Dinamika dobitka i gubitka gena nakon genske duplikacije

Mrnjavac, Andrea

University of Zagreb

Faculty of Science

Department of Biology

Andrea Mrnjavac

Large-scale inference of gene gain and loss dynamics following gene duplication

Master's thesis

Zagreb, 2019

This thesis was conducted at Genomic Microbiology Group, Institute of Microbiology, Christian-Albrechts-University, Kiel, Germany under supervision of professor Tal Dagan PhD and at Department of Biology of the University of Zagreb, Zagreb, Croatia under supervision of associate professor Damjan Franjević PhD. The thesis was submitted for the evaluation to the Department of Biology at the Faculty of Science, University of Zagreb in order to obtain the title of Master of Molecular Biology (mag.biol.mol.).

BASIC DOCUMENTATION CARD

University of Zagreb

Faculty of Science

Department of Biology

Master's Thesis

LARGE-SCALE INFERENCE OF GENE GAIN AND LOSS DYNAMICS FOLLOWING GENE DUPLICATION

Andrea Mrnjavac

Rooseveltov trg 6, 10000 Zagreb, Croatia

Gene turnover (gene gain and loss) is ever occurring process in genomes of both eukaryotes and prokaryotes. Forms of gene gain are: duplication of an existing gene in a genome, *de novo* evolution from noncoding regions of a genome or horizontal gene transfer from one genome to another. Uncovering gene gain and loss events in genomes' histories is usually done by comparing gene trees with species trees, that is, tree reconciliation. A caveat in the existing reconciliation algorithms is that their resulting inference largely depends on the input parameters set by the user which can by themselves be very error-prone. Therefore, in this thesis, we developed a simple parameter-free algorithm for inferring duplication events, mapping them on the branches a rooted species tree and inferring losses that followed the inferred duplication event. Our algorithm only assumes a rooted species tree and rooted gene trees. Developed algorithm was used to analyze genome evolution in prokaryotes and eukaryotes. Obtained results suggest differences in horizontal gene transfer rates between prokaryotic genomes evolution and eukaryotic genomes evolution and overall prevalence of duplication and loss processes.

(41 pages, 22 figures, 2 tables, 34 references, original in: English)

Thesis deposited in the Central Biological Library

Key words: phylogenomics, tree reconciliation, eukaryotes, prokaryotes, genome evolution, horizontal gene transfer

Supervisors: Assoc Prof. Damjan Franjević PhD (University of Zagreb)
Prof. Tal Dagan PhD (Christian-Albrechts-University Kiel)

Reviewers: Assoc Prof. Damjan Franjević PhD (University of Zagreb)
Prof. Dijana Škorić PhD (University of Zagreb)
Rosa Karlić PhD (University of Zagreb)

Thesis accepted: 31.1.2019.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Biološki odsjek Diplomski rad

DINAMIKA DOBITKA I GUBITKA GENA NAKON GENSKE DUPLIKACIJE

Andrea Mrnjavac

Rooseveltov trg 6, 10000 Zagreb, Hrvatska

Genomi eukariota kao i prokariota kontinuirano prolaze kroz procese dobivanja i gubljenja gena. Procesi dobivanja gena mogu biti: duplikacija postojećeg gena u genomu, *de novo* evolucija iz nekodirajućih dijelova genoma ili horizontalni transfer gena iz jednog genoma u drugi. Uobičajen pristup razotkrivanju događaja dobitka ili gubitka gena u povijestima genoma je usporedba stabla gena i stabla vrsta *(tree reconciliation).* Problem s postojećim algoritmima za tu svrhu je što su im rezultati uvelike ovisni o ulaznim parametrima koji su neprecizni. U ovom radu izrađen je jednostavan neparametarski algoritam za određivanje duplikacija gena, mapiranje duplikacije na odgovarajuću granu u stablu vrsta, te određivanje broja gubitaka kopija koji su uslijedili nakon te duplikacije. Prednost našeg algoritma je što su mu jedini ulazni parametri ukorijenjena stabla gena i stablo vrsta. Koristeći algoritam analizirana je dinamika dobivanja i gubljenja gena u evoluciji genoma prokariota i eukariota. Rezultati upućuju na veliku razliku frekvencije horizontalnog transfera gena u evoluciji prokariotskih i eukariotskih genoma te sveukupnu rasprostranjenost događaja duplikacije i gubljenja gena u evoluciji.

(41 stranica, 22 slika, 2 tablica, 34 literaturnih navoda, jezik izvornika: engleski)

Rad je pohranjen u Središnjoj biološkoj knjižnici

Ključne riječi: filogenomika, usporedba stabala, eukarioti, prokarioti, evolucija genoma, horizontalni prijenos gena

Voditelji: izv. prof. dr. sc. Damjan Franjević (Sveučilište u Zagrebu)
prof. dr. Tal Dagan (Christian-Albrechts-Sveučilište u Kielu)

Ocjenitelji: izv. prof. dr. sc. Damjan Franjević (Sveučilište u Zagrebu)
prof.dr.sc. Dijana Škorić (Sveučilište u Zagrebu)
doc.dr.sc. Rosa Karlić (Sveučilište u Zagrebu)

Rad prihvaćen: 31.1.2019.

Table of contents

1 Introduction

1.1 Gene families

Gene, or multigene, family is a set of all orthologous and/or paralogous genes descending from a common ancestral gene. Paralogy means homology caused by duplication of an existing gene within a genome while orthology is homology due to speciation, that is, species divergence (Graur, 2016).

The number of genes within families, termed gene family size, can vary greatly within a genome (that is, by the number of paralogs in *sensu sticto*) and among genomes (that is, how many genes are separated through speciation events). One reason for that are gain and loss dynamics within the lineages. These processes are described in the *birth-and-death* model: genes (paralogs) arise in families by gene duplication of an existing gene member and can be removed through nonfunctionalization or deletion (Hughes & Nei, 1989).

1.2 Gene duplication

Gene duplications occur spontaneously within genomes. Mechanisms of duplication are homologous recombination, nonhomologous recombination, slippage within the replication machinery or (retro)transposition.

In some cases, the entire genomes can be duplicated. This event is known as whole genome duplication (WGD). During eukaryotic evolution this event occurred several times (Graur, 2016). Paralogs that originated from whole genome duplication are named ohnologs after Susumu Ohno (Wolfe, 2000).

In his famous book *Evolution by Gene Duplication* (1970), Ohno stated the importance of duplication for creating evolutionary novelties: "*Only the cistron which became redundant was able to escape from the relentless pressure of natural selection, and by escaping, it accumulated formerly forbidden mutations to emerge as a new gene locus*". Creating new functions is important, but only a minority of duplicates will become fixed in the population. Fixation of a duplicated gene is caused by neofunctionalization, retention of the original function or subfunctionalization. In the case of neofunctionalization, one copy retains the original function while the other one is released from selective constraint and accumulates mutations which allow performing a new function, while subfunctionalization assumes division of ancestral functions among the gene copies due to differential partial silencing of descendant

copies. Pseudogenization or nonfunctionalization due to deleterious mutations, on the other hand, is the more likely scenario for the newly arisen gene copy.

1.3 Gene loss

Gene loss is one driving force of the great biodiversity we observe today. Experiments (in a variety of taxa: *Escherichia coli* and other bacterial species, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, *Drosophila melanogaster*, mice and human cancer cell lines) have shown that only 10-35 % of genes are essential under laboratory conditions, which is known as the gene knockout paradox (Albalat & Cañestro, 2016). This is a confirmation of another Ohno's (1985) hypothesis, the hypothesis of gene dispensability. Redundant gene copies are to some extent responsible for gene dispensability phenomenon. In that case, one copy accumulates deleterious mutations and the other copies can act as a back-up. Additionally, there are alternative pathways for many biological functions (Albalat & Cañestro, 2016). These allow for high rates of gene loss, which do not have to be permanent: pseudogenes can be putatively resurrected by gene conversion and vice versa. This can create gene content variability in populations and may play a role in the evolution of the population, e.g., via subsequent adaptive radiation through differential gene loss.

Several scientists argue that evolution is mainly driven by gene loss. In his "less is more" hypothesis Olson (1999) stated that since only a small fraction of mutations are advantageous and mutations are most likely to cause loss of function, loss must be the main driving force in adaptive evolution. More recently, Wolf & Koonin (2013) hypothesized that genome reduction is the dominant mode of evolution. The most extreme examples can be observed in mitochondria and chloroplasts which lost nearly all ancestral genes, or hydrogenosomes and mitosomes which lost their genome completely. In prokaryotes and eukaryotes genome reduction is most perceivable in the case of endosymbionts. Here, genome reduction is occurring *via* the "neutral gene loss ratchet", that is, due to a small effective population size and genetic drift that greatly influences the gradual loss of non-essential genes. In free-living organisms, like most abundant cellular life forms on earth: the cyanobacterium *Prochlorococcus* sp. and the alpha-proteobacterium *Pelagibacter ubique*, genes are lost via adaptive genome streamlining. Here gene loss is driven by purifying selection for rapid genome replication and minimization of resources required. Wolf & Koonin (2013) also proposed a biphasic model of evolution which states that gene gain occurs in bursts and provides material for speciation in various niches through a long phase of differential gene loss occurring in a clock-like manner.

Experimental studies supported that proposal by showing that deletion of 25% of genes in *Salmonella enterica* can result in fitness increase depending on the environment (Albalat &

Cañestro, 2016). Adaptive gene loss has also been reported in diverse prokaryotic and eukaryotic species. For example, in *homo sapiens* the loss of receptor that binds HIV can cause resistance to AIDS (Dean et al., 1996). However, it is assumed that the fixation of gene loss is mostly driven by neutral evolution (Albalat & Cañestro, 2016). Additionally, a recent comparative genomics study (Iranzo et al., 2017) that modeled the dynamics of gene family sizes confirmed that most gene families are neutral or only slightly beneficial. The authors revealed that duplication (defined as “any process that causes increase in copy number that is proportional to the preexisting copy number”) to loss ratio (without selection) is 1:8 for prokaryotes. Gene birth rate is estimated as 1 fixed duplication per gene per 100 million years in eukaryotes, though the retention rate for each family varies with function (Graur, 2016). In addition, dosage selection can also play a role in gene loss patterns: some genes are cryptically resistant to duplication because their function is sensitive to changes in stoichiometry, while for other genes, the increased dosage can positively influence the retention rate (Albalat & Cañestro, 2016).

1.4 Horizontal gene transfer

Apart from gene duplications, horizontal gene transfer (HGT) contributes greatly to the rate of gene birth processes in prokaryotes (Treangen & Rocha, 2011) (Dagan & Martin, 2007) in contrast to eukaryotes, where HGT is debated to be significant only in terms of endosymbiotic gene transfer (Martin, 2017). Homology due to HTG is called xenology (Graur, 2016).

1.5 *De novo* gene evolution

Another form of gene gain is *de novo* evolution from noncoding regions of the genome. It is very unlikely, but not impossible event and its prevalence has been heavily debated. Recent experiments showed that large fraction of artificial random nucleotide sequences inserted into *Escherichia coli* have influence on its fitness (Neme et al., 2017). It is proposed that process of *de novo* gene emergence is mainly driven by abundant transcription of genomic regions into noncoding RNAs, which may provide raw material for the evolution of novel protein coding genes (Graur, 2016) (Tautz & Domazet-Lošo, 2011).

1.6 Phylogenetic trees

Phylogenetic trees are used to describe evolutionary relationships among taxonomic units: DNA sequences, genes, proteins, organisms, populations, species or higher taxa. A tree is a graph in which any two nodes are connected by a single path. Networks are connected graphs in which at least two nodes are connected by two or more pathways, and are useful to describe reticulate evolution. The nodes represent taxonomic units, and branches represent relationships among them. Terminal nodes in a tree represent the real data, and are termed operational taxonomic units (OTUs), while internal nodes are representing inferred ancestral taxonomic units and are termed hypothetical taxonomic units (HTUs). Trees can be bifurcating (or binary), where every node has two immediate descendants, or multifurcating, where each node can have two or more immediate descendants. (Graur, 2016).

1.7 Tree inference

Given the set of OTUs, there is only one phylogenetic tree which correctly represents their evolutionary history and such a tree is called the true tree. Methods of tree reconstruction are aiming to give the best estimate of the true tree. There are three main approaches in tree inference methods: distance matrix, character state and maximum likelihood. In distance methods, tree topology and branch lengths are inferred from distances (for example, number of nucleotide substitutions) between two OTUs. Character state methods use raw character state data (e.g., nucleotides, amino acids, or presence/absence) for tree inference. The maximum parsimony approach estimates the best tree topology as the one requiring the smallest number of evolutionary changes, it uses discrete character states and the shortest set of pathways leading to the observed states is chosen as the best, most parsimonious tree. The maximum likelihood approaches use both character state and distance data. The likelihood of phylogenetic tree is conditional probability of observing the data (sequences) given the tree and a probabilistic model of character state changes. The likelihood of each possible tree is calculated as the product of likelihoods of all sites in the alignment, each of which is calculated as the sum of probabilities of all possible scenarios leading to the observed data. Maximum likelihood estimates the best tree as the one with the highest likelihood. Although the method depends on choosing the right substitution model and is computationally time-consuming, it uses full raw character dataset, unlike maximum parsimony which only uses informative sites, and is considered to outperform other tree inference methods (Graur, 2016).

1.8 Tree rooting

The result of most tree inference methods is an unrooted tree. Unrooted trees lack temporal directionality and we cannot determine ancestor-descendant relationships. To infer the common ancestor of all taxonomic units in a tree we must root it, that is, add one rooting node that represents the last common ancestor (LCA) of all the OTUs. Commonly used rooting methods are outgroup rooting and midpoint rooting. In outgroup rooting we use an outgroup, that is, an OTU for which we know that branched off earlier from the ingroup OTUs. The node connecting the outgroup and the ingroup is defined as the root. Alternatively, midpoint rooting can be used without choosing an outgroup. In this approach we assume the rate of evolution is uniform throughout the branches and the midpoint of the longest pathway between OTUs is defined as the root (Graur, 2016).

Recently, a new rooting method was presented, the MAD (*Minimal Ancestor Deviation*) method. Midpoint approach assumes that under strict molecular clock the middle of the path between any two OTUs should correspond to their last common ancestor. MAD evaluates all the branches in an unrooted tree by calculating the mean relative deviation of LCAs induced by the candidate root from the LCAs inferred as midpoints, that is, departure from molecular clock. The root is placed on the candidate branch that has the lowest mean relative deviation. MAD method works with any kind of binary tree and was shown to outperform other rooting methods (Tria et al., 2017).

1.9 Gene trees and species trees

Distinct biological species can arise by *cladogenesis*: splitting of lineages, *anagenesis*: change within a lineage, or *genome hybridization*: merging of lineages. Speciation, the process of species diversification occurs mainly through cladogenesis, therefore, bifurcating *species tree* is a good representation of evolutionary relationships among taxa (Graur, 2016). The evolution of a gene family can be described with phylogenetic tree that is a *gene tree*. Gene trees are usually very different than species trees. Together and in addition to species diversification, gene diversification is constantly happening. Gain and loss events cause variation of evolutionary histories among gene families and differences between gene trees and species tree. Besides orthologs, gene tree OTUs can include paralogs and xenologs. The origin of gene gain from gene duplication or horizontal gene transfer can be inferred in theory, but in practice this is a challenging task.

1.10 Tree reconciliation

There are many evolutionary scenarios that could explain the observed variety in multigene families and the inference of the one true tree is uncertain. One way to estimate gain and loss dynamics in histories of gene families is by comparing gene trees with the species tree. Basic scenarios that could influence gene tree topology are summarized in Figure 1.

