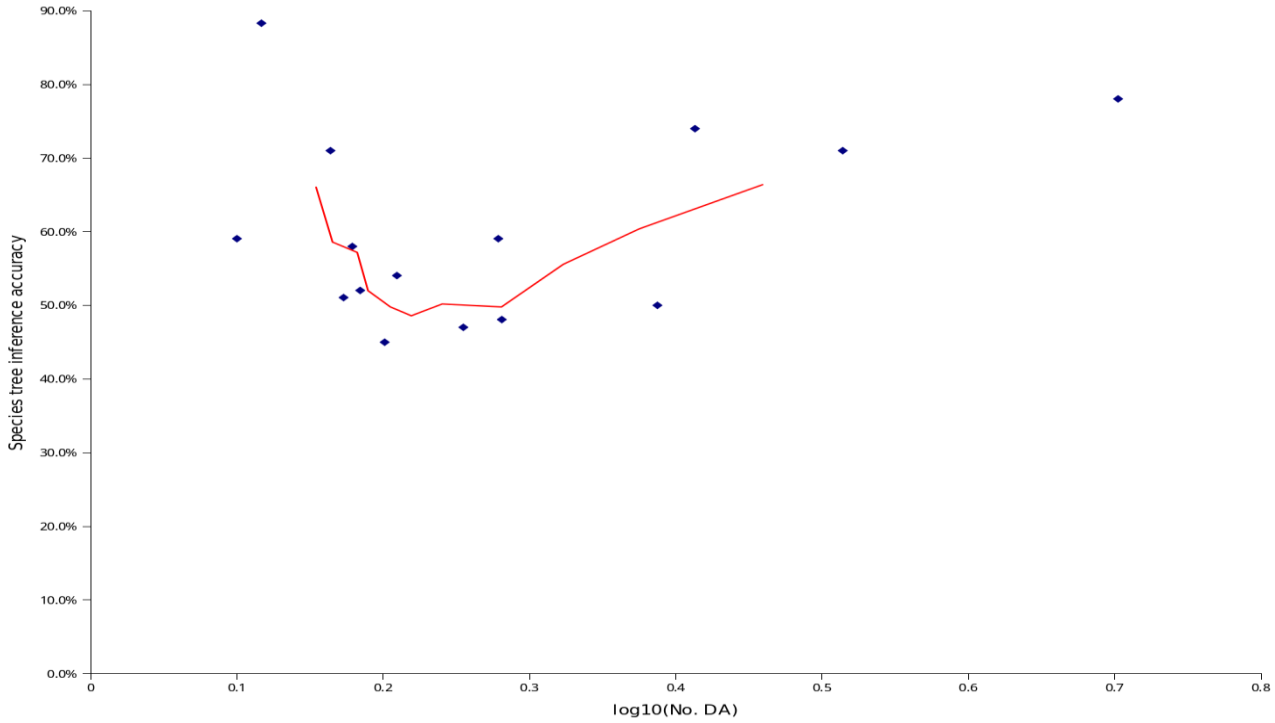


Figure 23. Relationship between species tree inference accuracy and number of domain architecture (log10 transformed). The 1494 gene families with Superfamily annotation of six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "15 bins". Each dot represents a bin. Number of duplications are estimated by Notung 2.6. Average number of domain architecture for each bin is used.

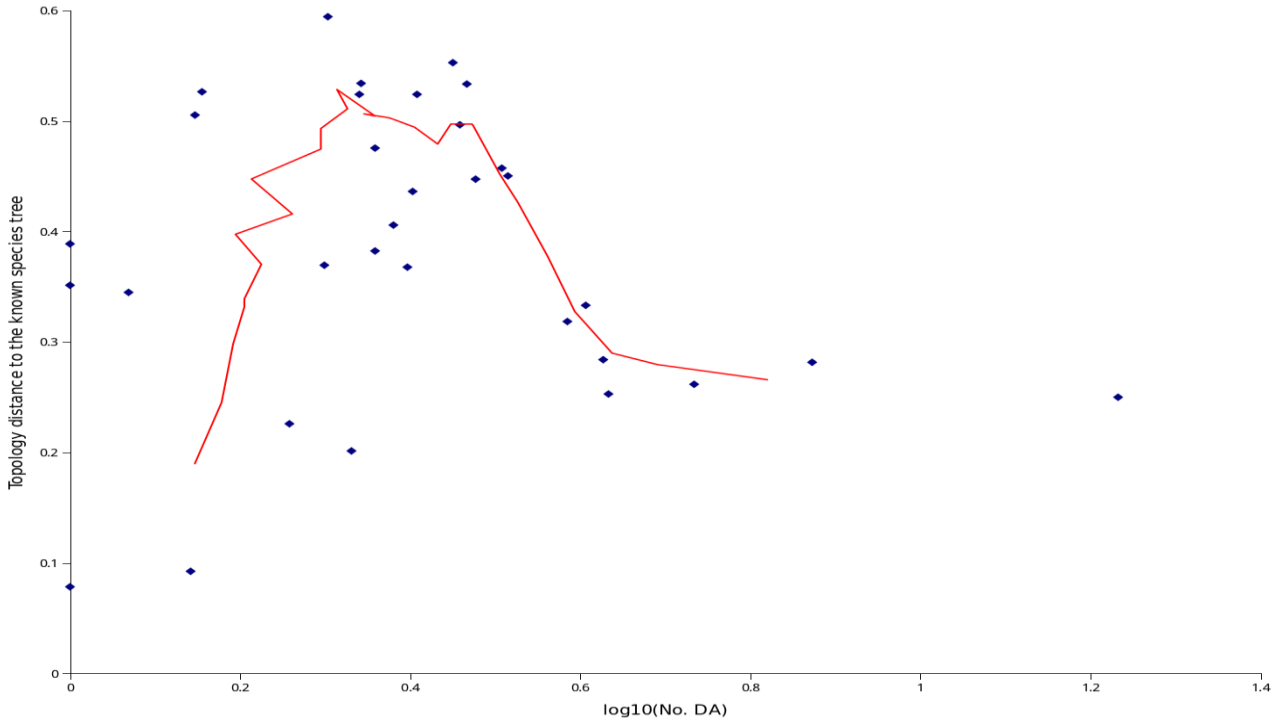


Figure 24. Relationship between average inferred species tree topology distance to the known species tree and gene duplication rate (log10 transformed). The 3178 gene families with Pfam annotation of six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "32 bins". Each dot represents a bin. Number of duplications are estimated by Notung 2.6. Average number of domain architecture for each bin is used.

