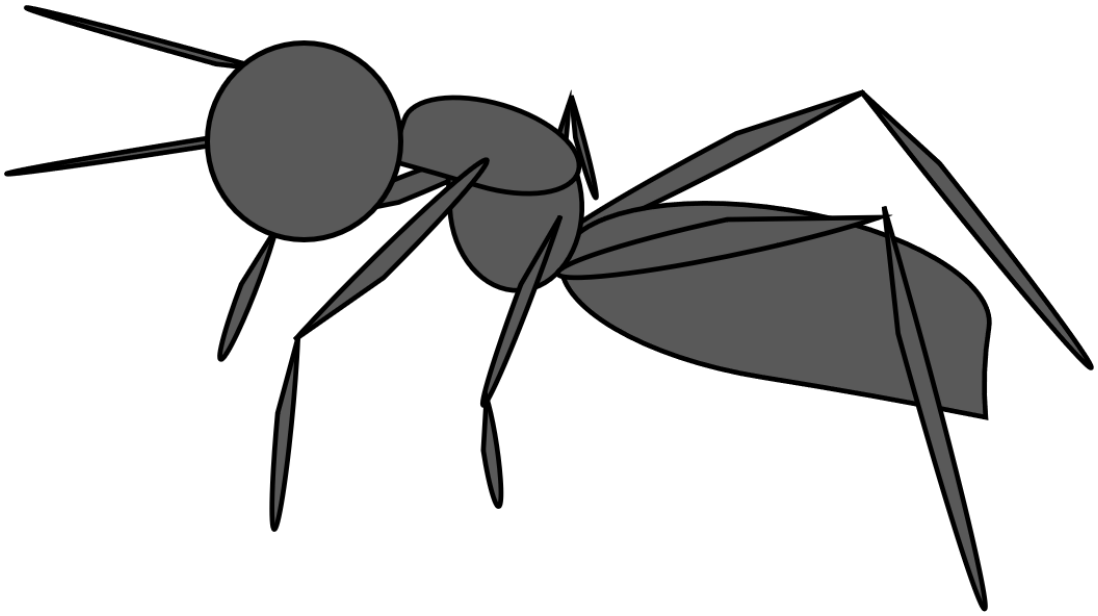
UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE



Project outside Course Scope MSc

Ants' UCEs Phylogeny

With focus on concatenated matrix and gene tree - species tree methods

Zelin Li

Advisor: Guojie Zhang and Joel Vizueta

January 21, 2022

This project has been submitted to Department of Biology, Faculty of Science, University of Copenhagen

Introduction

The ants have evolved a stunning global diversity with more than 15,000 extant species belonging to over 330 genera. Over their history, these colonial insects have been impressive innovators, evolving a huge diversity of advanced traits such as an advanced division of labor with specialized, morphologically distinct castes (queens, workers, and even soldiers). The GAGA project was launched in 2017 to generate and study high-resolution genomes for up to 200 ant species, covering most of the genomic and phenotypic diversity of the ants. Since then, 551 samples of 266 species from 130 genera were collected during the first phase of the project, resulting in the generation of 144 new genome assemblies. This dataset, together with the current publicly available genome data for ants, represents a total of 163 ants genome assemblies, which sets an unprecedented number of available genomes in ants, and even for any invertebrate family.

Ultra-conserved elements (UCEs) were originally defined as stretches of DNA with more than 200 base pairs that are identical, and with corresponding regions in two or more other sufficiently diverged vertebrate genomes[1]. In general, UCE is not necessarily to be a functional region in the genome, it can completely inside the non-coding region, in a gene, or as combination of multiple functional regions, for example the UCE-1 and UCE-2 in the Figure 1.