Explaining differences between a gene tree and a species tree by gain and loss events is termed reconciliation (Graur, 2016). Reconciliation methods aim to estimate the most likely or the most parsimonious scenario of gene family evolution. The methods working in parsimony framework, where each of the possible events has a fixed cost set by the user, choose the scenario with the lowest reconciliation cost as the best one. Bayesian and likelihood based methods work by estimating rates of possible gain and loss events and subsequently inferring the most likely reconciliation scenario. As we can see, reconciliation methods take a lot of assumptions: species tree, gene tree, gain and loss event rates or costs. Method evaluations showed that reconciliation scenario inference methods results are not robust to changes in parameters (Kamneva & Ward, 2014).

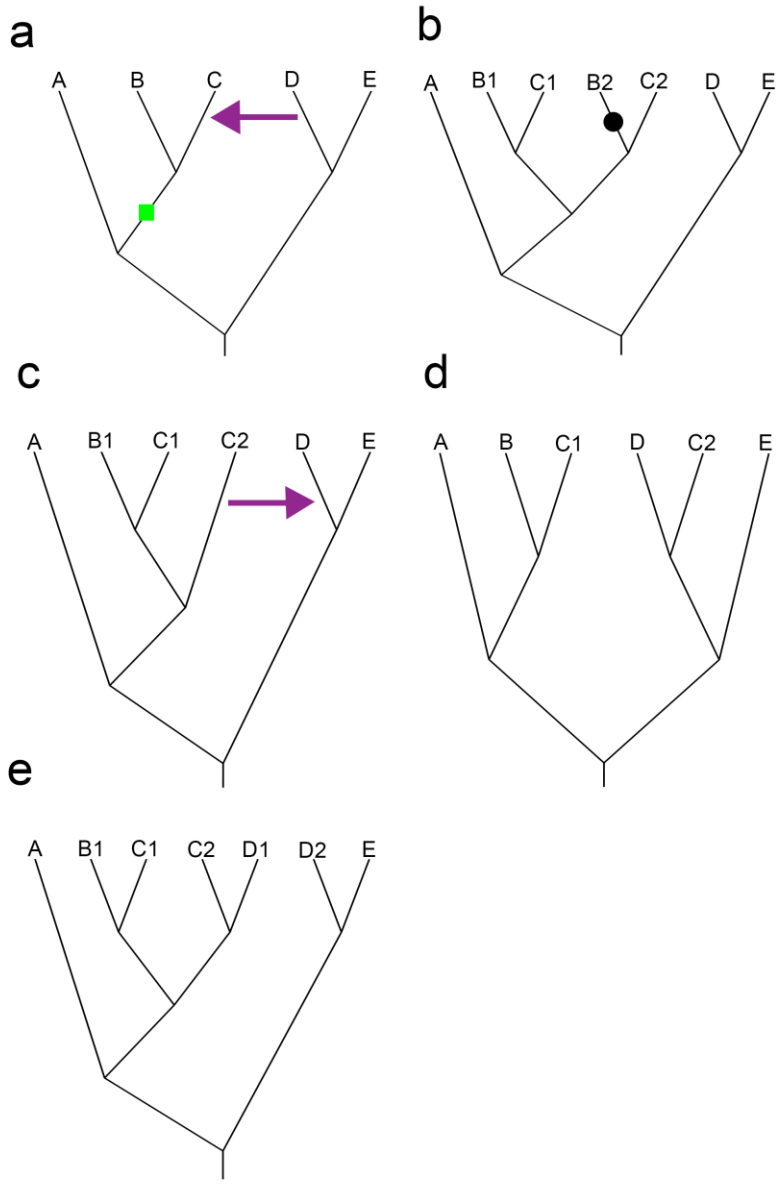


Figure 1. examples of gene tree topology changes due to basic gain and loss scenarios

a) rooted species tree

b) rooted gene tree topology if duplication occurred before speciation of species B and C but after the divergence of clade containing species B and C from the branch leading to A (green square on the tree a);

c) rooted gene tree topology if after the mentioned duplication one copy was lost in specie B (black dot on the tree b);

d) rooted gene tree topology if horizontal gene transfer occurred from specie D to specie C (purple arrow on the tree a);

e) rooted gene tree topology if in addition to mentioned duplication and loss, HGT of C2 occurred from specie C to specie D (purple arrow on the tree c)

2 Objective

In this master thesis I aim to explore gene gain and loss dynamics in prokaryotic and eukaryotic gene families. The commonly used approach for this problem is tree reconciliation. A caveat in the existing algorithms is that their resulting inference largely depends on the input parameters set by the user which can by themselves be very error-prone. A solution for this problem is using parameter-free approach. Such an approach would allow us to estimate general trends in gene gain and loss events along the evolution of gene families.

My main objective is thus to develop a simple algorithm for large-scale inference of gene gain and loss dynamics. Applying the algorithm to prokaryotic and eukaryotic gene families will be useful in order to obtain new insights on genome evolution in eukaryotes and prokaryotes.

3 Materials and methods

Two datasets were analyzed, one of eukaryotic gene families and one of prokaryotic gene families. As the algorithm input preparation, two species trees were reconstructed: eukaryotic and cyanobacterial one in addition to gene trees, as summarized in table 1. The algorithm input preparation is described in detail in the following paragraphs. The two datasets were generated with slightly different pipelines due to the use of different sources of data.

3.1 Eukaryotic dataset preparation

Altogether 31 opisthokonta (the best supported eukaryotic supergroup) species (14 animals and 17 fungi) that represent major taxonomic clades were selected. Complete EggNOG database (a database of orthologous groups and functional annotation), version 4.5 (Huerta-Cepas et al., 2016) was downloaded and 18,481 protein families with at least 3 species (from 31 chosen species) present were identified. For each of the families, corresponding protein sequences were extracted from the database, and turned into multi-FASTA format. Sequences were aligned with MAFFT, version v7.027b (Katoh & Standley, 2013) using an accurate option (L-INS-i). Finally, maximum likelihood gene trees were reconstructed with PhyML, version 20120412 (Guindon & Gascuel,2003), using parameters '-b -4 -s SPR' which give SH-like branch support. SH-like branch support implemented in PhyML is fast and nonparametric measure derived from SH multiple tree comparison procedure (Shimodaira & Hasegawa, 1999). SPR (*Subtree Pruning and Regrafting*) is a tree rearrangement algorithm used in search for an optimal tree structure (Hordijk & Gascuel, 2005). These trees were obtained from unpublished work by Fernando Tria. Majority-rule consensus (which means splits of frequency 50% or more were adopted) species tree was reconstructed from 117 universal single-copy gene trees using Consense, version EMBOSS:6.6.0.0 PHYLIPNEW:3.69.650 (Felsenstein, 1993). All gene trees were rooted with MAD (Tria et al., 2017), and the branch in the species tree, inferred as the root in majority of universal single-copy gene trees, was selected as the root. Branch lengths for the species tree were estimated as the medians of corresponding branch lengths from the universal single-copy gene trees.

3.2 Cyanobacterial dataset preparation

Genome assembly files were downloaded from NCBI Reference Sequence (RefSeq) database (version May 2016), using FTP protocol with "wget" command, for all (379)

cyanobacteria complete genomes. Multi-FASTA files of all annotated protein sequences for each genome were transformed into a BLAST (Altschul et al.,1990) database with "makeblastdb" command. All created databases were protein blasted (blastp, version 2.2.26) against all protein multi-FASTA files. Pairs of proteins that are each other's best BLAST hits (reciprocal BBHs) are very likely to be orthologous (Tatusov et al., 1997). rBBHs with E-value equal or lower than $1x10^{-5}$ were globally aligned using needle, version EMBOSS:6.6.0.0 (Rice et al., 2000). Pairs with 30% or more identical amino acids were clustered into protein families using Markov clustering algorithm MCL, version 12-135 (Enright et al., 2002). Mysql database with all the clustered proteins and their sequences in 379 cyanobacterial species and strains was obtained from the unpublished work by Tal Dagan. 47 section IV and V species with reliable genomes (less than 300 scaffolds and unimodal distribution of gene GC content and codon adaptation index) were chosen for further analysis. 300 universal single-copy families for the chosen set of species, that were present in more than 300 cyanobacteria genomes (which should be a strong evidence of orthology), were identified. Alignments for each of the mentioned families were produced with MAFFT using default parameters, version v7.123b and concatenated. Maximum likelihood species tree was inferred from the concatenated alignment, together with 100 bootstrap replicates, using PhyML, version 20131022 and parameter "-s SPR". Alignments were made with MAFFT using default parameters for all protein families present in the 47 species set with 4 or more protein family members. Quality of the alignments was tested with Heads or Tails (HoT) method (Landan & Graur, 2007). HoT method is based on the assumption that perfect sequence alignments should be independent of the input sequence orientation, therefore HoT measures the agreement between original alignment and the one produced from reversed sequences. 8,288 alignments that had mean column score equal or higher than 95 % were used for maximum likelihood gene tree inference using IQ-tree, multicore version 1.5.5 for Linux 64-bit (Nguyen et al., 2015), with parameter "-bb 1000" that gives ultrafast bootstrap approximation support for the branches (Minh et al., 2013). All trees were rooted with MAD.

3.3 Simple approach to identify and date duplications and subsequent losses

All reliable internal branches (branch was considered reliable if it had SH-like branch support value >=0.80 for eukaryotic trees and if it had ultrafast bootstrap approximation value >=95 for cyanobacterial trees) in all gene trees were examined/tested to represent a lineage in which duplication event happened in the following way:

1. In rooted bifurcating tree, each internal branch has two descendant clades. First sign that duplication happened is finding OTUs from the same species in

both of the descendant clades. If this is the case, there are two possibilities considering descendant OTUs content (figure 2).

2. If OTUs from only one species are descending from the examined branch (figure 2b) it is a clear case of inparalogs (although it is possible that the duplication is actually older than the speciation, but appears to be recent because of gene conversion). If there are OTUs from more than one species descending from the examined branch (figure 2c) it is possible duplication happened in the ancestral lineage of all the species in such clade or that HGT happened between left and right side of such clade. Unfortunately there is no precise way to tell these two scenarios apart. There are few algorithms designed for this problem, but each of them depends on error-prone assumptions, as mentioned in the introduction. Being aware that more HGT would lead us to infer more deep duplications and higher loss frequency, in this approach duplication event is matched to the last common ancestral branch of all the species in the clade. This is not absolutely correct even in the absence of HGT, it is possible that duplication is older than the LCA of species in clade, and we could narrow down the number of possible candidate branches for the given duplication event by taking into account speciation event preceding the duplication.

3. With this approach duplication event is matched to the branch in the species tree with speciation events as reference points which allows inference of subsequent loss. For each inferred duplication event, frequency of species that lost one and two copies afterwards is inferred in a following way: when we identify branch in which duplication occurred in the species tree we can ask what are the species descending from that branch and what is the frequency of these species that lost zero, one or both copies after assumed duplication in a gene tree.

3.4 Algorithm application and further analyses

This approach was applied on two different datasets: prokaryotic and eukaryotic. Basic output values are summarized in table 1.

Furthermore, effect of tree quality on the approach was examined on both datasets: gene trees were separated into two subsets: ‘high quality’ and ‘low quality’. ‘High quality’ subset of trees contained the trees whose median of all bootstrap values in the tree was equal or higher than 0.5 and whose root had value of ambiguity index equal or lower than 0.95, and the ‘low quality’ subset contained the rest of them. Distributions of inferred duplication events for each branch in the species tree were compared. Results of this analysis are presented in supplementary material (figures 21 and 22).

Additionally, to get a better perspective on gene family evolution, emergences of gene families were matched to the branch in the species tree taking LCA branch to all the species present in a given family to represent lineage in which emergence of a said family happened. For eukaryotic single-copy families, subsequent losses were counted in the previously described way. Loss frequencies for multi-copy and single-copy families were compared.

Inferred number of duplications was normalized by branch length (only for eukaryotes) and number of inferred duplications per family was calculated.

The algorithm and the analysis described was executed using custom scripts with MATLAB version R2015b. Code available in the supplementary material.

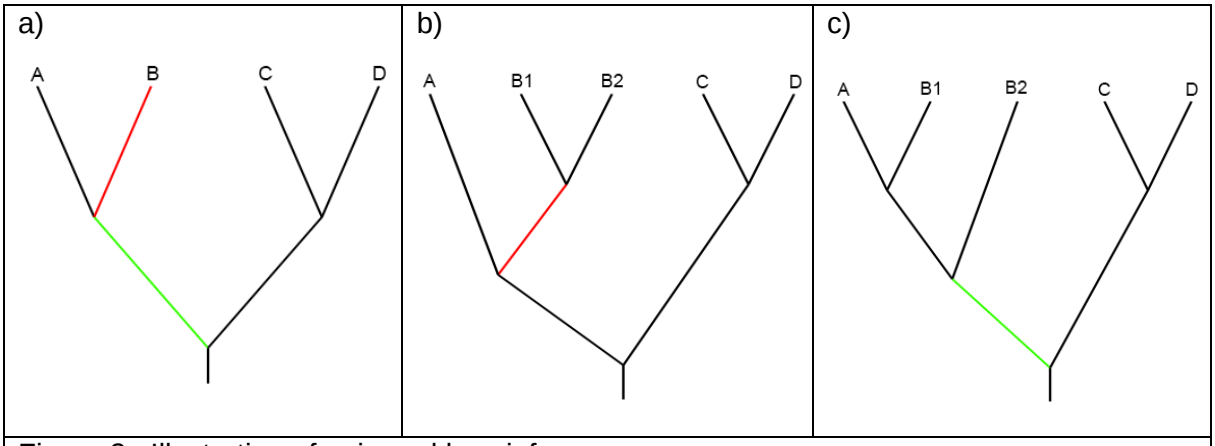


Figure 2. Illustration of gain and loss inference

a) Rooted species tree

b) Rooted gene tree where inspected branch (red branch) has genes from only one specie in both left and right descendant clades, in this case B. We can assume that duplication event which caused such topology occurred in a lineage represented by red branch in the species tree, that is, terminal branch leading to B.

c) Rooted gene tree where inspected branch (green branch) has genes from the same specie (B) in both left and right descendant clades, which is a sign of duplication event, and in this case one of descendant clades also involves a gene from another specie, in this case A. We can assume duplication occurred in the LCA of the species A and B, the lineage represented by the green branch in the species tree. To explain given tree topology we must also assume that one of duplicated copies was lost in the specie A, but no loss occurred in the specie B. For a given inferred duplication event, loss inference would be: 0% of species that lost both copies after inferred duplication, 50% of species that lost one copy after inferred duplication and 50% of species that lost zero copies after inferred duplication.

Table 1. Summary of input and output values for eukaryotic and cyanobacterial datasets		
Dataset	Eukaryotes	Cyanobacteria
Number of species (genomes) considered	31 species (14 animals, 17 fungi)	47 species (sections IV and V)
Input	8,257 gene trees (multi-copy) (and 10,105 partial single-copy trees) consensus species tree	8,288 gene trees (of which 1,803 multi-copy) concatenated species tree
Number of clades diverging from trusted branches with members from at least one species on both sides	46,041	2,569
Number of such clades containing members of only one species (in-paralogs) (figure 2, b)	23,391 (50.8 %)	399 (15.5 %)
Number of such clades containing members from more than one species (figure 2, c)	22,650 (49.2 %)	2,170 (84.5%)

4 Results

4.1 Species trees

Two species trees were reconstructed: eukaryotic, as presented in figure 3, and cyanobacterial, in figure 4.