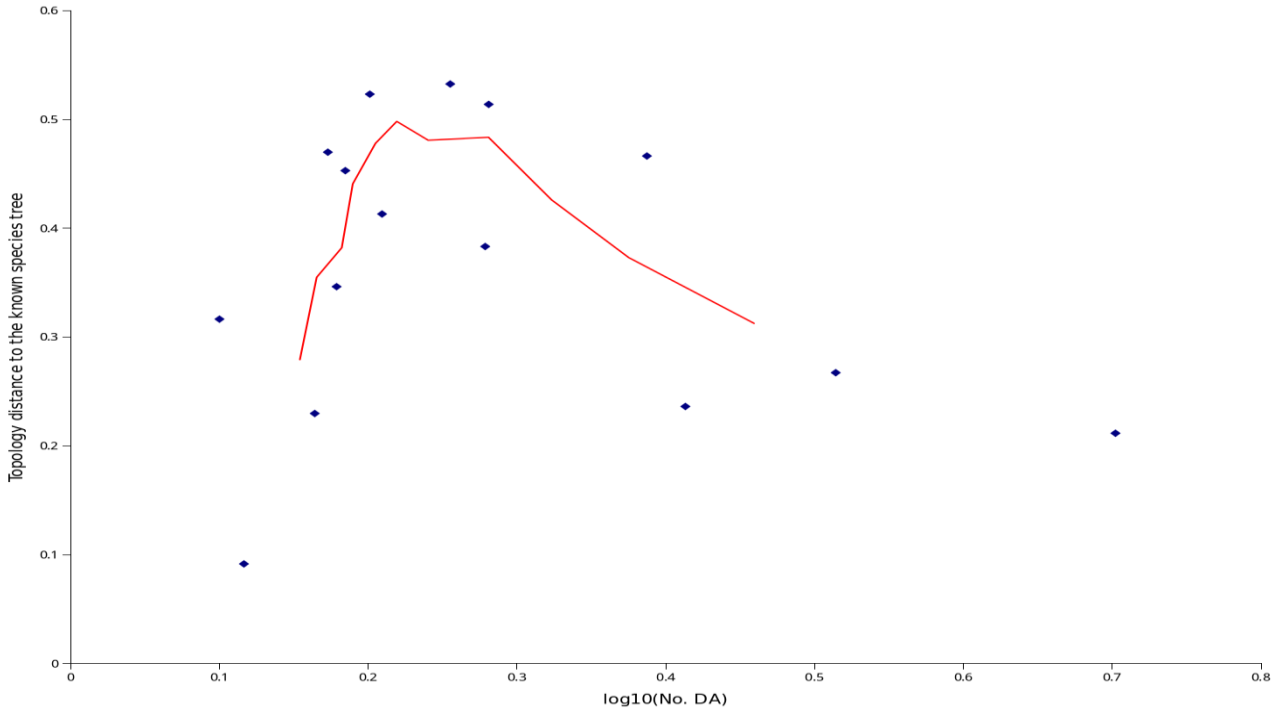


Figure 25. Relationship between average inferred species tree topology distance to the known species tree and gene duplication rate (log10 transformed). The 1494 gene families with Superfamily annotation of six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "15 bins". Each dot represents a bin. Number of duplications are estimated by Notung 2.6. Average number of domain architecture for each bin is used.

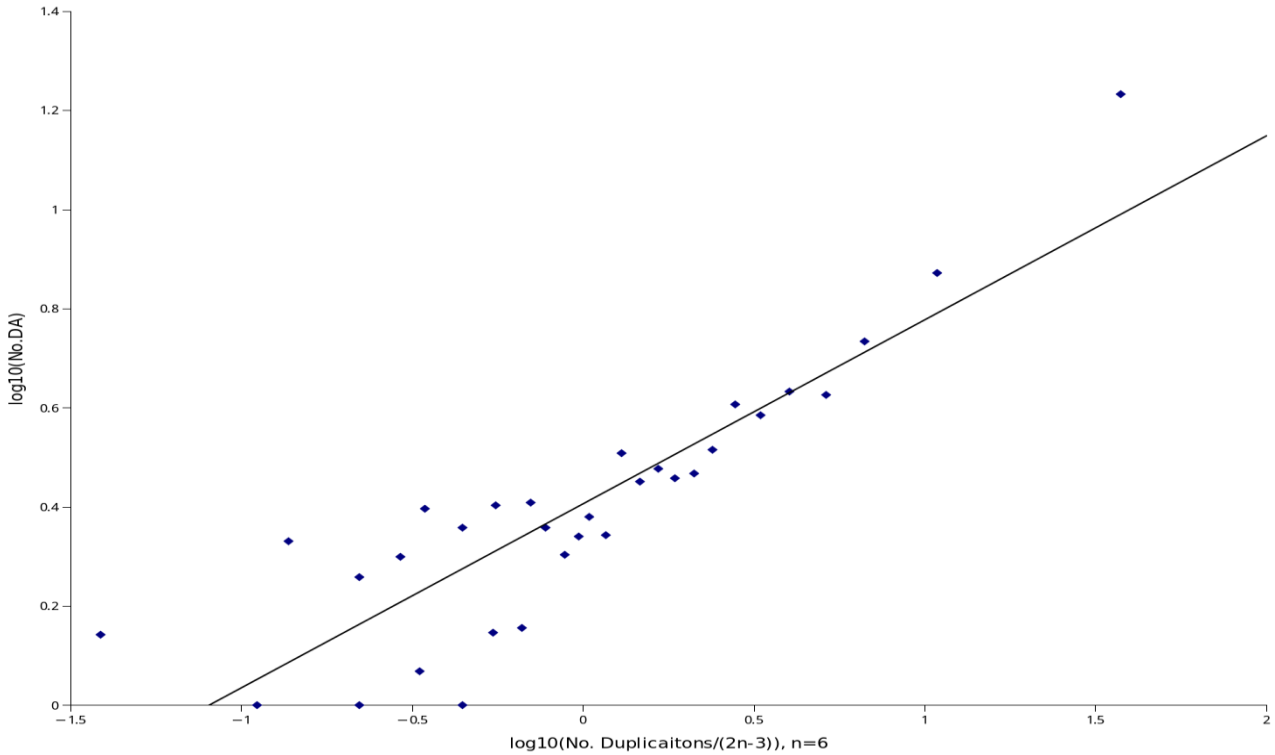


Figure 26. Relationship betwee gene duplicationa rate (log10 transformed) and number of domain architecture (log transformed). The 3178 gene families with Pfam annotation of six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "32 bins". Each dot represents a bin. Number of duplications are estimated by Notung 2.6. Average number of domain architecture for each bin is used.