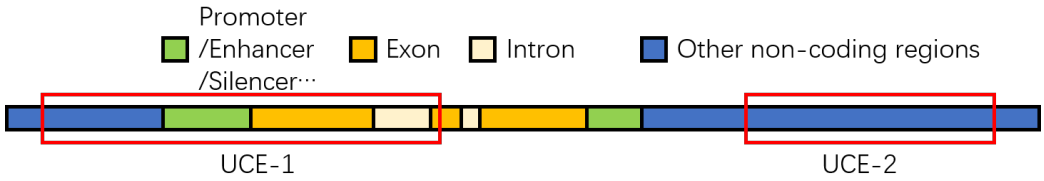


Figure 1: Example of genome regions that can be defined as UCEs

UCE definition and usability has expanded, and it has been described in many other taxa, including ants. In consequence, the targeted enrichment of UCEs has grown rapidly in popularity[2]. Specifically, an enhanced UCE bait set has been recently designed for the order *Hymenoptera* and especially for ants, which include 9446 baits, targeting 2524 conserved loci[3], and this bait set was used for the phylogenetic analysis on GAGA project's 163 ants.

Here, we are using complete ant genomes and the *Hymenoptera* UCE baits to extract the UCE regions in order to reconstruct the phylogenetic relationships of the sequenced ant species in the GAGA project. In this project, we have 163 ants genomes assemblies; most of them sequenced with long-read technologies (PacBio and stLFR) and some assembled at the chromosome level (best case would be chromosome level). The assemblies sources can be seen in Figure 2.

Methods

The code generated under this project were uploaded in a github repository[1] (the pipeline is not fully automated).

[1]GitHub: GAGA pipeline

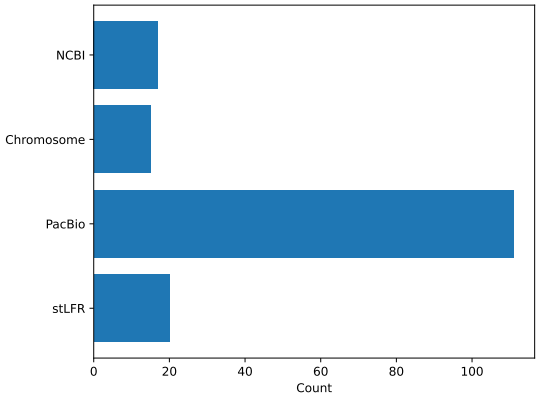


Figure 2: Different level's genome assemblies. NCBI: genomes from NCBI database; Chromosome: chromosome level's assemblies; PacBio and stLFR: third-generation sequencing long reads' assemblies

To conduct ants' UCEs phylogeny research, I used the phyluce pipeline to extract the UCEs from genomes[4, 5] by probing on the hymenoptera-v2-ANT-SPECIFIC-uce-baits[3]. In the phyluce pipeline, MAFFT[6] was used to align each set of UCEs, then the alignments were trimmed (first internal trimming, then gblocks trimming[7]), and 2515 UCEs' (trimmed or untrimmed) alignments files (in nexus format) were generated.

Notice that I also try to use the alignments without trimming to construct phylogenetic trees, the results are better than trimmed alignments' result for gene tree - species tree method but are worse in concatenated matrix method, this fact can be explained because: For the gene tree - species tree method, it construct individual tree for each UCE, while trimmed alignments (especially for gblocks trimming) may delete a great amount of gaps which are useful for genetics distance calculation. For the concatenated matrix method, it consumes big matrix for overall genetics distance calculation, so that trimming would be essential to eliminate or reduce the noise.

Based on the alignments, two methods were used for further analysis:

Concatenated matrix: This approeach consists in concatenating all UCE alignments following phyluce pipeline: First, we created three matrix containing all UCE present in more that 75%, 90% and 95% of the species. Then concatenate the remaining alignments by building the concatenated data matrix (one file for all UCEs). We use this matrix to construct phylogenetic tree with IQ-TREE[8], its configuration is:

```
iqtree2 -s $in --prefix tree${percent}p -B 1000 -alrt 1000 -m MFP -T 40 --
    safe
```

In this configuration, two bootstrapping methods were used: ultrafast bootstrap approximation (parameter `-B`) and SH-aLRT (SH-like approximate likelihood ratio test, parameter `-alrt`). Both methods' bootstrap replicates were set to 1000, and this is the minimum requirement for SH-aLRT. Ultrafast bootstrap was used because it is fast[9] and it was planned to use on the gene tree - species tree methods too, thus it is convenient to compare the results across methods. SH-aLRT was used is because it is based on the log ratio between the likelihood value of the current tree and that of the best alternative[10]. The MFP, short for ModelFinder Plus, tells IQ-TREE to compute the log-likelihoods of an initial parsimony tree for many different models, then it

chooses the model that minimizes the BIC (Bayesian information criterion) score, and the remaining analysis will use the selected model. `-T` specify the threads and `--safe` activate safe likelihood kernel to avoid numerical underflow.