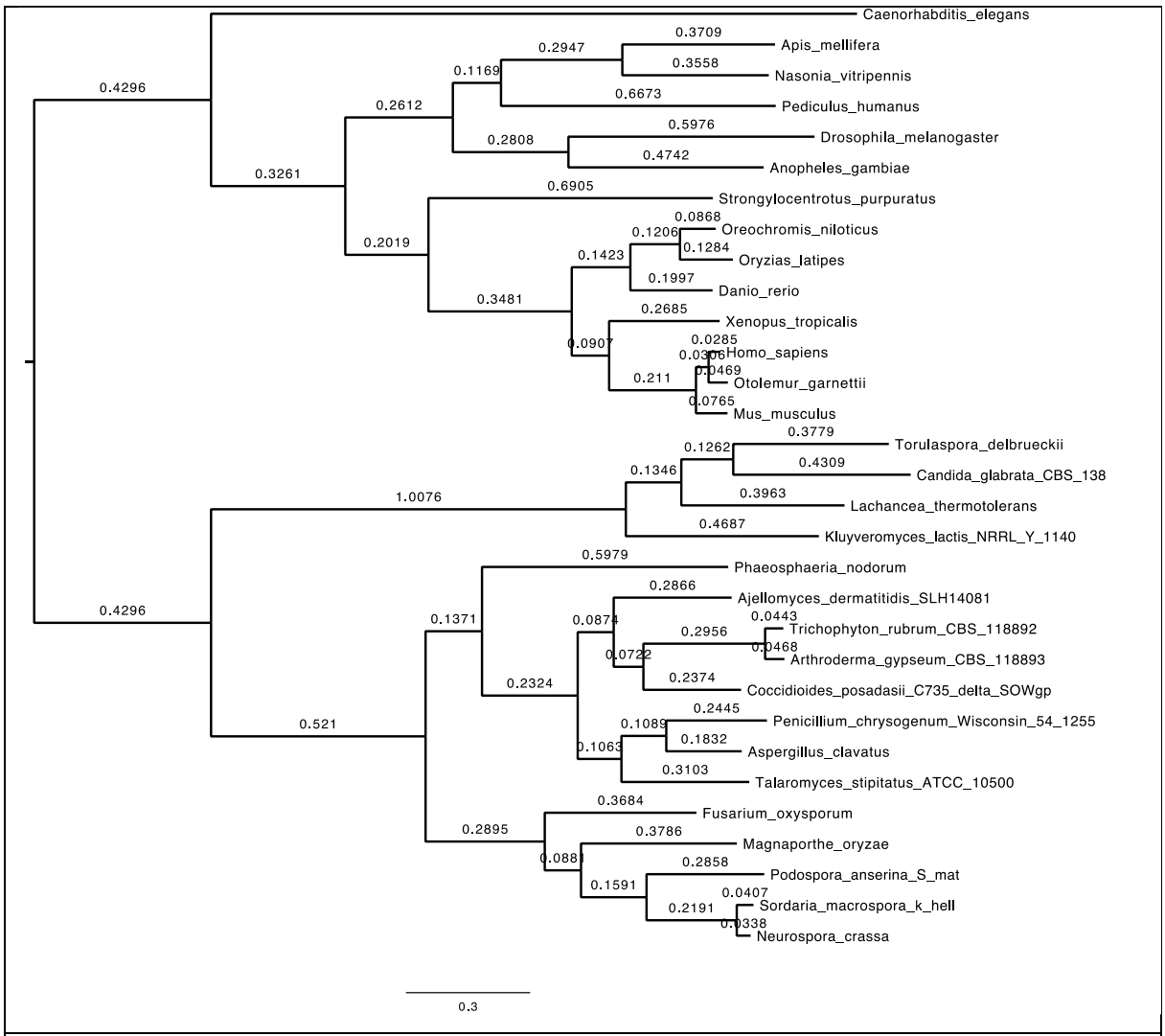


Figure 3. Eukaryotic species tree

Branch lengths (numbers on branches) represent evolutionary distances

For eukaryotic species tree the root is clearly positioned between fungi and metazoa as confirmed by Tria et al. (2017), while the root of cyanobacterial species tree is highly ambiguous (MAD ambiguity index 0.99822).

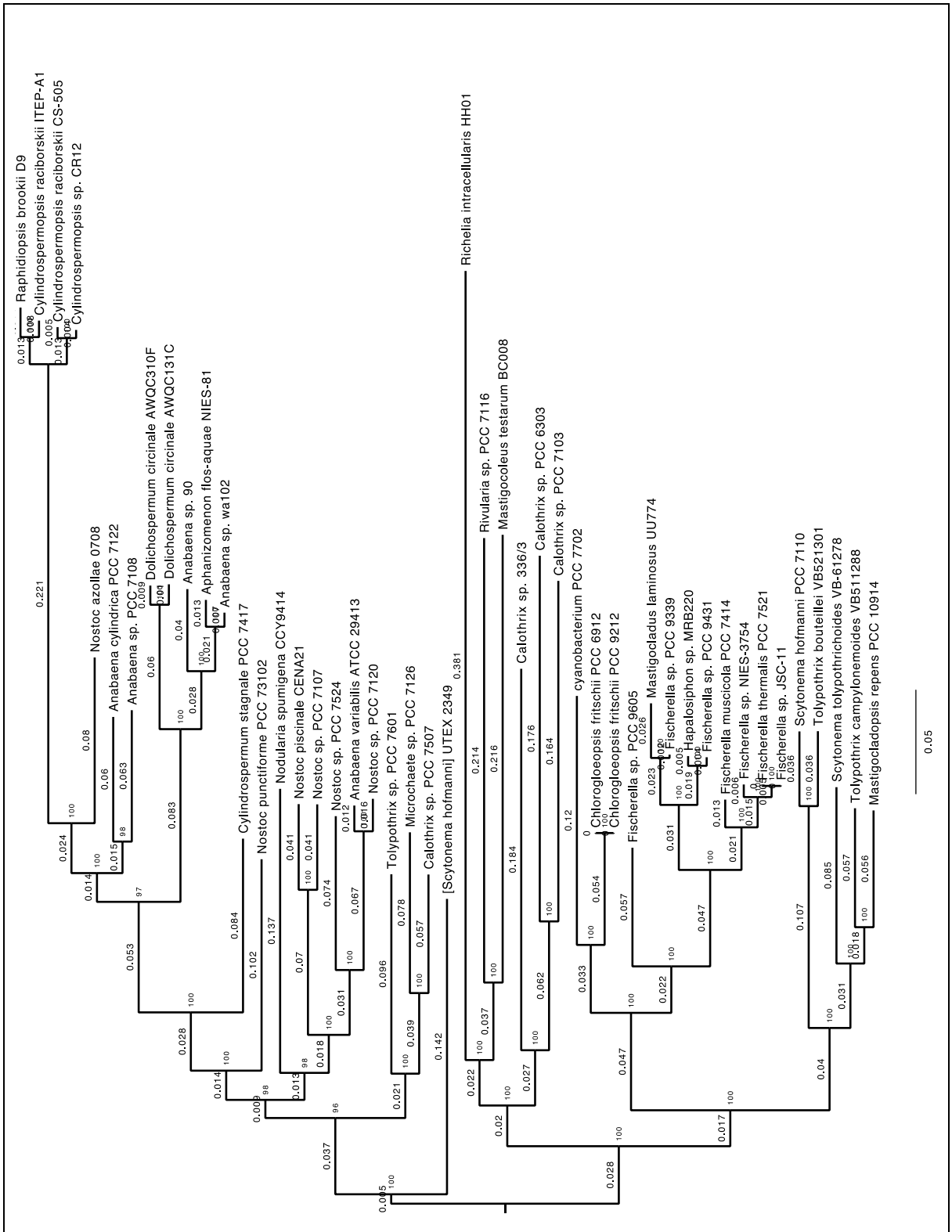


Figure 4. Cyanobacteria (sections IV and V) species tree

Branch lengths (numbers on branches) represent evolutionary distances while numbers at nodes represent bootstrap support on scale 0-100

Heterocystous cyanobacteria are a monophyletic group of multicellular filamentous organisms, closely related to plastid ancestor (Dagan et al., 2013). They consist of sections IV

and V and heterocystous cyanobacteria species tree we inferred suggests section V is polyphyletic. This is opposed to previous assumptions (as in Dagan et al., 2013).

In the presented trees, branch lengths, derived from estimated number of amino acid substitutions between taxonomic units, represent evolutionary distances.

4.2 Mapping duplications to the species trees

Using the developed algorithm (described in section 3.3), the occurrences of gene duplications were inferred in respect to speciation events presented is species trees. That is, inferred duplication events were assigned to the appropriate species tree branches. The inferred numbers of duplication events were visualized in form of branch lengths in the cyanobacteria species tree (figure 5) and the eukaryotes species tree (figure 6).

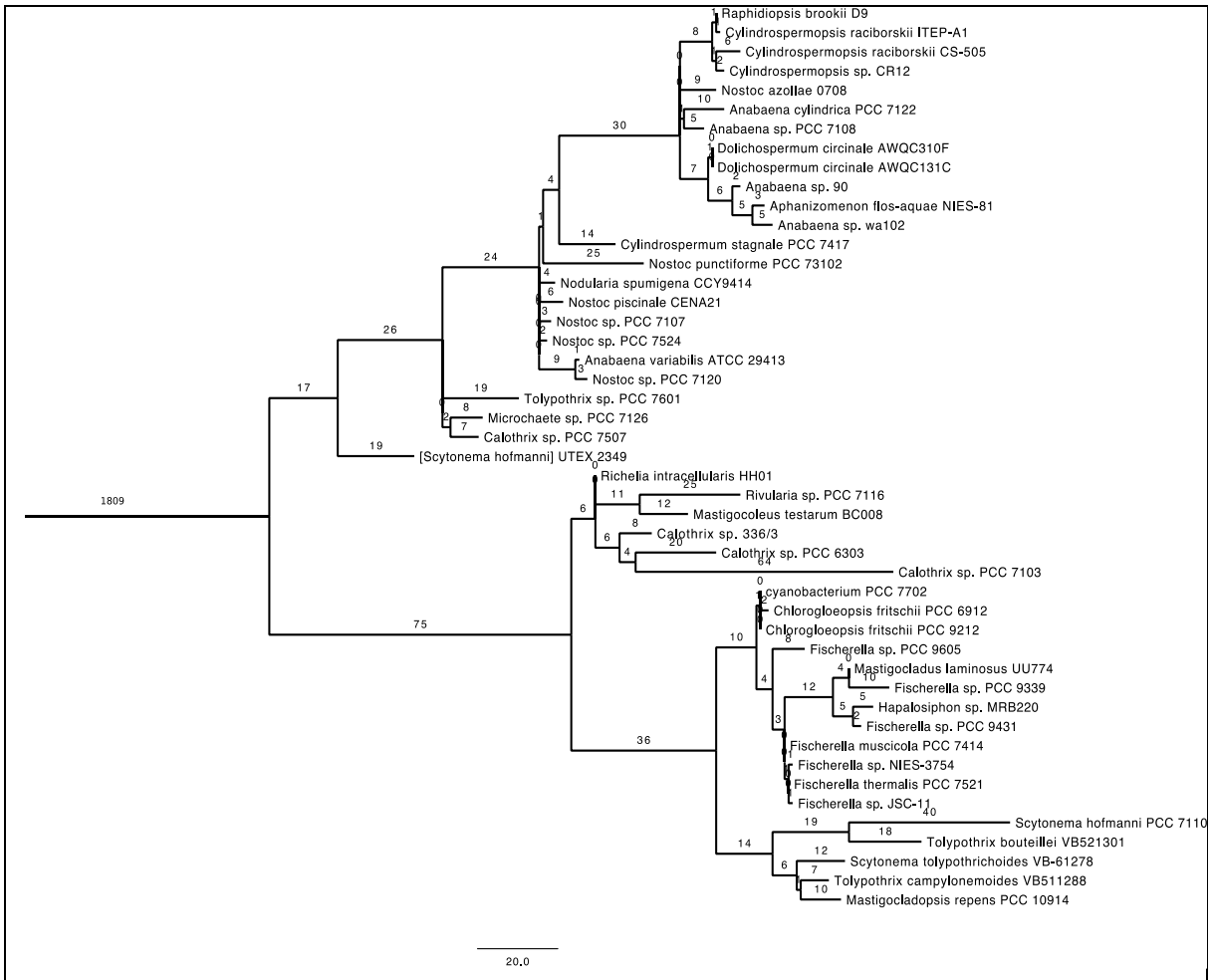


Figure 5. Number of inferred duplication events per branch in cyanobacteria species tree represented as branch length

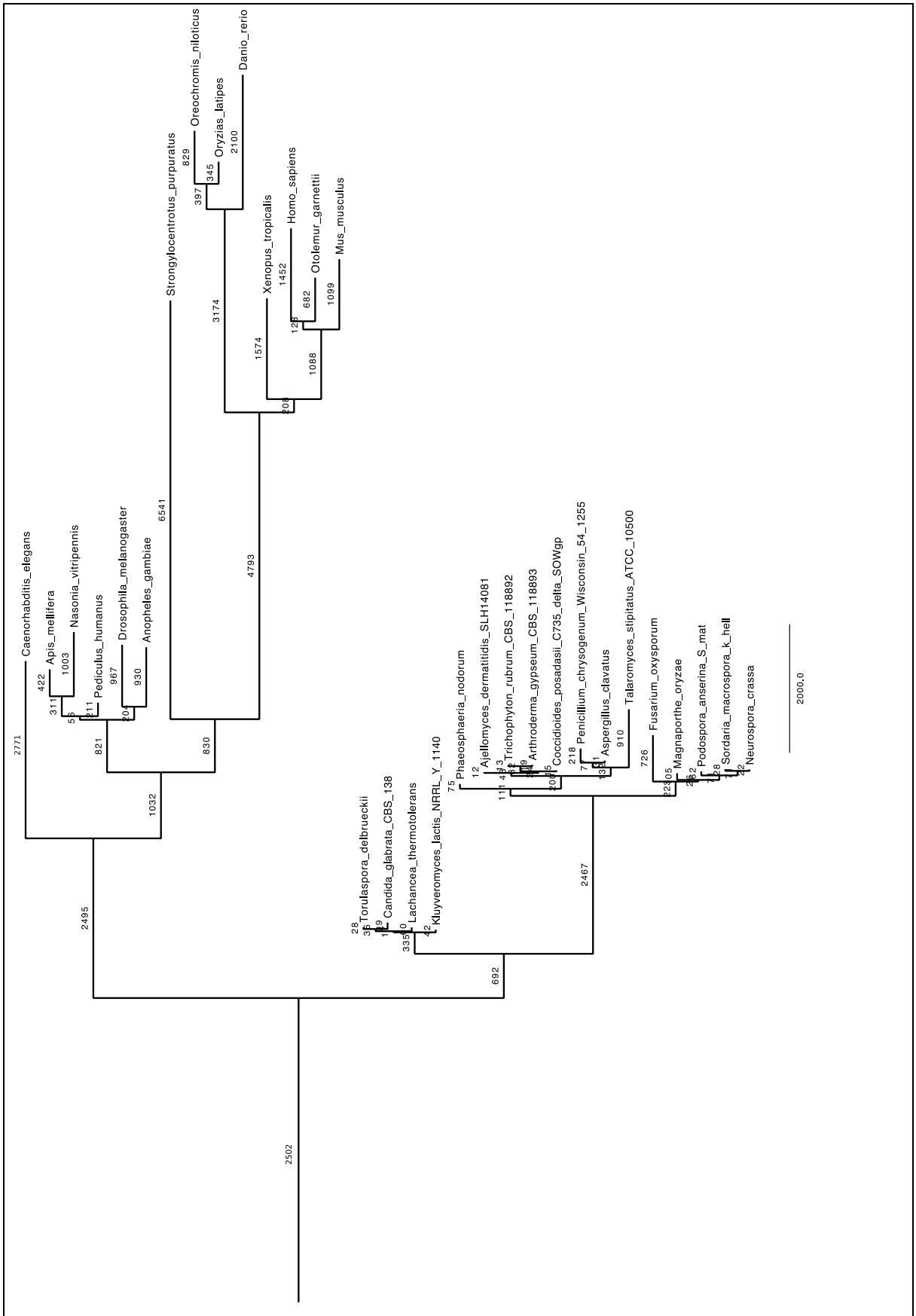


Figure 6. Number of inferred duplication events per branch in eukaryotes species tree represented as branch length

The majority of gene duplications in cyanobacterial gene families were mapped to the root of the species tree, that is, they were inferred as ancestral to heterocystous cyanobacteria (figure 5). Most of the other duplications were mapped along the branches leading to the species with the largest genomes (table 2 in supplementary material). In contrast, the distribution of inferred duplications along branches of eukaryotic species tree does not have such extreme outliers (figure 6). Anyway, notable is the difference in duplication patterns between fungi and animals: there are more duplication events inferred in metazoan evolutionary history.