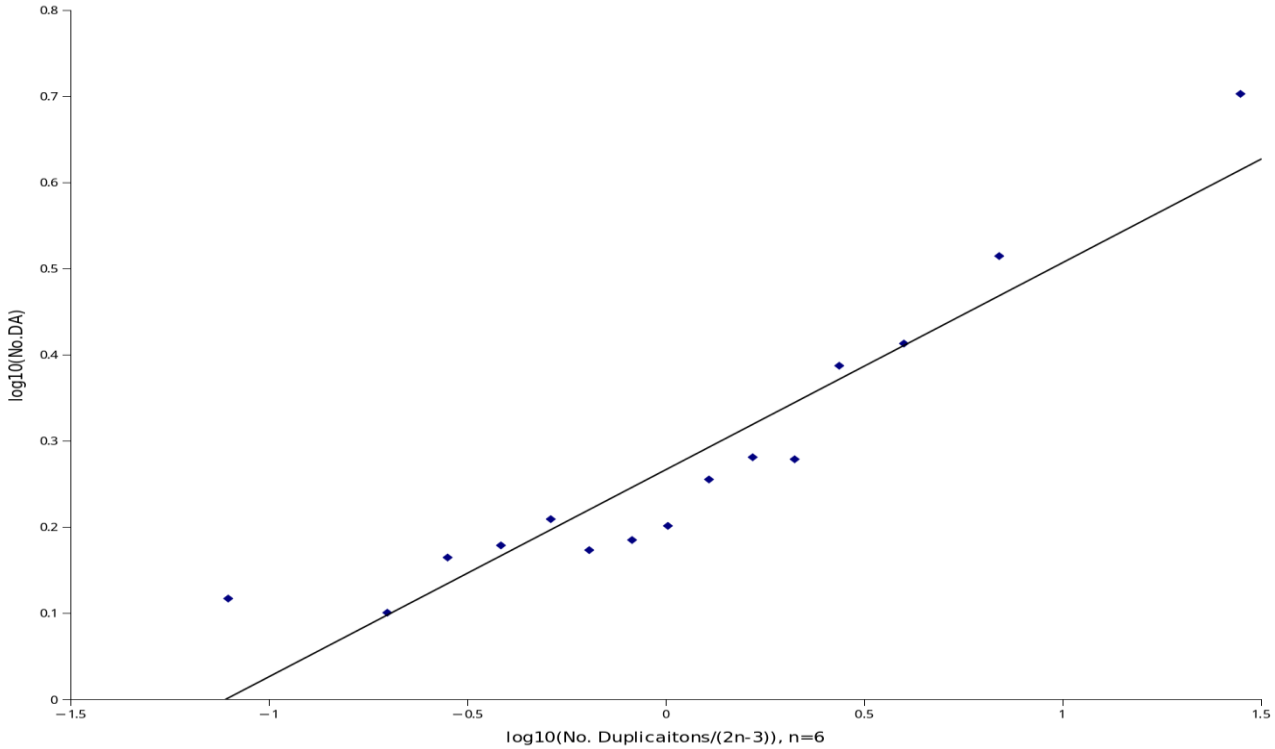


Figure 27. Relationship between gene duplication rate (log10 transformed) and number of domain architecture (log transformed). The 1494 gene families with Superfamily annotation of six taxa coverage are sorted according to the "descending order of number of duplications", then grouped into "32 bins". Each dot represents a bin. Number of duplications are estimated by Notung 2.6. Average number of domain architecture for each bin is used.