Gene tree - species tree: This approach consists in reconstructing the gene tree for each UCE loci separately, and then combine all trees to reconstruct the species tree. Instead of concatenate the UCEs, this method consider each UCE as a 'gene', then construct phylogenetic tree of each UCE separately, and in the end combine all the trees to one final tree. Starting with the untrimmed alignments, I converted the nexus files to single line sequence fasta files, then each UCE tree was constructed by IQ-TREE with configuration similar to concatenated matrix method but without the -alrt 1000, because downstream analysis may have error due to more than one bootstrap confidence, only using the -B is because it's faster, and -T was set to 1 for parallel scheduling system. After that, 2510 trees were constructed (5 UCEs have less than 4 species, so they were not able to bootstrap and generate tree); I filtered out the UCE trees with less than 10% (16), 25% (41), 50% (82), 75% (122), 90% (147) of species (total 163 is 100%), which were included and put into one newick file, then I filtered out the low support nodes (bootstrap support less than 10%, alternatively it could be less than 20% or more) in the trees by using newick utilities[11]. Finally, ASTRAL[12] was used to combine all trees with default configuration and the final tree's topology (branch length) was adjusted by using IQ-TREE with the fixed topology retrieved by ASTRAL (parameter `-g`) and concatenated matrix method's alignment (parameter `-s`).

Results and Discussion

We obtained the most supported topology by using the concatenated matrix approach. Three phylogenetic trees generated from this approch are shown in Figure 3. The best tree is Figure 3a, which filtered out the UCE's alignments with less than 75% of all 163 species included.

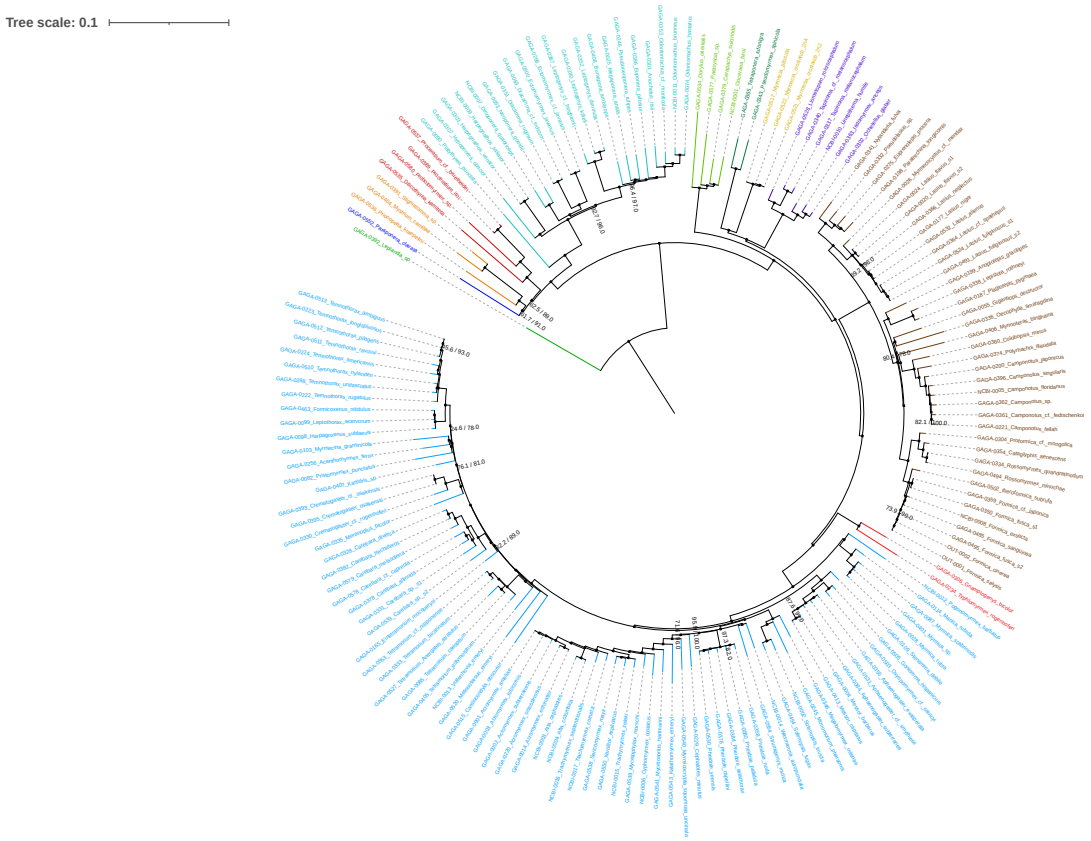
On the other hand, using the gene tree - species tree method, I have tested different filters on two steps, the first step is filtering (remove) the UCE trees with less than certain amount of species, the second step is filtering (remove) the nodes within the trees that have low support.