Assuming duplication events occur spontaneously and continuously, we expect more duplication events in longer branches. There is very weak correlation between branch length and number of duplications inferred for that branch in eukaryotic species tree: Sprearman r=0.348, p-value=0.006. To obtain a perspective on how the number of duplications per branch is influenced by the branch length, figure 7 shows estimates of “duplication rate” per branch. Inferred number of duplications per branch (visible in figure 6) was divided by the original branch lengths of the species tree (visible in figure 3). Such duplication number normalization for cyanobacterial dataset was not necessary because of the mentioned distribution of duplications per branch.

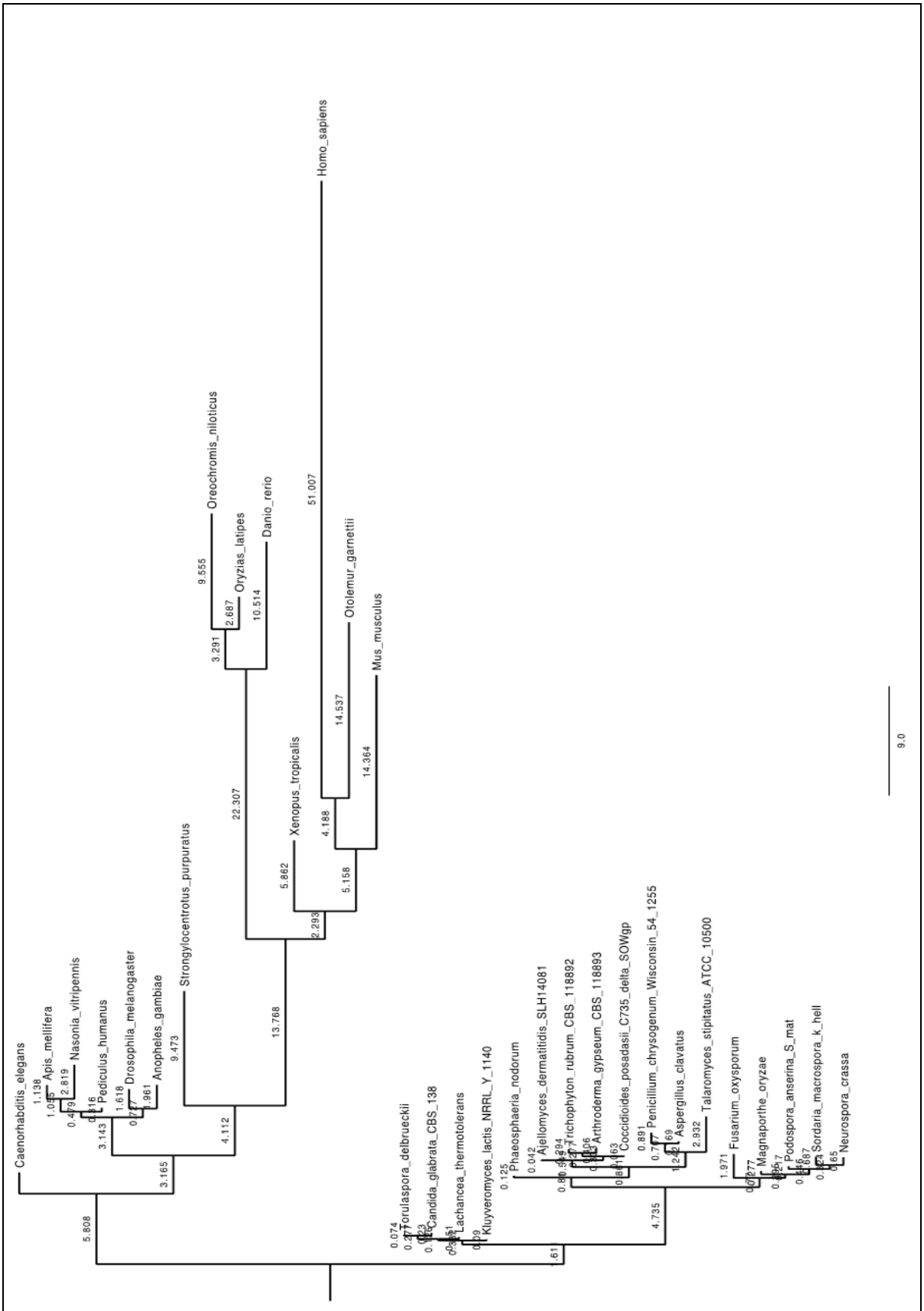


Figure 7. Number of inferred duplication per branch in eukaryotes species tree divided by the original branch length, visualized as branch length

4.3 Mapping gene family emergence events to the species trees

From the same species trees and gene trees, using the algorithm analogous to the one for mapping duplication events along species tree branches, fraction of family emergence per branch was estimated and presented in figure 8 for cyanobacterial dataset.

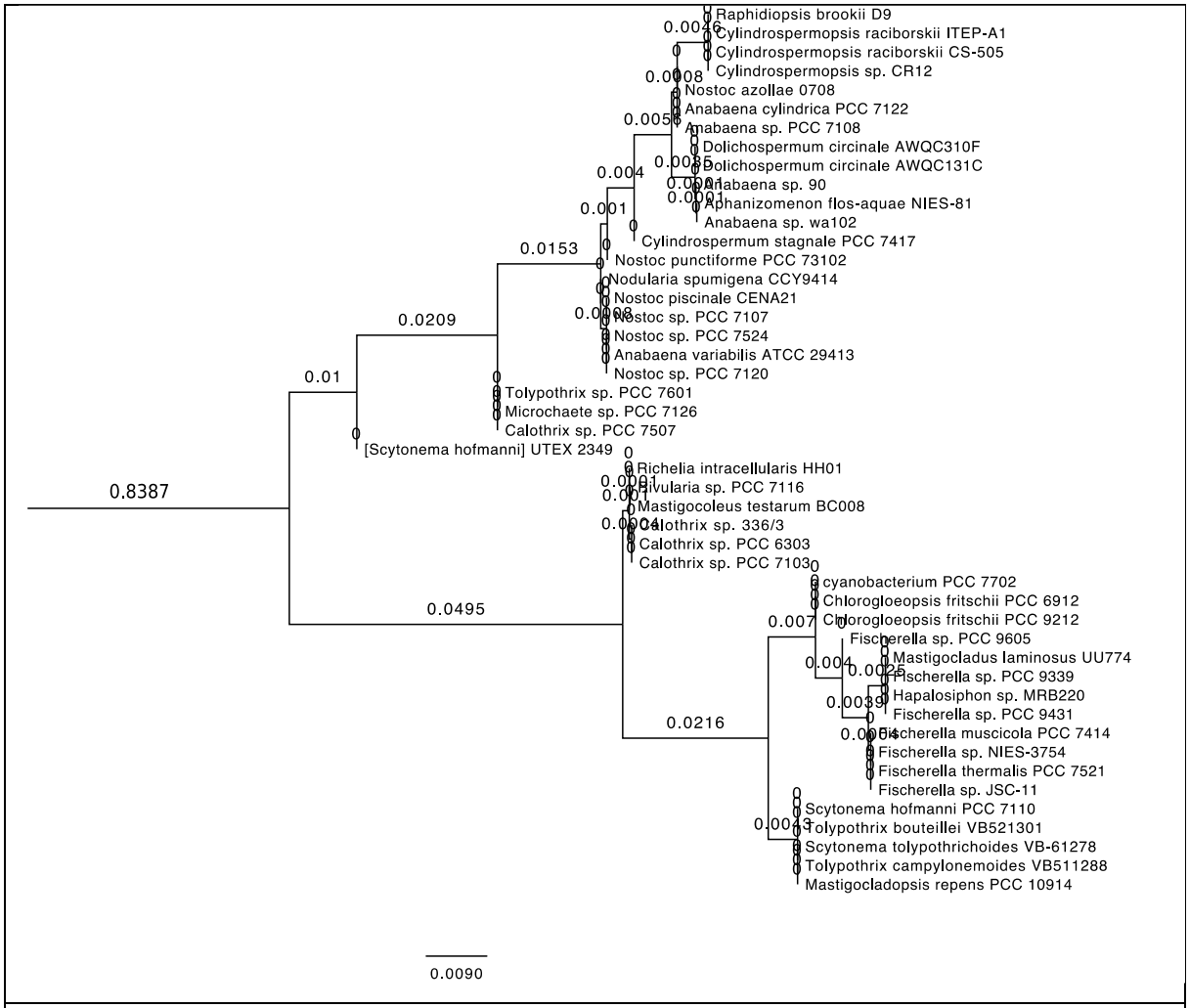


Figure 8. Cyanobacteria species tree with branch length representing fraction of gene families emerging at a given branch (for each branch: number of inferred number of families emerging divided by the total number of gene families)

Estimated number of family emergences per branch was compared to the inferred number of duplications per branch to examine the correlation as presented in figure 9.

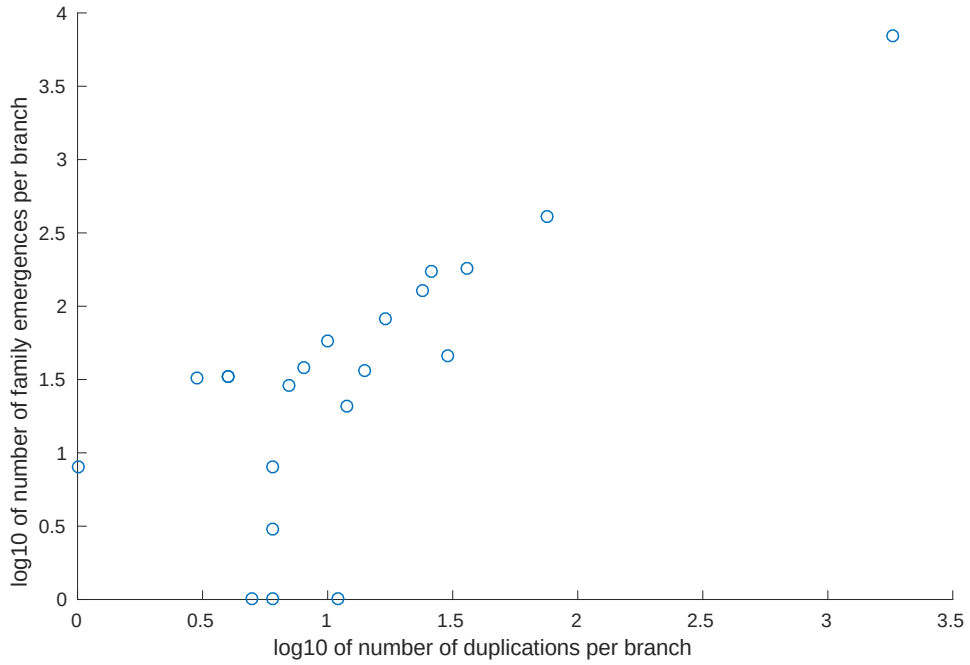


Figure 9: Number of cyanobacteria family emergences per branch compared to number of duplications per branch: Spearman r=0.67, P=3.17x10^{-7}

For eukaryotic dataset fraction of family emergence per branch was estimated and presented separately for multi-copy gene families and partial single-copy families in figures 10 and 11 respectively.

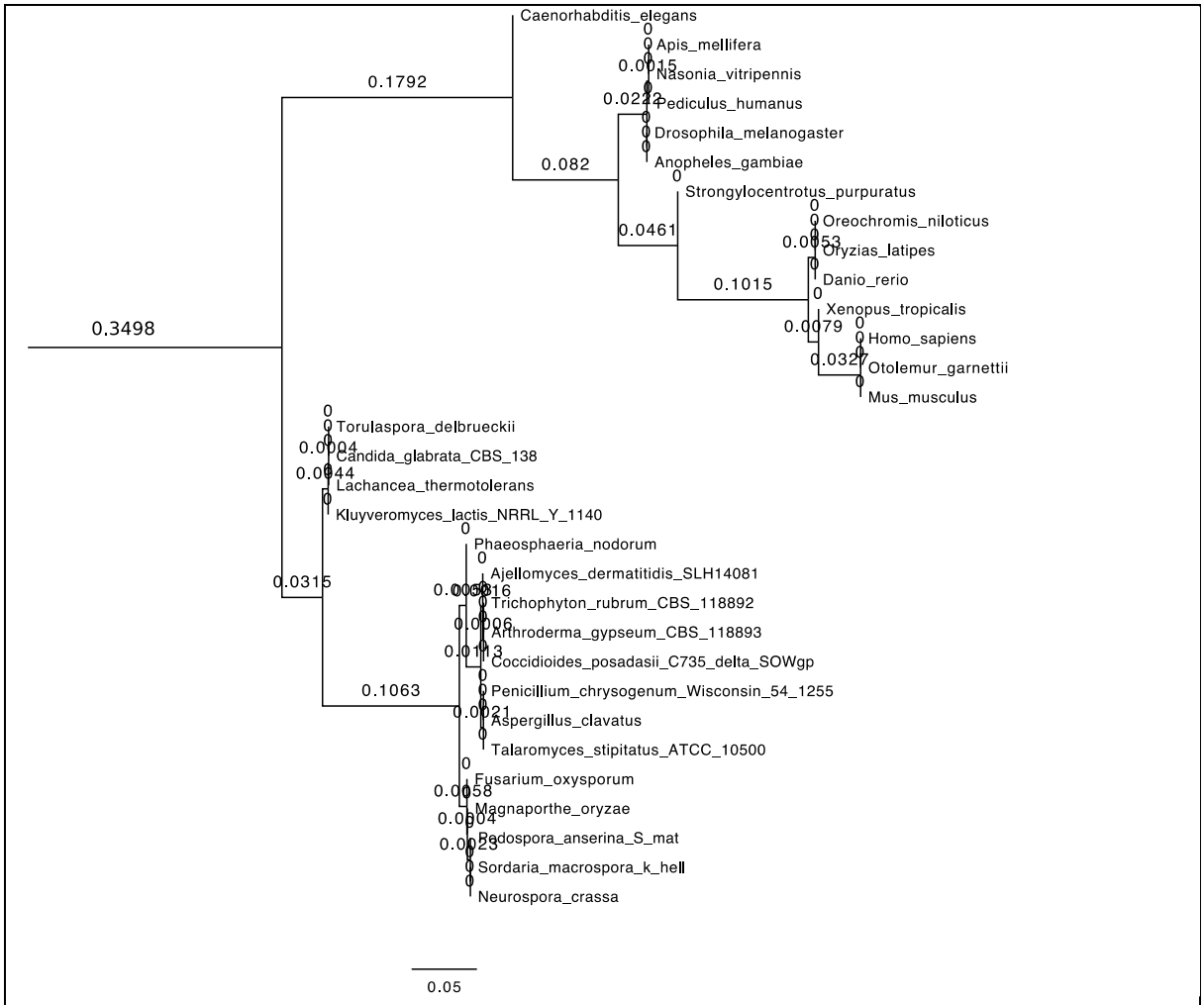


Figure 10: Eukaryotes species tree with branch length representing fraction of multi-copy gene families emerging at a given branch (for each branch: number of inferred number of families emerging divided by the sum of multi-copy gene families)