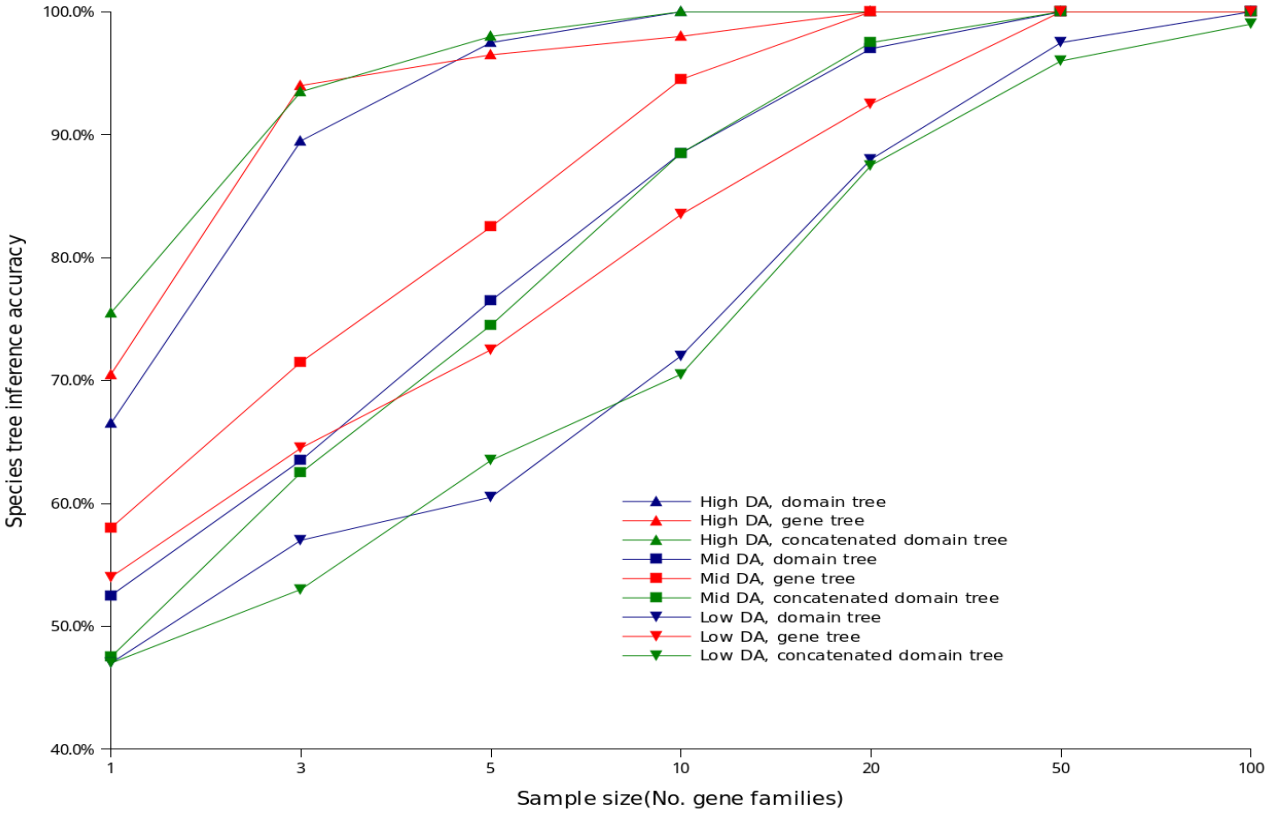


Figure 28. Accuracy of species tree inference from gene tree, domain tree, concatenated domain tree under different sample sizes (number of gene families) and different Domain Architecture diversity levels (High, Mid, Low). Domains are annotated by Pfam.

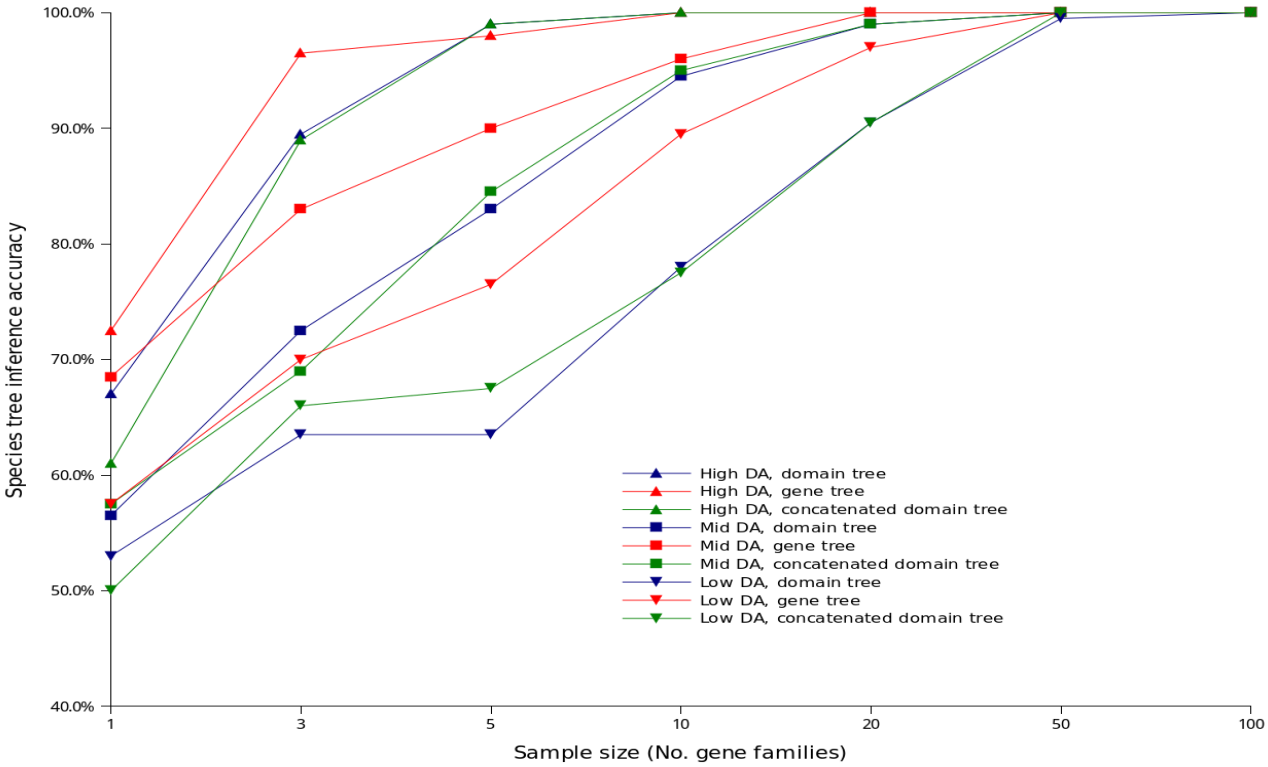


Figure 29. Accuracy of species trees inference from gene tree, domain tree, concatenated domain tree under different numbers of sample size (number of gene families) and different Domain Architecture diversity levels (High, Mid, Low). Domains are annotated by Superfamily.