Fix node's filter as: ≤10% support

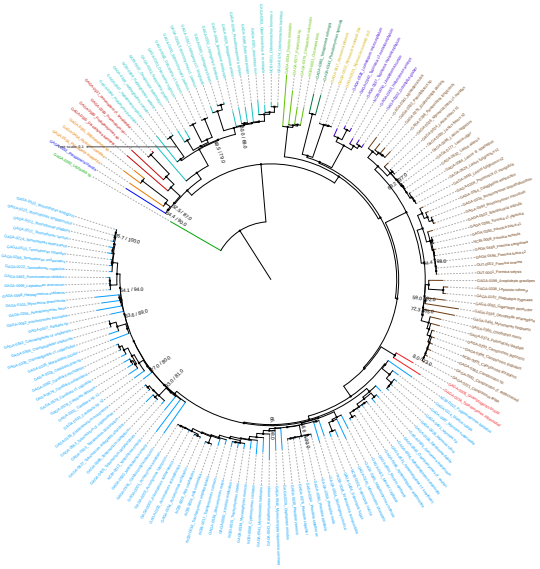
filter	count*	support sum**	support mean***
<16	21	1741%	82%
<41	22	1855%	84%
<82	23	1906%	82%
<122	21	1609%	76%
<147	23	1843%	80%

Table 1: Fix the node's filter to filter out nodes with support less or equal to 10%, then testing different UCE-tree filter (filter out trees with less than 16,41,82,122,147 species); *,**,***: support <100% node's count, support sum and support mean

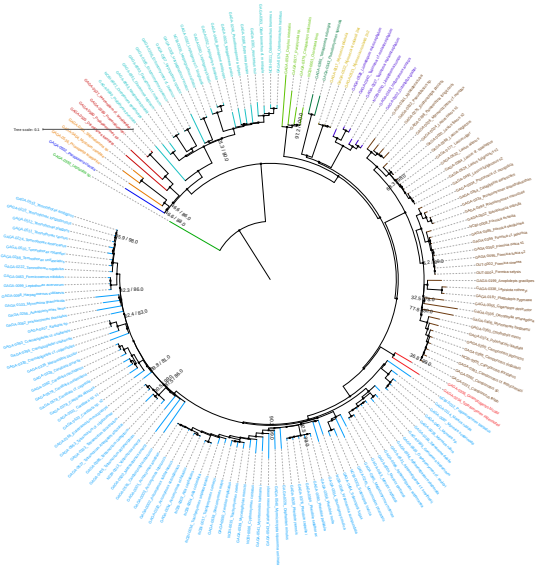
Therefore, I have done two experiment for finding the best filters of these two steps. The first experiment fix the node's filter and trying to find the best UCE-tree filter



(a) Filtered out UCE's alignments with less than 75% of (163) species



(b) Filtered out UCE's alignments with less than 90% of (163) species



(c) Filtered out UCE's alignments with less than 95% of (163) species

Figure 3: Three concatenated matrix method's trees; the trees here in the graph will only show the two bootstrap methods result when a node having ultrafast bootstrap support less than 98%, and for each node, the first support (left one) is ultrafast bootstrap, the second support is SH-aLRT

based on bootstrap support over the same retrieved topology, the result is shown in Table 1.

From the result, it is easily to figure out that the best UCE-tree filter is to filter out the trees with less than 41 species included. Thus, I set the second experiment to fix the UCE-tree filter as "<41" and trying to find the best filter, the result is shown in Table 2.