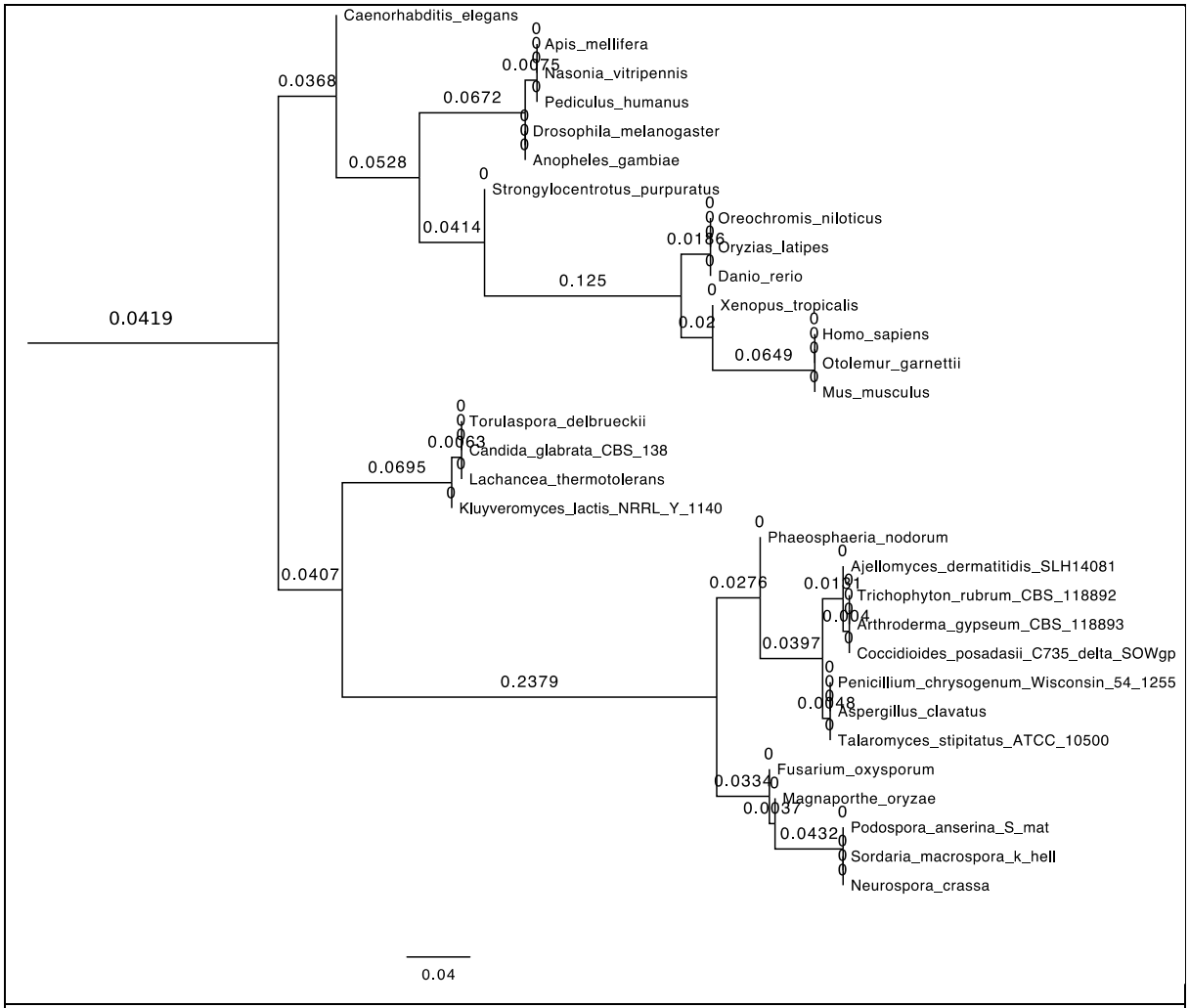


Figure 11: Eukaryotes species tree with branch length representing fraction of single-copy gene families emerging at a given branch (for each branch: number of inferred number of families emerging divided by the sum of single-copy gene families)

Correlation between emergence patterns in single-copy and multi-copy gene families was inspected as visible in figure 12.

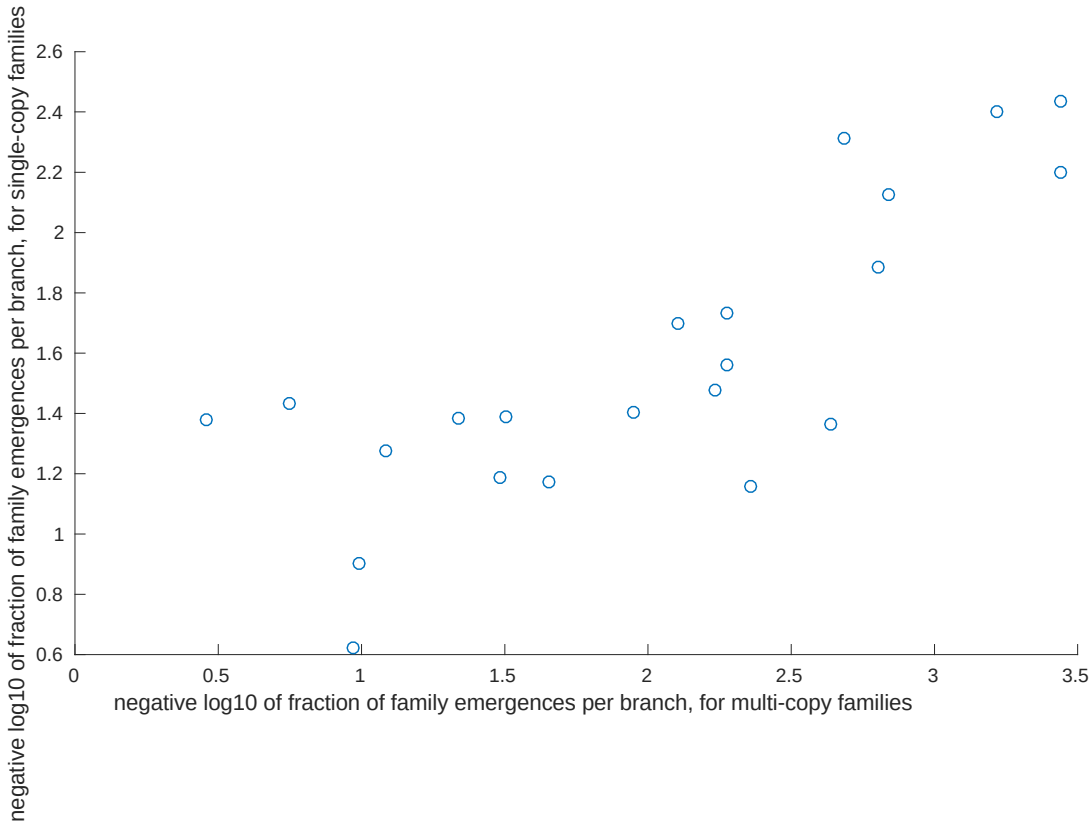


Figure 12: Number of family emergences per branch compared in eukaryotic multi-copy and single-copy families (Spearman r=0.89, $P=2.9x10^{-11}$)

Estimated number of eukaryotic multi-copy family emergences per branch was compared to the inferred number of duplications per branch to examine the correlation as presented in figure 13.

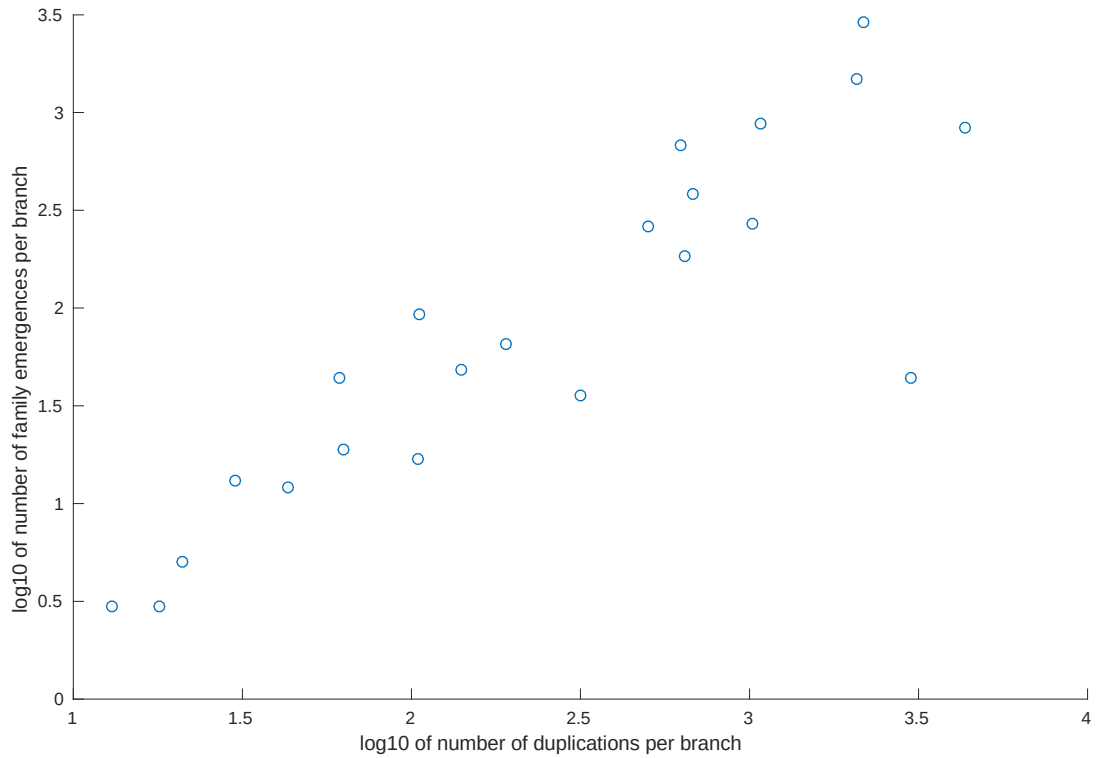


Figure 13: Number of family emergences per branch compared to number of duplications per branch in eukaryotic multi-copy families (Spearman r=0.66, $P=7.37 \times 10^{-5}$)

4.4 Gene family sizes

Number of inferred duplications per gene family was counted and cumulative distributions are presented separately for cyanobacterial and eukaryotic datasets in figure 14.

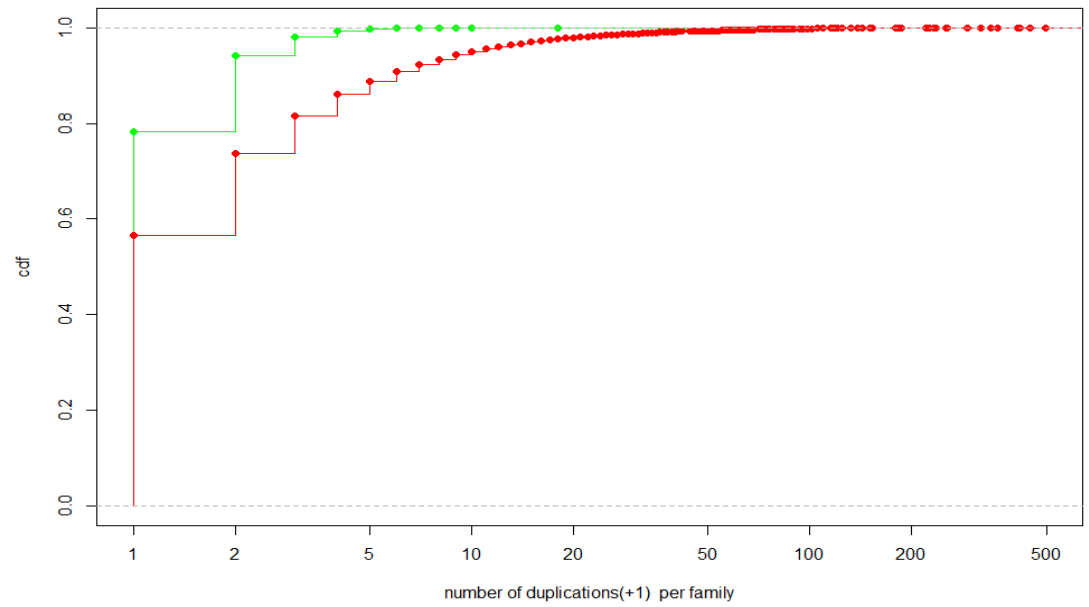


Figure 14: Inferred number of duplications (+1 so zero values can be shown on log scale) per gene family presented on log scale: red line for eukaryotic families and green line for cyanobacterial families.

4.5 Loss inference

Previously presented process of mapping duplication events to species tree branches allowed us to infer subsequent losses of gene copies assumed to exist in the ancestor. The fraction of species which lost one or both copies was calculated following each inferred duplication event. Distributions of losses after inferred duplications in each of the internal branches were presented in figures 15 and 16 for cyanobacterial and eukaryotic dataset respectively.