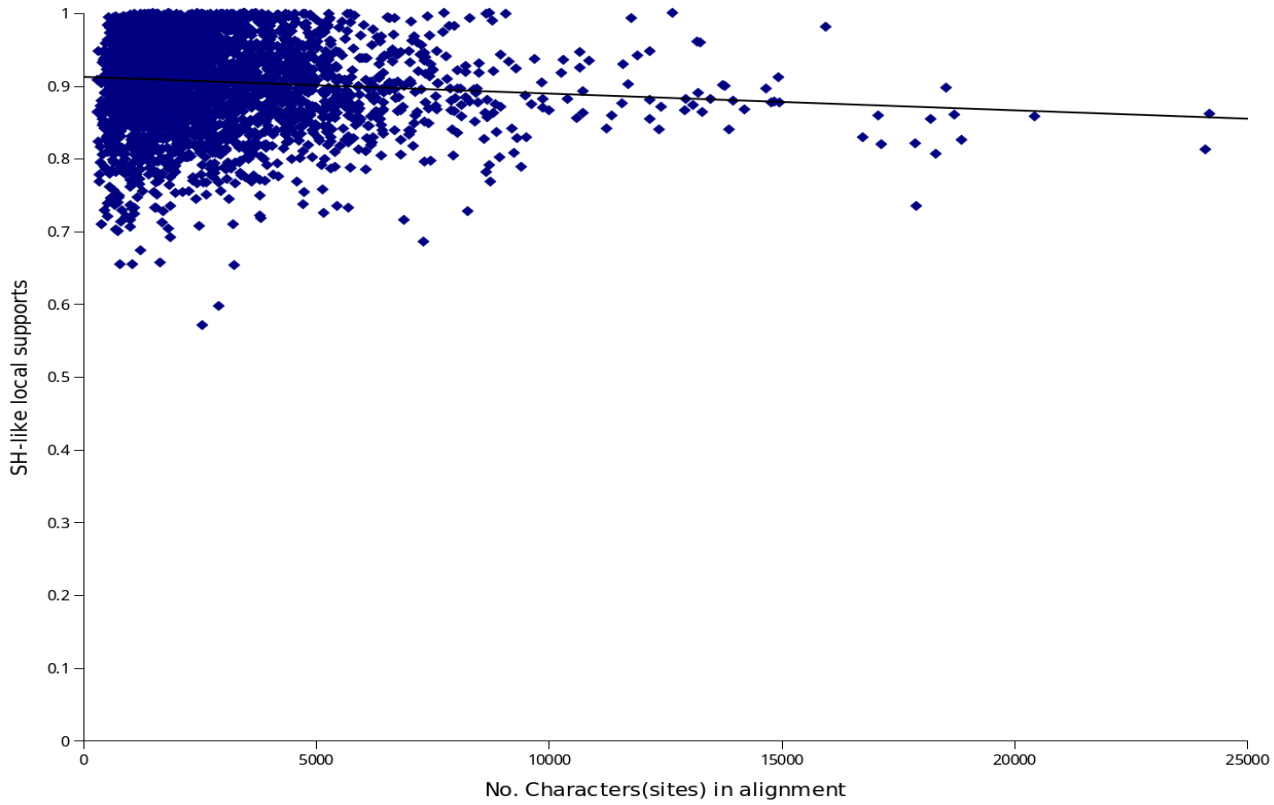


Figure 30. Relationship between SH-like local supports and number of characters(sites) in alignment. Each dot represents a gene family and a total of 3178 gene families with Pfam annotation of six taxa coverage are plotted here. R^2=0.00730, $p<0.01$.

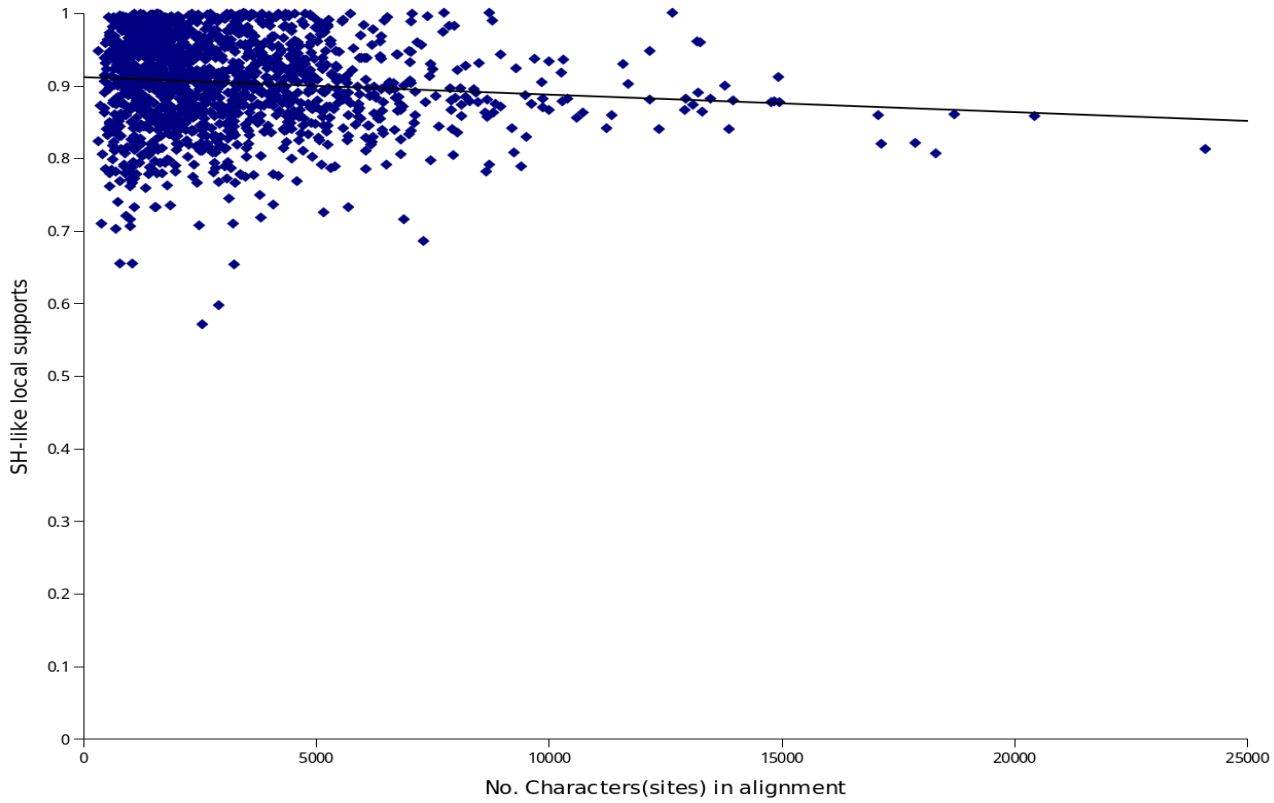


Figure 31. Relationship between SH-like local supports and number of characters(sites) in alignment. Each dot represents a gene family and a total of 1494 gene families with Superfamily annotation of six taxa coverage are plotted here. R^2=0.00953, $p<0.01$.