Fix UCE-tree filter as: <41 species

filter	count*	support sum**	support mean***
≤10%	22	1855%	84%
≤20%	22	1861%	84%
≤30%	21	1750%	83%
≤40%	22	1731%	78%

Table 2: Fix the UCE-tree filter to filter out trees with less than 41 species included, then testing on different node's filter (filter out trees' nodes with ≤10%,20%,30%,40% support); *,**,***: support <100% node's count, support sum and support mean

In conclusion, we found that for gene tree - species tree method, the best tree have filters of <41 species and ≤20% support nodes in UCE-trees. Moreover, the retrieved topologies were similar in both apporaches, therefore confirming the validity of the phylogenetic relationships retrievied here. By adjusting the tree's topology with the 75% concatenated matrix alignments and adding the ultrafast bootstrap support value to the nodes, I got the final version tree shown in Figure 4 (other retrieved topologies' trees are shown in Appendix Figure 5).

Finally, the obtained ant phylogeny here is comparable to a previous one using 2000 conserved genes from the *Hymenoptera* BUSCO dataset, therefore validating both approaches and providing a robust phylogeny for the ant species sequenced under GAGA project. The phylogeny would be valuable for downstream analyses that will shed light into understanding global trends in ant.

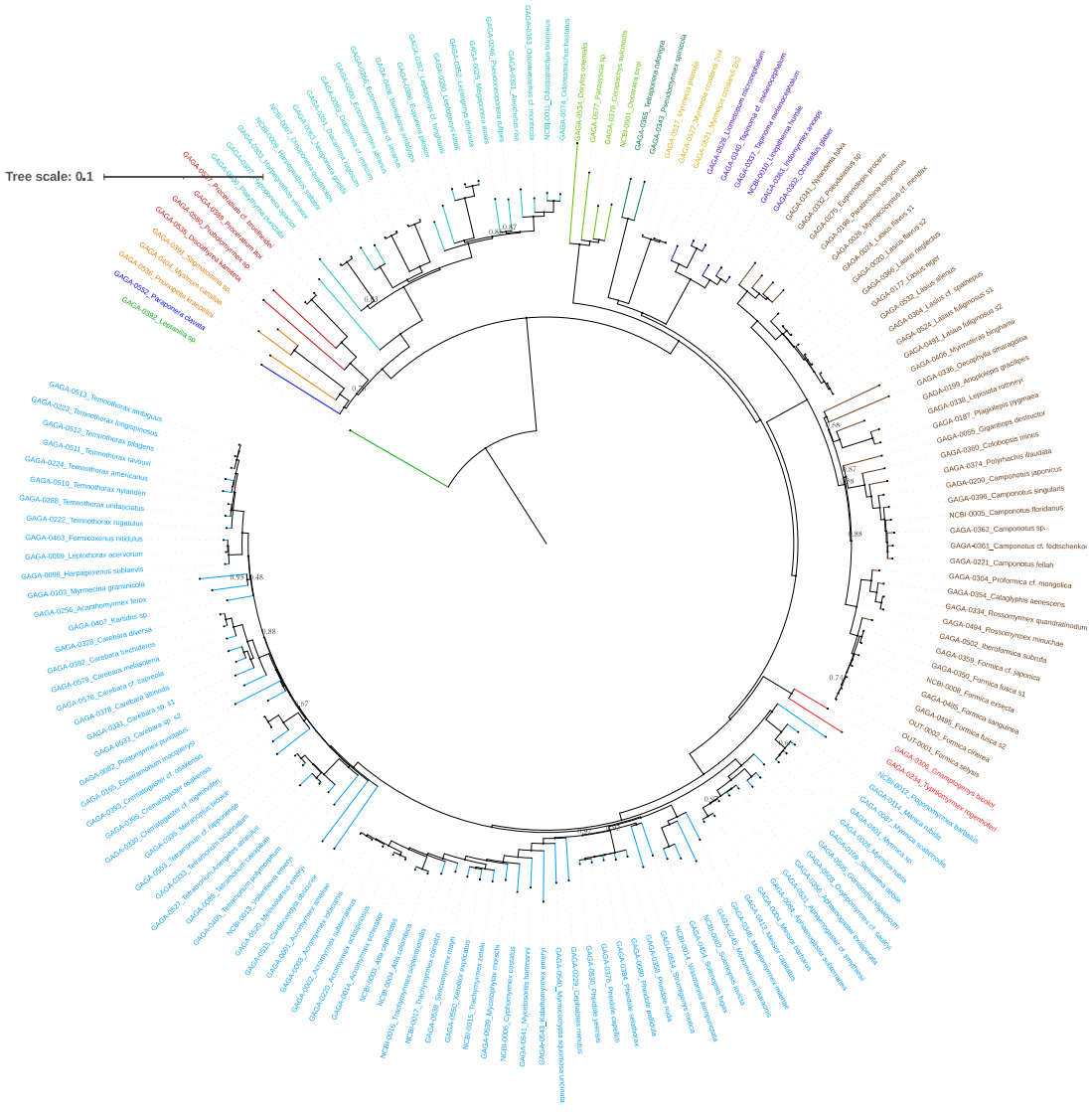


Figure 4: The final version of gene tree - species tree method's tree; it will only show the ultrafast bootstrap support that less than 98%

References

[1] Sol Katzman et al. "Human Genome Ultraconserved Elements Are Ultraselected". In: *Science* 317.5840 (2007), pp. 915–915. DOI: 10.1126/science.1142430.

[2] Brant C. Faircloth et al. "Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales". In: *Systematic Biology* 61.5 (Jan. 2012), pp. 717–726. ISSN: 1063-5157. DOI: 10.1093/sysbio/sys004.

[3] Michael G. Branstetter et al. "Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera". In: *Methods in Ecology and Evolution* 8.6 (2017), pp. 768–776. DOI: https://doi.org/10.1111/2041-210X.12742.

[4] Brant C. Faircloth. "PHYLUCE is a software package for the analysis of conserved genomic loci". In: *Bioinformatics* 32.5 (Nov. 2015), pp. 786–788. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv646.