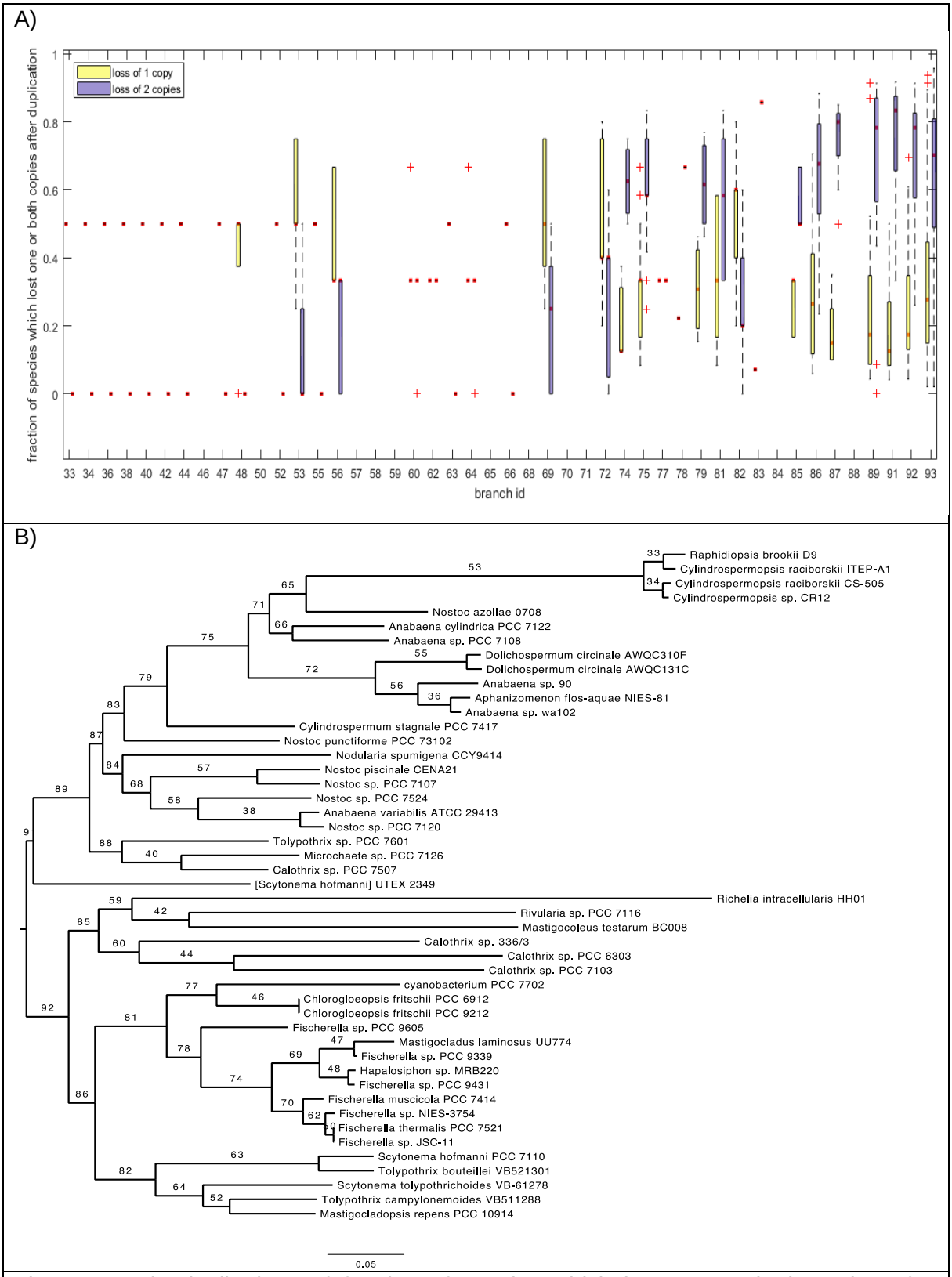
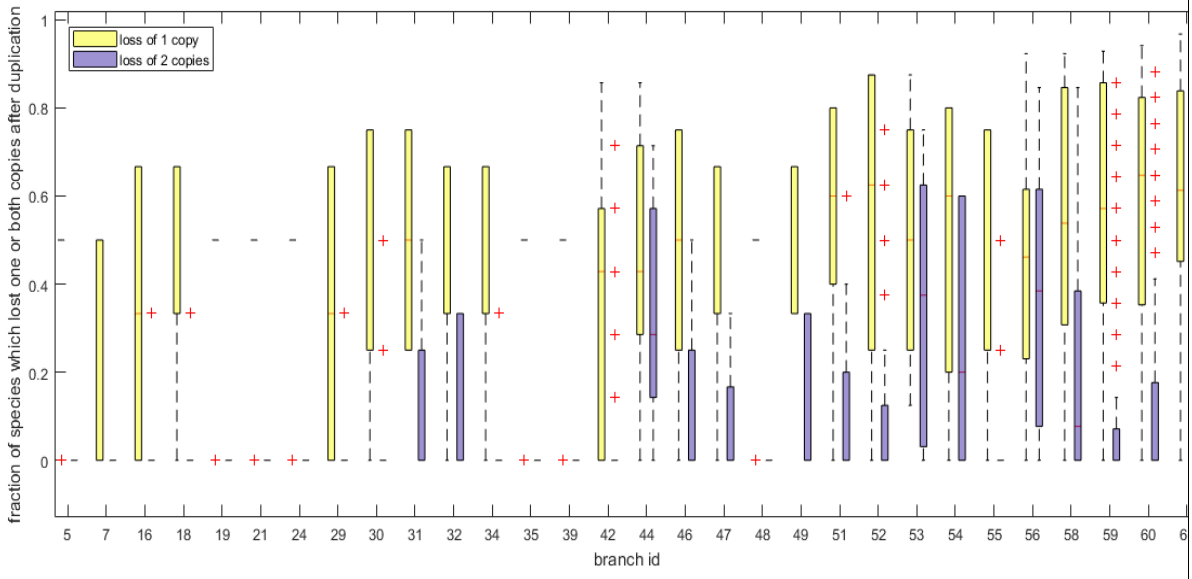


Figure 15: A) Distributions of fraction of species which lost one or both copies after duplications inferred in each of internal branches of cyanobacteria species tree: strong correlation between branch depth and frequency of loss of both copies (Spearman r=0.88, P=1.1x10-12), and negative correlation between branch depth and loss of 1 copy (Spearman r=-0.75, P=2.3x10-7)

B) cyanobacteria species tree with branch ids corresponding to boxplots in A)

A)



B)

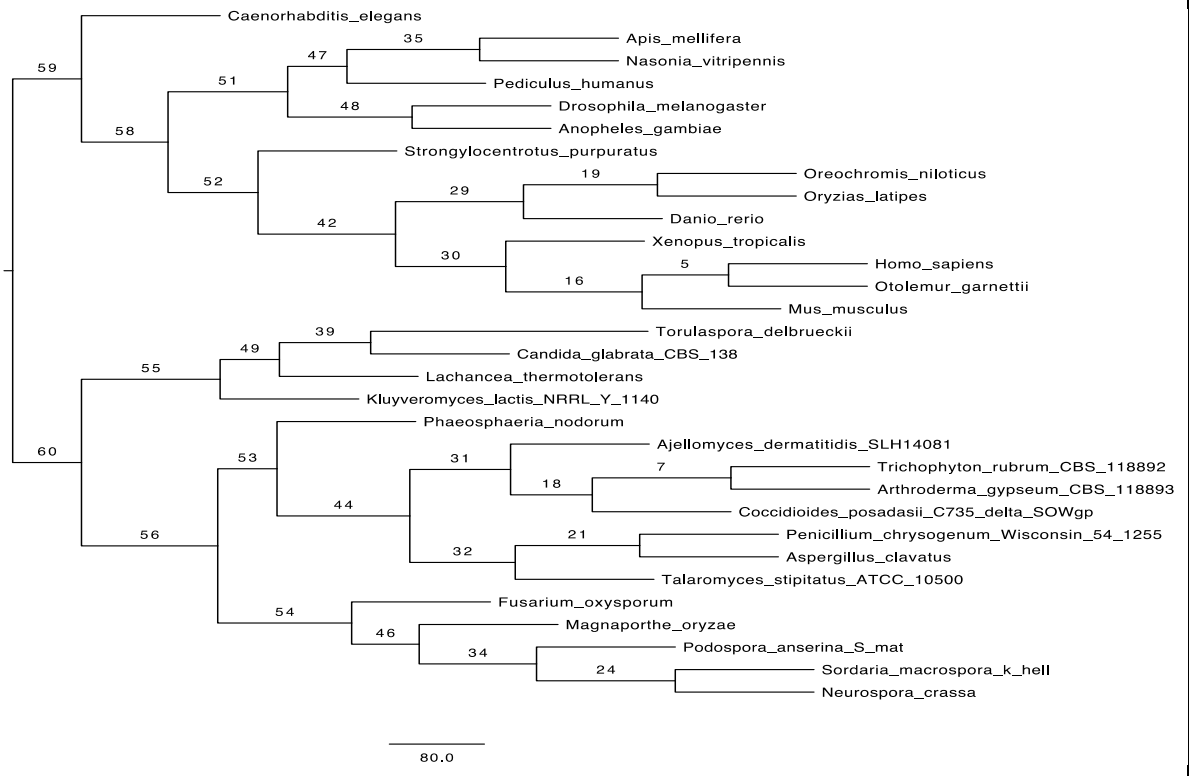


Figure 16. A) Distributions of fraction of species which lost one or both copies after duplications inferred in each of internal branches of eukaryotes species tree: no correlation between branch depth and amount of loss following duplication in a given branch (Spearman correlation was calculated separately for medians of distributions of fraction of species which lost 1 copy after duplication and separately for medians of distributions of fraction of species which lost both copies after duplication). Also, no correlation between number of duplications in branch and subsequent losses. B) eukaryotes species tree with branch ids corresponding to boxplots in A)

Analogously to loss inference after duplication, for eukaryotic partial single-copy families losses following gene family emergence were inferred. Separate distributions for each of the internal branches are presented in figure 17.

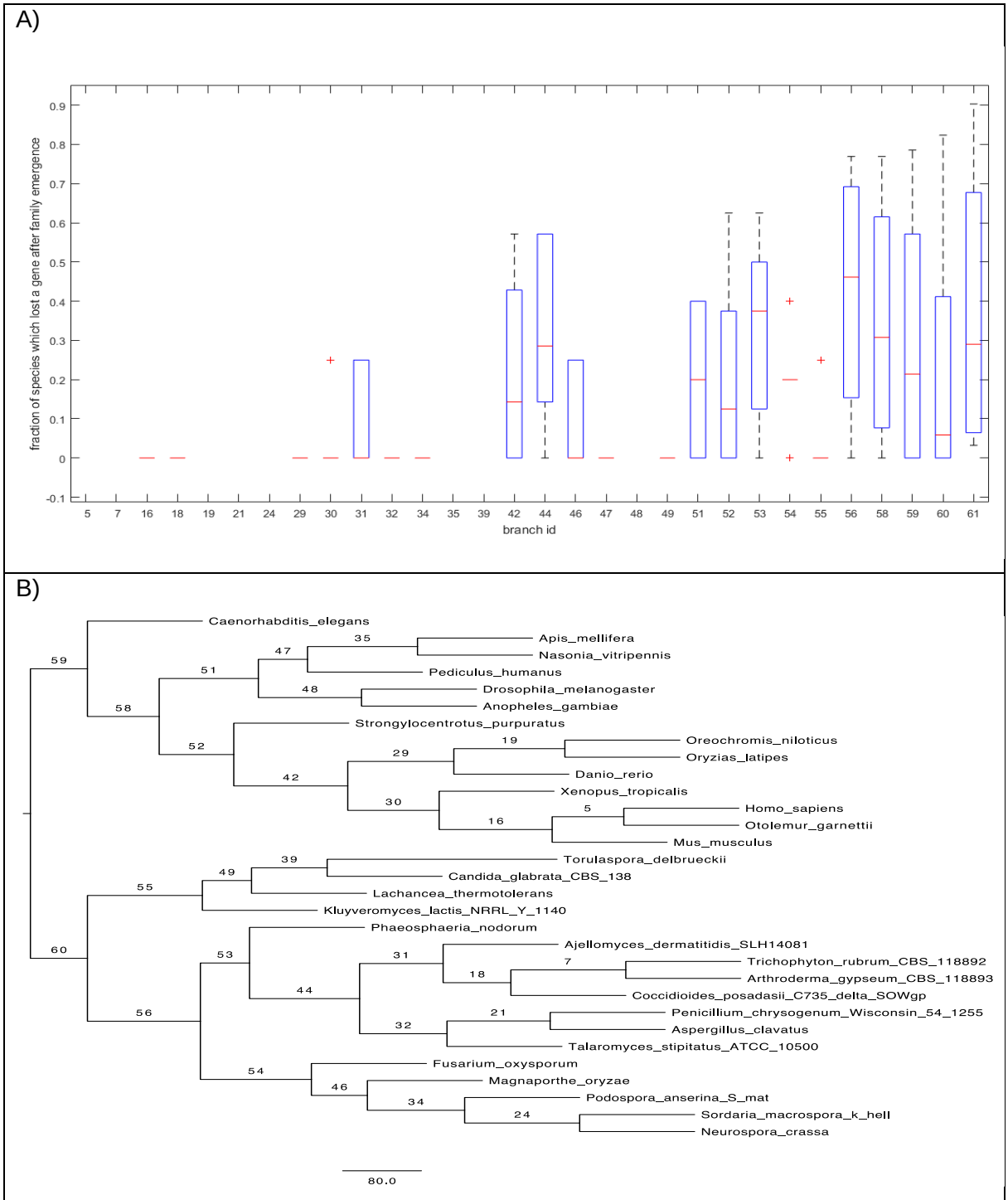


Figure 17. A) Distribution of losses after partial single-copy gene family emergence inferred in each of internal branches of eukaryotes species tree: weak correlation between number of emerging families in a branch and amount of loss following (Spearman r=0.5, P=0.018) B) eukaryotes species tree with branch ids corresponding to boxplots in A)

Differences in cumulative distributions of gene losses in cyanobacterial and eukaryotic species are presented in figures 18, 19 and 20.

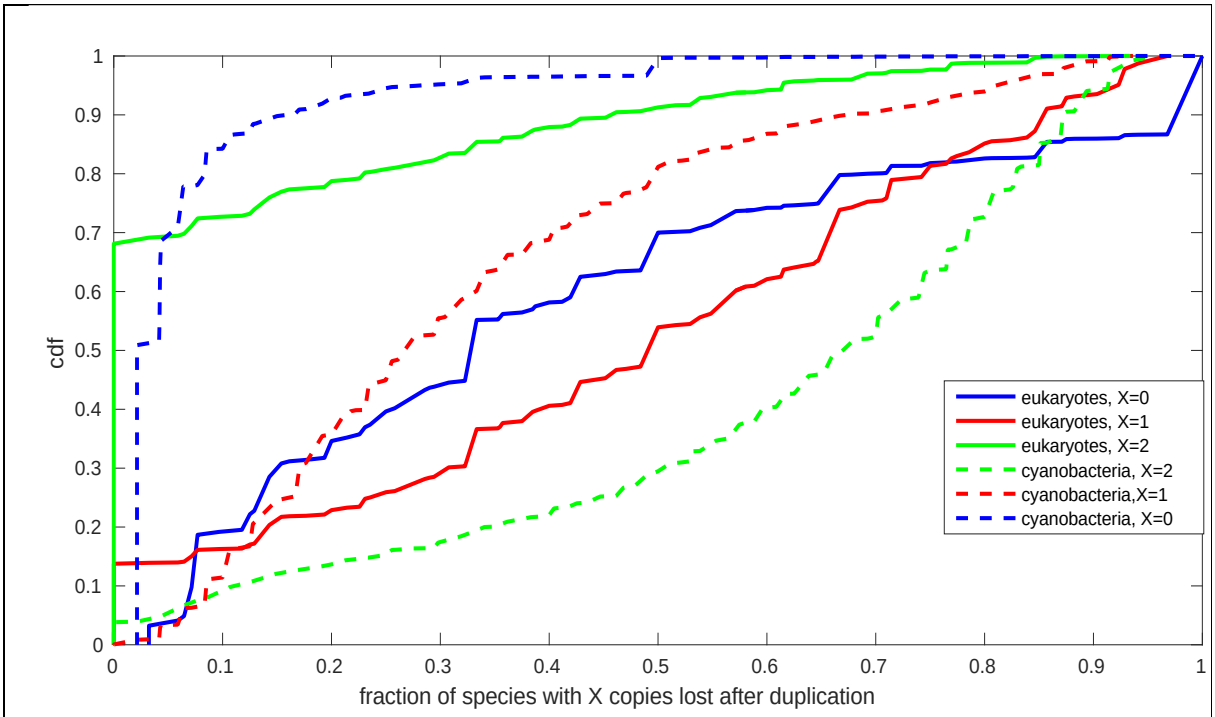


Figure 18. Cumulative distributions of inferred losses after inferred duplications in eukaryotic and cyanobacterial datasets

Line presenting loss of both gene copies after duplication in eukaryotes (full green line) in almost 70 % of cases shows value of 0 species. That means that for all species in almost 70 % of cases, after duplication remains at least one gene copy. It also means that median value for this distribution is zero because all species keep at least one gene copy in more than 50 % of cases. In contrast to that, the line showing the corresponding distribution for cyanobacteria (green dashed line) shows value zero in only about 3 % of cases of duplication which means in only about 3 % of cases, at least one of the copies is retained in all of the species in whose ancestor duplication occurred. Median value of this distribution is about 70 %, which means that in 50 % of cases up to 70 % of species lost both copies. Red lines represent loss of one copy after duplication and blue lines are presenting loss of zero copies after duplication, that is, proportion of species that retained both of the copies after duplication. Blue dashed line tells us that in almost 90 % of cases less than about 13 % of species retains both copies in cyanobacterial dataset and distribution reaches its maximum at 50 % of species.

Figures 19 and 20 are modified versions of figure 18. Figure 19 additionally shows the distribution of gene loss in partial single-copy families (purple dotted line). In figure 20 distributions of losses in eukaryotic dataset are separated for animals and fungi.