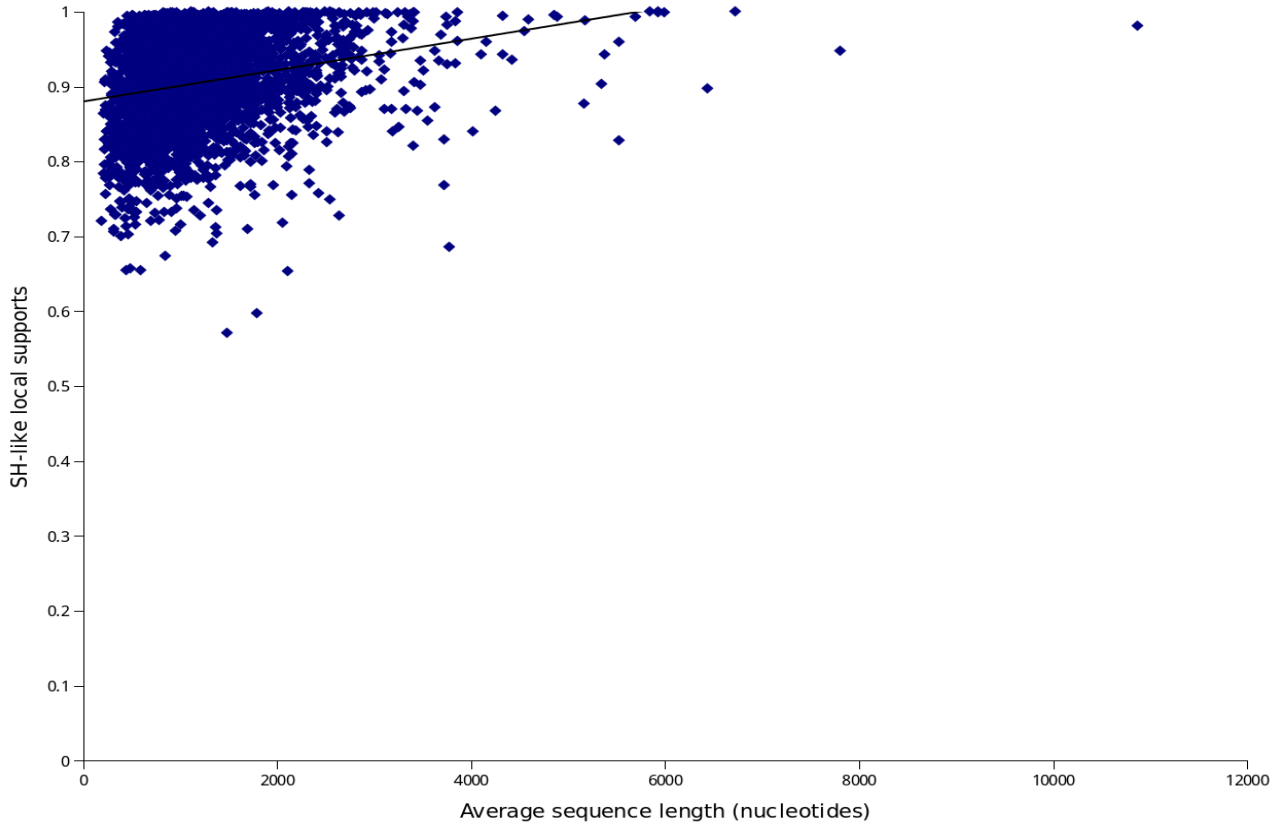


Figure 32. Relationship between SH-like local supports and average sequence length(nucleotides). Each dot represents a gene family and a total of 3178 gene families with Pfam annotation of six taxa coverage are plotted here. R^2=0.05982, $p<0.01$.

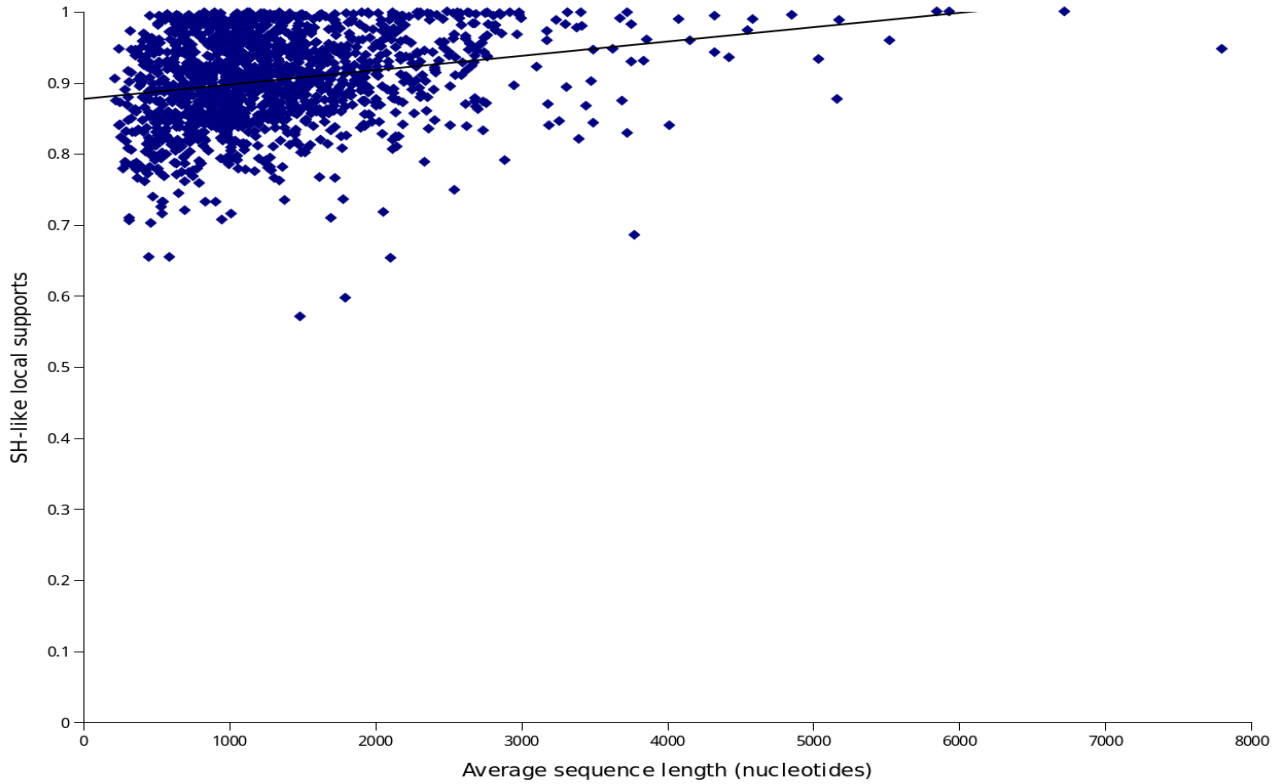


Figure 33. Relationship between SH-like local supports and average sequence length(nucleotides). Each dot represents a gene family and a total of 1494 gene families with Pfam annotation of six taxa coverage are plotted here. R^2=0.06095, $p<0.01$.