[5] Brant C. Faircloth et al. "Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales". In: *Systematic Biology* 61.5 (Jan. 2012), pp. 717–726. ISSN: 1063-5157. DOI: 10.1093/sysbio/sys004.

[6] Kazutaka Katoh et al. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform". In: *Nucleic Acids Research* 30.14 (July 2002), pp. 3059–3066. ISSN: 0305-1048. DOI: 10.1093/nar/gkf436.

[7] J. Castresana. "Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis". In: *Molecular Biology and Evolution* 17.4 (Apr. 2000), pp. 540–552. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a026334.

[8] Bui Quang Minh et al. "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era". In: *Molecular Biology and Evolution* 37.5 (Feb. 2020), pp. 1530–1534. ISSN: 0737-4038. DOI: 10.1093/molbev/msaa015.

[9] Diep Thi Hoang et al. "UFBoot2: Improving the Ultrafast Bootstrap Approximation". In: *Molecular Biology and Evolution* 35.2 (Oct. 2017), pp. 518–522. ISSN: 0737-4038. DOI: 10.1093/molbev/msx281.

[10] Stéphane Guindon et al. "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0". In: *Systematic Biology* 59.3 (May 2010), pp. 307–321. ISSN: 1063-5157. DOI: 10.1093/sysbio/syq010.

[11] Thomas Junier and Evgeny M. Zdobnov. "The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell". In: *Bioinformatics* 26.13 (May 2010), pp. 1669–1670. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq243.

[12] Chao Zhang et al. "ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees". In: *BMC Bioinformatics* 19.6 (May 2010), p. 153. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2129-y.

Appendix

(a) 75%,16 (b) 90%,16 (c) 95%,16 (d) 75%,41

(e) 90%,41 (f) 95%,41 (g) 75%,82 (h) 90%,82

(i) 95%,82 (j) 75%,122 (k) 90%,122 (l) 95%,122

(m) 75%,147 (n) 90%,147 (o) 95%,147

Figure 5: 15 trees that from gene tree - species tree method. The percentage means: the genetics distance of the tree is calculated from matrix with alignments that having greater or equal to that % of total species; the number means: the topology of the tree is from the ASTRAL tree that is generated from UCE-trees which filtered out the nodes with $\leq$10% support and less than that amount of species