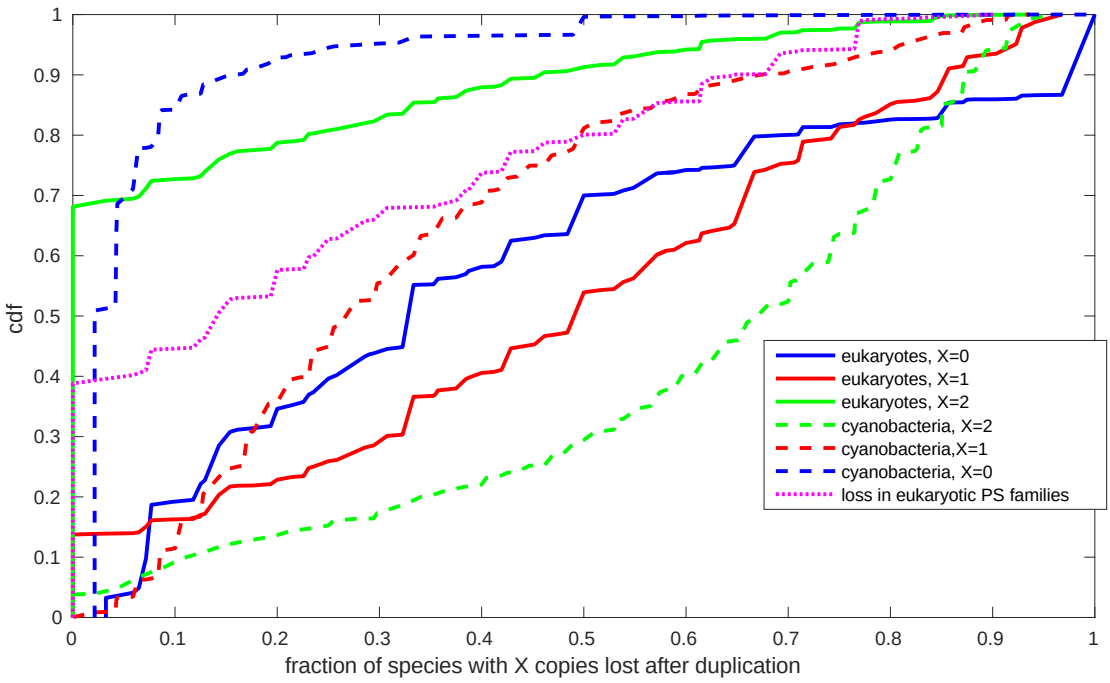


Figure 19 Cumulative distributions of inferred losses after inferred duplications in eukaryotic and cyanobacterial datasets and inferred losses after family emergences for eukaryotic partial single copy families

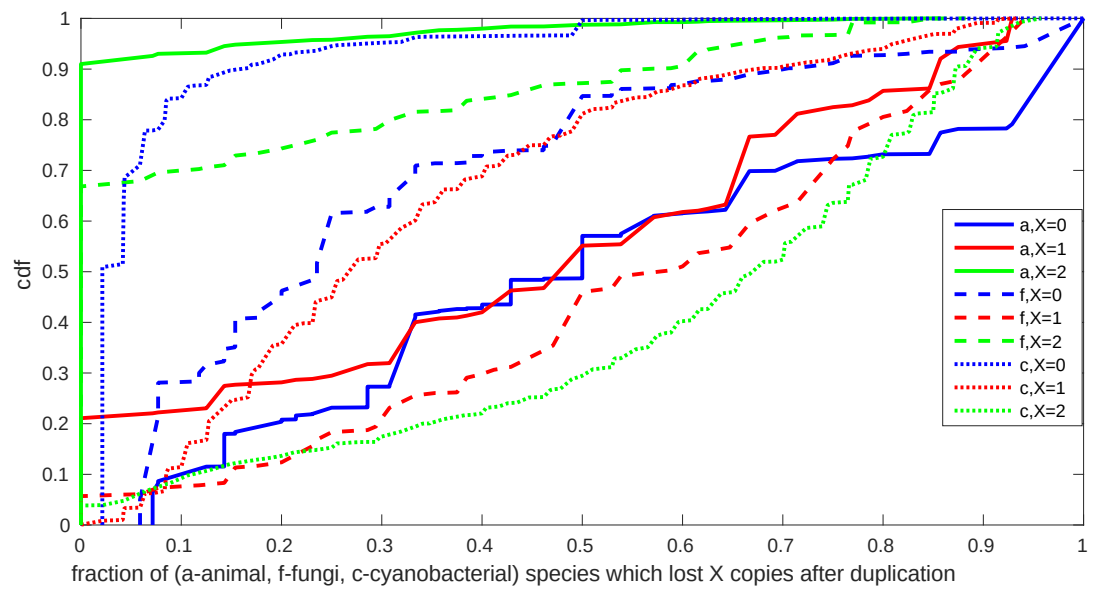


Figure 20 Cumulative distributions of inferred losses after inferred duplications in eukaryotic and cyanobacterial datasets, with separately counted losses for animal and fungi branches in eukaryotic dataset

5 Discussion

5.1 Simple algorithm for duplication/loss inference

In this thesis, we developed a simple algorithm for inferring duplication events, mapping them on a rooted species tree branches and inferring losses that followed an inferred duplication event. As an input, the algorithm takes a trusted inferred rooted species tree and a set of inferred rooted gene trees containing the species from the species tree. This algorithm treats all gene gain events as duplications and the consequences of it are gain events that are in reality HGT mapped wrongly to the deeper branch in the species tree and subsequent overestimation of losses. This is because HGT between left and right part of a clade would look like duplication event in the clade's LCA and eventually we would falsely infer losses. The algorithm was applied to two datasets: eukaryotic and prokaryotic one.

5.2 Mapping duplications to the species trees

The majority of duplications in cyanobacterial gene families were mapped to the root of the species tree, that is, they seem to be ancestral to the clade of heterocystous cyanobacteria. Most of the other duplications were mapped along the branches leading to the species with the largest genomes. We can hypothesize that rooting branch is such a major outlier in terms of number of inferred duplications per branch because of three possible reasons. The first one would be HGT (for prokaryotes normal form of gene gain) between left and right clade diverging from the root of the species tree. The second one would be uncertain position of the heterocystous cyanobacterial species tree root. Finally, the third reason could be the great number of duplications that are truly ancestral to divergence of heterocystous cyanobacteria and which might have helped the development of their special features. It is reasonable to expect number of in-paralogs to correlate with the genome size (e.g. Larsson et al., 2011). Large genomes contain higher number of gene families than small genomes do. In this study, the number of duplications was inferred for each branch for all the gene families (e.g. figures 5, 6). While exhaustively searching for gene duplications in gene families, one will have higher probability to find more duplication events in ancestors of larger genomes, as it was here shown for Cyanobacteria. If a genome contains more gene families it has higher probability to have duplications (in numerous families) inherited from a large ancestral genome. Defined from both angles, one could say that large genome tend to have more duplications, also meaning that higher number of duplications will lead to lager genomes. In summary, the inferred duplications for cyanobacterial dataset are an estimate that should be interpreted with caution. On the other hand, the distribution of inferred duplications along branches of

eukaryotic species tree does not have such extreme outliers. Additionally, we can take these estimates with greater reliability than for cyanobacterial dataset because there should be hardly any HGT to interfere with inference of in-genome duplications, moreover, the root of the eukaryotic species tree is unambiguous. Figure 7 shows estimates of “duplication rate” per branch, number of original number of inferred duplications per branch normalized by the branch length to obtain a perspective on how number of duplications per branch is influenced by a branch length. In both figures 6 and 7 we can observe duplication events are widespread in opisthokonta evolution with about 2500 duplications dated to the LCA. Interesting is the branch leading to *Pezizomycotina* species of fungi with 2467 duplications: these duplications seem to occur in rather large numbers compared to other fungi branches. It might represent whole genome duplication, genome hybridization or phase of genome expansion and these duplications might have been important for subsequent diversification of *Pezizomycotina* species. *Pezizomycotina* having much more duplications than *Saccharomycotina* could also be due to inheritance from the large LCA of the Fungi genomes, which could have had a lot of duplications. *Pezizomycotina* are morphologically more plesiomorphic than *Saccharomycotina* (yeasts). There are numerous unicellular yeast species, which represent rather derived fungal group. *Saccharomycotina* could, thus also have smaller genomes than *Pezizomycotina* due to numerous losses during their evolutionary history (e.g. Hibbet et al. 2007, Mohanta & Bae 2015). In the metazoan half of the tree duplication seems to be more prominent than in the fungi one: from LCA of all metazoan species to highest duplication rates in vertebrata. Branches leading to vertebrata and fish show high duplication rates and could represent 2R and 3R whole genome duplication events that occurred in vertebrata evolution (Graur, 2016). The branch leading to *Homo sapiens* has the highest duplication rate. This is in congruence with findings that gene turnover rate greatly accelerated in primates. It is even proposed duplications and losses are mainly responsible for morphological and behavioral differences between human and chimpanzee (Graur, 2016).

5.3 Mapping gene family emergence events to the species trees

Analogous algorithm to the one for mapping duplication events was used to map gene family emergences to corresponding branches in the species tree: the family was estimated to emerge in the LCA of all species in which was present. The proportion of family emergences in a branch is to some degree correlated with the number of duplications in it. This observation is expected since de novo gain provides material for further duplications and more duplications allow for more duplication-and-divergence scenarios of family emergence.

5.4 Gene family sizes

Inferred number of duplications per gene family was also counted and we can observe great variation of estimated numbers of duplications among families: from zero to almost 500 overall. This could be partly due to different selection pressures for different families, different mechanisms of gene amplification and a consequence of mostly stochastic interplay of duplication and loss. Eukaryotes have more gene families than prokaryotes, which is expected since they have larger genomes. In addition, eukaryotic gene families have more inferred duplication events per family. This could, to minor extent, be due to different clustering methods used for family inference in eukaryotic and cyanobacterial dataset. Another explanation for such observed difference could be eukaryotes having higher propensity for gene retention.

5.5 Loss patterns

After each of duplications mapped to internal branch, the proportion of species which lost none, one or both of copies inferred to exist in the ancestor was calculated. Figures 15 and 16 show distributions of losses after duplications inferred in each of the internal branches of cyanobacterial and eukaryotic species tree respectively. Cyanobacterial dataset exhibits a strong correlation between branch depth in a tree and the loss pattern, that is, deeper the branch, greater the fraction of species which lost both copies after inferred duplication and smaller the fraction of species that kept at least one copy. Eukaryotic dataset shows no such correlation. This could again be due to different HGT patterns in eukaryotes and prokaryotes. Inferred duplication events that are in fact HGT would increase the number of estimated losses due to wrongly placing “duplication” event to the LCA of a donor and a recipient. Because of this, it’s hard to infer a realistic loss pattern in prokaryotes. Eukaryotic loss pattern suggests prevalence of loss events along eukaryotic evolution although further research is needed for more precise estimates.

Figure 17 summarizes the differences in loss patterns between eukaryotes and cyanobacteria: cyanobacterial gene families display much higher loss frequencies. It is in a way a comparison of eukaryotic and prokaryotic evolution. Clearly, it’s not the same, and the main suspect in this case is HGT. Unfortunately HGT and its influence on the method prohibits us from estimating the real loss rates in prokaryotes. Figure 20 additionally shows different evolutionary patterns of animals, unicellular fungi and cyanobacteria.

5.6 Influence of the input tree quality on the algorithm

The influence of gene tree quality on the algorithm developed was also inspected and figures 21 and 22 (Supplementary material) suggest tree quality does not seem to influence the distribution of inferred duplications per branch, hence, all gene trees were used as an input for the algorithm developed.

6 Conclusion

1. In this thesis, we developed a simple algorithm for inferring duplication events, mapping them on a rooted species tree branches and inferring losses that followed an inferred duplication event. This algorithm treats all gene gain events as duplications but we are aware how presence of HGT would influence the results, so they can still be interpreted correctly. On the other hand, this algorithm's only assumptions are rooted species tree and rooted gene trees. This is an advantage in confront to other commonly used reconciliation methods that depend on more subjective assumptions.
2. Even with this simplistic approach we can observe very different ways of gene family evolution in prokaryotes and eukaryotes. In prokaryotes HGT rate is, it seems, significantly higher than in eukaryotes which results in higher loss frequencies and more deep duplications inferred. Since in eukaryotes (especially multicellular) HGT is effectively non-existing, results are more plausible. In eukaryotes duplication and loss are continuously occurring processes, and in prokaryotes duplication, loss and HGT are prevalent.

7 References

Albalat R, Cañestro C. 2016 Evolution by gene loss. Nat.Rev.Genet. 17(7):379-91

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990 Basic local alignment search tool. J.Mol.Biol. 215:403-410

Dagan T, Martin W. 2007 Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Procl.Natl.Acad.Sci.USA 104:870-875

Dagan T, Roettger M, Stucken K, Landan G, Koch R, Major P, Gould SB, Goremykin VV, Rippka R, Tandeau de Marsac N, Gugger M, Lockhart PJ, Allen JF, Brune I, Maus I, Pühler A, Martin WF.2013 Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids., Genome Biol Evol. 5(1):31-44.

Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, Goedert JJ, Buchbinder SP, Vittinghoff E, Gomperts E, Donfield S, Vlahov D, Kaslow R, Saah A, Rinaldo C, Detels R, O'Brien SJ. 1996 Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Science 273:1856-1862

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

Felsenstein J. Seattle (WA): University of Washington; 1993. PHYLIP (phylogeny inference package).

Graur D. 2016 Molecular and Genome Evolution. Sinauer, Sunderland

Guindon S, Gascuel O. 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52(5):696-704

Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PF, Eriksson, OE, Lumbsch HT. 2007 A higher-level phylogenetic classification of the Fungi. Mycological research, 111(5), 509-547

Hordijk W, Gascuel O. 2005 Improving the efficiency of SPR moves in Phylogenetic Tree Search Algorithms based on Maximum-Likelihood. Bioinformatics 21 (24): 4338-4347

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. 2016 eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44(D1):D286-93

Hughes AL, Nei M. 1989 Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci USA. 86:958-962.

Iranzo J, Cuesta JA, Manrubia S, Katsnelson MI, Koonin EV. 2017 Disentangling the effects of selection and loss bias on gene dynamics. Proc.Natl.Acad.Sci.USA 114(28):E5616-E5624

Kamneva OK, Ward N. 2014 Reconciliation Approaches to Determining HGT, Duplications, and Losses in Gene Trees. Meth Microbiol 41:183-199

Katoh K, Standley DM. 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772-80

Landan G, Graur D. 2007 Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24(6):1380-3

Larsson J, Nylander JA, Bergman B. 2011 Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. BMC Evol Biol. 11:187

Martin WF. 2017 Too much eukaryote LGT. Bioessays. 39(12)

Minh BQ, Nguyen MA, von Haeseler A. 2013 Ultrafast approximation for phylogenetic bootstrap. Mol.Biol.Evol. 5:1188-95

Mohanta TK, Bae H. 2015 The diversity of fungal genome. Biological procedures online, 17(1), 8.

Neme R, Amador C, Yildirim B, McConnell E, Tautz D. 2017 Random sequences are an abundant source of bioactive RNAs or peptides. Nat.Ecol.Evol. 1(6):0217

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol. 32(1):268-274

Ohno S. 1970 Evolution by Gene Duplication. Springer,New York

Ohno S. 1985 Dispensable genes. Trends Genet 1:160–164

Olson M V. 1999 When less is more: Gene loss as an engine of evolutionary change. Am.J.Hum.Genet. 64:18-23

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276-277

Shimodaira H, Hasegawa M. 1999 Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116

Tatusov RL, Koonin EV, Lipman DJ. 1997 A genomic perspective on protein families. Science 278:631-637

Tautz D, Domazet-Lošo T. 2011 The evolutionary origin of orphan genes. Nat.Rev.Genet. 12(10):692-702

Treangen TJ, Rocha EP. 2011 Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 7(1):e1001284

Tria FDK, Landan G, Dagan T. 2017. Phylogenetic rooting using minimal ancestor deviation. Nat.Ecol.Evol. 1, 0193.