REFERENCES

Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acid Res. 2010;38:W7-W13.

Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comp. Biol. 2009;5:e1000262.

Anfinsen CB, Haber E, Sela M, White Jr, FH.The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proceedings of the National Academy of Sciences of the United States of America.1961; 47(9), 1309.

APG III [Angiosperm Phylogeny Group III]. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Botanical Journal of the Linnean Society.2009; 161: 105–121.

Baker WJ, Savolainen V, Asmussen-Lange CB, Chase MW, Dransfield J, Forest F, Harley MM, Uhl NW, Wilkinson M. Complete generic-level phylogenetic analyses of palms (Arecaceae) with comparisons of supertree and supermatrix approaches. Syst Biol. 2009;58:240-256.

Bansal MS, Burleigh JG, Eulenstein O, Wehe A. Heuristics for the gene-duplication problem: a Θ(n) speed-up for the local search. RECOMB.2007; LNCS 2007;4453:238-252.

Bansal MS, Burleigh JG, Eulenstein O. Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models. BMC Bioinformatics 2010;11:S42.

Bapteste E, Boucher Y, Leigh J, Doolittle WF. Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol. 2004;12:406-411.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy S, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL. The Pfam protein families database. Nucleic Acids Res. 2002;30:276-280.

Bininda-Emonds O, Sanderson MJ. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. Syst. Biol. 2001;50:565-579.

Blackburne BP, Whelan S. Class of multiple sequence alignment algorithm affects genomic analysis. Molecular Biology and Evolution.2013;30(3), 642-653.

Buerki S, Forest F, Salamin N, Alvarez N. Comparative performance of supertree algorithms in large data sets using the soapberry family (Sapindaceae) as a case study. Systematic Biology 2011;60:32-44.

Burleigh JG, Driskell AC, Sanderson MJ. Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. Syst. Biol. 2006;55:426-440.

Burleigh JG, Bansal MS, Eulenstein O, Hartmann S, Wehe A, Vision TJ. Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees. Syst. Biol. 2011;60:117-125.

Chen F, Mackey AJ, Vermunt JK, Roos DS. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PloS one 2007;2(4):e383.

Chen K, Durand D, Farach-Colton M. Notung: a program for dating gene duplications and optimizing gene family trees. J. Comput. Biol. 2000;7:429-447.

Cohen-Gihon I, et al. Evolution of domain promiscuity in eukaryotic genomes—a perspective from the inferred ancestral domain architectures. Mol Biosyst. 2011;7:784-792.

Cotton JA, Page RDM. Going nuclear: gene family evolution and vertebrate phylogeny reconciled. P. Roy. Soc. Lond. B Biol. 2002;269:1555-1561.

Cotton JA, Page RDM. Reconciled trees for supertree construction. Pages 107-125 in Phylogenetic supertrees: combining information to reveal the tree of life. Dordrecht (The Netherlands): Kluwer Academic Press; 2004.

Cranston C, Hurwitz B, Ware D, Stein L, Wing RA. Species trees from highly incongruent gene trees in rice. Syst. Biol. 2009;58:489-500.

Dalquen DA, Altenhoff AM, Gonnet GH, Dessimoz C. The Impact of Gene Duplication, Insertion, Deletion, Lateral Gene Transfer and Sequencing Error on Orthology Inference: A Simulation Study. PloS one.2013; 8(2), e56925.

de Queiroz A, Gatesy J. The supermatrix approach to systematics. Trends Ecol. Evol. 2007;22:34-41.

Dessimoz C, Gabaldón T, Roos DS, Sonnhammer ELL, Herrero J. Toward community standards in the quest for orthologs. Bioinformatics, 2012;28:900-904.

Dwivedi B, Gadagkar SR. Phylogenetic inference under varying proportions of indel-induced alignment gaps. BMC Evol Biol. 2009;9:211.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 2004; 32(5), 1792-1797.

Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nuc. Acids Res. 2002;30:1575-1584.

Fukami-Kobayashi K, Minezaki Y, Tateno Y, Nishikawa K. A tree of life based on protein domain organizations. Mol Biol Evol 2007;24:1181-1189.

Gingerich DJ, Hanada K, Shiu SH, Vierstra RD. Large-scale, lineage-specific expansion of a bric-a-brac/tramtrack/broad complex ubiquitin-ligase gene family in rice. The Plant cell 2007, 19(8):2329-2348.

Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed by globin sequences. Syst. Zool 1979;28:132-163.

Górecki P, Burleigh JG, Eulenstein O. GTP supertrees from unrooted gene trees: linear time algorithms for NNI based local searches. In Bioinformatics Research and Applications. 2012; 102-114.

Guigó R, Muchnik I, Smith TF. Reconstruction of ancient molecular phylogeny. Mol. Phylogenet. Evol. 1996;6:189-2

Hilu KW, Borsch T, Müller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R, Sauquet H, Neinhuis C, Slotta TAB, Rohwer JG, Campbell CS, Chatrou LW. Angiosperm phylogeny based on matK sequence information. Am. J. Bot 2003;90:1758-1776.

Hunter S, Apweiler R, Attwood TK, et al.(38 co-authors). InterPro: the integrative protein signature database. Nucleic Acids Res. 2009;37:D211-D215.

Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models.Nat. Rev. Genet. 2010. 11: 97-108.

Jackson SE. How do small single-domain proteins fold?. Folding and Design. 1998. 3(4), R81-R91.