Wolf YI, Koonin EV. 2013 Genome reduction as the dominant mode of evolution. Bioessays 35(9):829-37

Wolfe K. 2000 Robustness – it’s not where you think it is. Nature Genet. 25:3-4

8 Supplementary material

I. Figure 21. Effect of tree quality on inferred number of duplications per branch for cyanobacterial dataset
II. Figure 22. Effect of tree quality on inferred number of duplications per branch for eukaryotic dataset
III. Table 2. Genome sizes of species in cyanobacterial tree in megabases
IV. MATLAB code

I.

Effect of tree quality on inferred number of duplications per branch: it's important to note there is a vast difference between number of observations in cyanobacterial and eukaryotic datasets, and because of law of large numbers, we should trust eukaryotic dataset results more. This is also reflected in lower p-value for eukaryotic dataset.

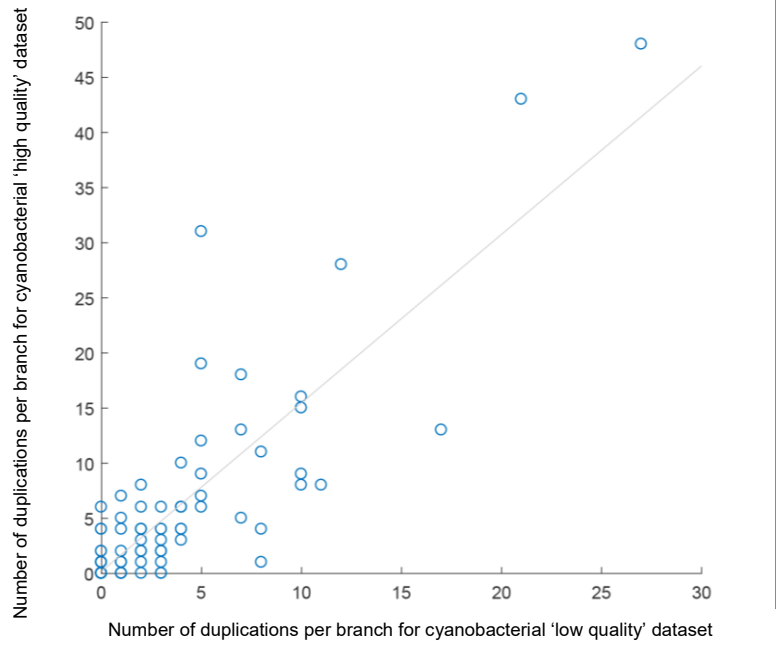


Figure 21. Effect of tree quality on inferred number of duplications per branch for cyanobacterial dataset

Spearman r= 0.77, P= 1.99×10^{-19} , rooting branch point (832,977) was excluded from the analysis because of high leverage

II.

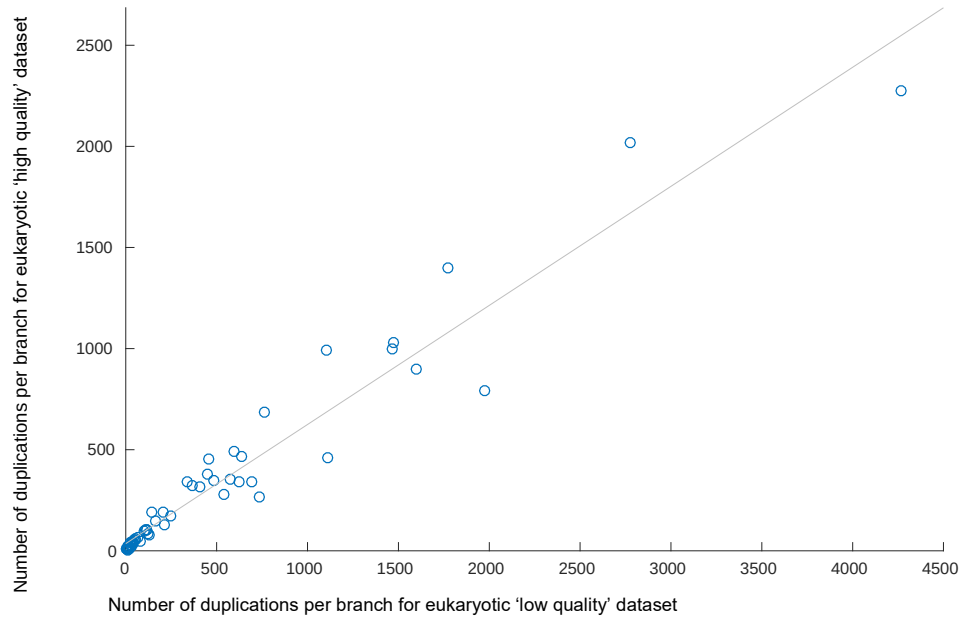


Figure 22. Effect of tree quality on inferred number of duplications per branch for eukaryotic dataset

Spearman r= 0.978, P= 7.389×10^{-42}

III.

Table 2: Genome sizes of species in cyanobacterial tree in megabases	
Organism	**Size (MB)**
Mastigocoleus testarum BC008	12.7002
Scytonema hofmanni PCC 7110	12.2944
Calothrix sp. PCC 7103	11.5844
Tolypothrix bouteillei VB521301	11.5723
Tolypothrix campylonemoides VB511288	10.6272
Scytonema tolypothrichoides VB-61278	10.0085
Tolypothrix sp. PCC 7601	9.97514
Nostoc punctiforme PCC 73102	9.05919
Rivularia sp. PCC 7116	8.72877
Mastigocladus laminosus UU774	8.56018
[Scytonema hofmanni] UTEX 2349	8.13344
Fischerella sp. PCC 9605	8.07918
Fischerella sp. PCC 9339	8.00826
Chlorogloeopsis fritschii PCC 6912	7.75174
Chlorogloeopsis fritschii PCC 9212	7.64985
Cylindrospermum stagnale PCC 7417	7.61059
Hapalosiphon sp. MRB220	7.42972
Nostoc sp. PCC 7120	7.21179
Fischerella sp. PCC 9431	7.16743
Anabaena variabilis ATCC 29413	7.10575
Nostoc piscinale CENA21	7.09456
Anabaena cylindrica PCC 7122	7.06328
Calothrix sp. PCC 7507	7.02322
Calothrix sp. PCC 6303	6.96039
Fischerella muscicola PCC 7414	6.90295
Nostoc sp. PCC 7524	6.71887
Mastigocladopsis repens PCC 10914	6.46565
Calothrix sp. 336/3	6.42013
Nostoc sp. PCC 7107	6.32982
Anabaena sp. PCC 7108	5.88674
Aphanizomenon flos-aquae NIES-81	5.85128
Fischerella sp. NIES-3754	5.82686
Anabaena sp. wa102	5.78203
Microchaete sp. PCC 7126	5.74226
'Nostoc azollae' 0708	5.48614
Nodularia spumigena CCY9414	5.46227
Fischerella thermalis PCC 7521	5.43827
Fischerella sp. JSC-11	5.38000
Anabaena sp. 90	5.30567
cyanobacterium PCC 7702	4.90315
Dolichospermum circinale AWQC131C	4.44565
Dolichospermum circinale AWQC310F	4.40585
Cylindrospermopsis raciborskii CS-505	3.87903
Cylindrospermopsis sp. CR12	3.72396
Cylindrospermopsis raciborskii ITEP-A1	3.60584
Richelia intracellularis HH01	3.24376
Raphidiopsis brookii D9	3.18651

IV.

```
s_tree=lineread('eu_tree.treefile');
spe_tree=newick(s_tree);
load('ids_eu.mat');
root=strcmp(group,'f');
sp_tree=tree2root(spe_tree,spe_tree.split(find(ismember(spe_tree.split,[root';~root'],'rows')),:));%roote
d species tree
species=str2double(regexp(sp_tree.ids,'^\d+','match','once'));
fungi=species(root);



n_nodes=sp_tree.root.full.n_nodes(1);
sp_brch_dupl=zeros(n_nodes,1);
fam_emergence=zeros(n_nodes,1);


list=dir('*.nwk'); %listing all gene trees
list(find([list.bytes]==0))=[];
dupl_per_tree=zeros(numel(list),1);
roott_AI=nan(numel(list),1);
tree_boot=nan(numel(list),1);
n_b_hits=0;
allone=0;
n_int=0;
n_ints=0;
allloss_0={};
allloss_1={};
allloss_2={};

losses_per_brch_1={};
losses_per_brch_2={};
for i=1:numel(list)
    [rooted_g,stats,info]=MAD(lineread(list(i).name));  %rooting gene trees
    rt=newick(rooted_g);
%     if stats.root_AI>0.95
%         continue
%     end
%     if nanmedian(rt.bsp)<0.5
%         continue
%     end

    roott_AI(i)=stats.root_AI;

    tree_boot(i)=nanmedian(rt.bsp);

    spid=str2double(regexp(rt.ids,'^\d+','match','once')); %species id-s
%     fam_num=fam_num+ismember(species,spid)';

    %%family emergence

%     if sum(sp_tree.root.split(sp_tree.root.n_branches,ismember(species,spid)))==0 ||
sum(sp_tree.root.split(sp_tree.root.n_branches,ismember(species,spid)))==numel(unique(spid)) %are
species on the same side of the root of the species tree
%                     spe_fam_clade=tree2clade(sp_tree,find(ismember(species,spid)));
%                     split_f=ismember(species,species(spe_fam_clade));
%                     branch_f=ismember(sp_tree.root.split, [split_f;~split_f], 'rows');
%
%                     if sum(branch_f)==2
```

```
%
spl=sp_tree.root.full.split(end,:)==repmat(sp_tree.root.full.split(end,sp_tree.root.full.brn2node(end,end)
),1,sp_tree.root.full.n_nodes(1));
%                       if split_f==spl(1:sp_tree.notu)
%                          branch_f=sp_tree.root.n_branches;
%                       else
%                          branch_f=sp_tree.root.n_branches-1;
%                       end
%                    end
%                    fam_emergence(branch_f)=fam_emergence(branch_f)+1;
%
%                 else
%                      fam_emergence(end)=fam_emergence(end)+1;
%    end
%




   losses_0=nan(rt.full.n_join,1);
   losses_1=nan(rt.full.n_join,1);
   losses_2=nan(rt.full.n_join,1);


   for j=1:rt.full.n_join %number of internal nodes/clustering steps in the tree



         lnid=rt.root.full.join_nodes(j,1);
         bid2=find(diff(rt.root.full.split(:,rt.root.full.join_nodes(j,1:2))')); %id-s of two branches where
inspected nodes are not together
         sp2=rt.root.full.split(bid2,:)==repmat(rt.root.full.split(bid2,lnid),1,rt.root.full.n_nodes(1)); %two
candidate branches encoded so that lnid is always 1
         m=sum(sp2,2);
         if m(1)>m(2)
            bid2=fliplr(bid2);
            sp2=flipud(sp2);

         end
         sp2(2,:)=~sp2(2,:);

         bb=(any(sp2(:,1:rt.notu)));
          b=find(ismember(rt.root.split,[bb;~bb], 'rows'));


         if isempty(b)
           bbb=1;
         else
         bbb=rt.root.bsp(b(1)); % bootstrap value of the branch
         end
         if isnan(bbb)
             bbb=1;
         end

         %%%%%%%%%%
         if bbb>=0.8
         left=spid(sp2(1,1:rt.notu));
         right=spid(sp2(2,1:rt.notu));
         if numel(intersect(left,right))>0
            int=intersect(left,right);
```

```
            n_int=n_int+1;

            if numel(union(left,right))>1
               n_ints=n_ints+1;



                if
sum(sp_tree.root.split(sp_tree.root.n_branches,ismember(species,union(left,right))))==0 ||
sum(sp_tree.root.split(sp_tree.root.n_branches,ismember(species,union(left,right))))==numel(union(lef
t,right)) %are species on the same side of the root of the species tree
                  spe_clade=tree2clade(sp_tree,find(ismember(species,union(left,right)))); %clade in the
sp_tree with species that are in the "symmetric" clade
                  split=ismember(species,species(spe_clade));
                  branch=ismember(sp_tree.root.split, [split;~split], 'rows');

                  if sum(branch)==2

spl=sp_tree.root.full.split(end,:)==repmat(sp_tree.root.full.split(end,sp_tree.root.full.brn2node(end,end)
),1,sp_tree.root.full.n_nodes(1));
                     if split==spl(1:sp_tree.notu)
                        branch=sp_tree.root.n_branches;
                     else
                        branch=sp_tree.root.n_branches-1;
                     end
                  else
                     branch=find(branch);
                  end
                  dupl_per_tree(i)=dupl_per_tree(i)+1;
                   sp_brch_dupl(branch)=sp_brch_dupl(branch)+1;


losses_per_brch_1{branch,(end+1)}=sum(ismember(species(spe_clade),setdiff([left,right],int)))/numel(
spe_clade);

losses_per_brch_2{branch,(end+1)}=sum(~ismember(species(spe_clade),[left,right]))/numel(spe_clad
e);
                  losses_0(j)=numel(int)/numel(spe_clade);

losses_1(j)=sum(ismember(species(spe_clade),setdiff([left,right],int)))/numel(spe_clade);
                  losses_2(j)=sum(~ismember(species(spe_clade),[left,right]))/numel(spe_clade);

                    %number of losses this duplication assumes

                else
                     dupl_per_tree(i)=dupl_per_tree(i)+1;
                     losses_0(j)=numel(int)/numel(species);
                     losses_1(j)=sum(ismember(species,setdiff([left,right],int)))/numel(species);
                     losses_2(j)=sum(~ismember(species,[left,right]))/numel(species);
                     sp_brch_dupl(end)=sp_brch_dupl(end)+1;


losses_per_brch_1{n_nodes,(end+1)}=sum(ismember(species,setdiff([left,right],int)))/numel(species);

losses_per_brch_2{n_nodes,(end+1)}=sum(~ismember(species,[left,right]))/numel(species);
                    %number of losses this duplication assumes


                end %if species are on the same side of the root

%                  end %ancestral bcrh
```

```
            else

                dupl_per_tree(i)=dupl_per_tree(i)+1; %num of duplication events per gene tree
                split=ismember(species,int);
                branch=ismember(sp_tree.root.split, [split;~split], 'rows');
                sp_brch_dupl(branch)=sp_brch_dupl(branch)+1;

                end %if union>1
          end %if int>0
        end%bbb>0.80
    end %j, iterating internal branches
    allloss_0{i}=losses_0;
    allloss_1{i}=losses_1;
    allloss_2{i}=losses_2;

end
losses_0=cat(1,allloss_0{:});
losses_0(isnan(losses_0))=[];
losses_1=cat(1,allloss_1{:});
losses_1(isnan(losses_1))=[];
losses_2=cat(1,allloss_2{:});
losses_2(isnan(losses_2))=[];
```

Curriculum vitae

I was born on 8th of September in 1992, in a beautiful town of Split, Croatia. Love for science has brought me to Zagreb where I obtained bachelor's degree in molecular biology from Department of Biology, Faculty of Science, University of Zagreb in 2015. During my bachelor studies I volunteered as an undergraduate teaching aid in Zoology course, participated several times in "Night of Biology" manifestation (science popularization event) and volunteered in several student associations. I was member of "Biology Students Association – BIUS" where I expanded knowledge about working in the field. In addition, I was member of another student association, "KSET", where I learned how to create video content. After obtaining bachelor's degree, I continued master studies in molecular biology at the same institution. Soon, I realized bioinformatics and molecular evolution are my biggest passions. To expand knowledge and gain practical experience in those fields, I took an internship in Prof. Dr. Tal Dagan's Genomic Microbiology Group, Institute of Microbiology, Christian-Albrechts-University Kiel, Germany. My work in Professor Dagan's group provided me with an excellent overview on how science works: I started with one question which opened another question and eventually it resulted in this master thesis. Striving for new experiences and skills, and always in need to challenge myself, I completed speleology training organized by Speleological department of Croatian Mountaineering Society "Željezničar" in 2018.