Jansen RK, Raubeson LA, Boore JL et al. (15 co-authors). Methods for obtaining and analyzing whole chloroplast genome sequences. Methods Enzymol. 2005; 395:348–384

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, dePamphilis CW. Ancestral polyploidy in seed plants and angiosperms. Nature 2011;473:97-100.

Jordan G, Goldman N. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol Biol Evol. 2012;29:1125-1139.

Katz LA, Grant JR, Parfrey LW, Burleigh JG. Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life. Syst Biol. 2012;61:653-660.

Kersting AR, Bauer EB, Moore AD, Grath S: Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. Genome Biology and Evolution, 2012; 4(3), 316-329.

Kumar S, Filipski A. Multiple sequence alignment: In pursuit of homologous DNA positions. Genome Res. 2007;17:127-135.

Kuzniar A, van Ham RCHJ, Pongor S, Leunissen JAM. The quest for orthologs: finding the corresponding gene across genomes. Trends Genet. 2008;24:539-551.

Li L, Stoeckert CJ Jr., Roos DS.OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178-2189.

Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. Genome Res 2008;18:298-309.

Maddison WP. Gene trees in species trees. Syst. Biol. 1997;46:523-536.

Martin AP, Burg TM. Perils of paralogy: using HSP70 genes for inferring organismal phylogeny. Syst. Biol. 2002;51:570-587.

McCormack JE, Huang H, Knowles LL. Maximum-likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. Syst. Biol. 2009;58:501-508.

McGowen MR, Clark C, Gatesy J. The vestigial olfactory receptor subgenome of odontocete whales: phylogenetic congruence between gene-tree reconciliation and supermatrix methods. Syst. Biol. 2008;57:574-590.

McMahon MM, Sanderson MJ. Phylogenetic supermatrix analysis of Genbank sequences from 2228 papilionoid legumes. Syst. Biol. 2006;55:818-836.

Ness RW, Graham SW, Barrett SC. Reconciling gene and genome duplication events: Using multiple nuclear gene families to infer the phylogeny of the aquatic plant family Pontederiaceae. Mol. Biol. Evol. 2011;28:3009-3018.

Page RDM, Charleston MA. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol. Phylogenet. Evol. 1997;7:231-240.

Page RDM, Cotton JA. Vertebrate phylogenomics: reconciled trees and gene duplication. Pac. Symp. Biocomput 2002;7:536-547.

Page RDM. Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. Mol. Phylogenet. Evol. 2000;14:89-106.

Pamilo P, Nei M. Relationships between gene trees and species trees. Mol. Biol. Evol. 1988;5:568-583.

Phuong TM, Do CB, Edgar RC, Batzoglou S. Multiple alignment of protein sequences with repeats and rearrangements. Nucleic Acids Res. 2006;34:5932-5942.

Price M, Dehal P, Arkin A. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 2010;5:e9490.

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. Plant Cell. 2009;21:3718-3731.

Puigbo P, Garcia-Vallve S, McInerney JO. TOPD/FMTS: a new software to compare phylogenetic trees. Bioinformatics 2007;23:1556-1558.

Rokas A, Williams B, King N, Carroll S. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 2003;425:798-804.

Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature.2013.

Sanderson MJ, McMahon MM. Inferring angiosperm phylogeny from EST data with widespread gene duplication. BMC Evol. Biol. 2007;7:S3.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature, 2006;440(7082), 341-345.

Semon M, Wolfe KH. Reciprocal gene loss between *Tetraodon* and zebrafish after whole genome duplication in their ancestor. Trends in Genetics, 2007;23(3), 108-112.

Shiu S-H, Karlowski WM, Pan R, Tzeng YH, Mayer KF, Li W-H. Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. Plant Cell 2004;16:1220-1234.

Slowinski JB, Knight A, Rooney AP. Inferring species trees from gene trees: a phylogenetic analysis of the Elapidae (serpentes) based on the amino acid sequences of venom proteins. Mol. Phylogenet. Evol. 1997;8:349-362.

Slowinski JB, Page RDM. How should species phylogenies be inferred from sequence data? Syst. Biol. 1999;48:814-825.

Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker JB, Moore MJ, Carlsward BS, Bell CD, Latvis M, Crawley S, Black C, Diouf D, Xi Z, Rushworth CA, Gitzendanner MA, Sytsma KJ, Qiu YL, Hilu KW, Davis CC, Sanderson MJ, Beaman RS, Olmstead RG, Judd WS, Donoghue MJ, Soltis PS. Angiosperm phylogeny: 17 genes, 640 taxa. Am. J. Bot. 2011;98:704-730.

Sennblad B, Schreil E, Sonnhammer ACB, Lagergren J, Arvestad L. Primetv: a viewer for reconciled trees. BMC Bioinformatics 2007;8(1); 148.

Stamatakis A. RAxML-VI-HPC: maximum likelihood phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 2006;22:2688-2690.

Steel M, Linz S, Huson DH, Sanderson MJ. Identifying a species tree subject to random lateral gene transfer. Journal of Theoretical Biology, 2013;322, 81-93.

Wang M, Caetano-Anollés G. Global phylogeny determined by the combination of protein domains in proteomes. Mol Biol Evol 2006;23:2444-2454.

Wehe A, Bansal MS, Burleigh JG, Eulenstein O. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics 2008;24:1540-1541.

Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J. SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. 2009;37:D380-D386.

Wolf YI, Rogozin IB, Grishin NV, Koonin EV. Genome trees and the tree of life. Trends Genet 2002;18:472-479.

Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Science 2008;319:473-476.

Xu G, Ma H, Nei M, Kong H .Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. Proc Natl Acad Sci USA.2009; 106: 835–840.

Yang S, Doolittle RF, Bourne PE. Phylogeny determined by protein domain content. Proc Natl Acad Sci USA 2005;102:373-378.

Yang Z. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 1998;47:125-133.

Yoder JB, Briskine R, Mudge J, Farmer A, Paape T, Steele K, Weiblen GD, Bharti AK, Zhou P, May GD, Young ND, Tiffin P. Phylogenetic signal variation in the genomes of *Medicago* (fabaceae).Syst Biol. 2013 May 1;62(3):424-38.