

# Supplementary Methods

## Detailed description of Read2Tree

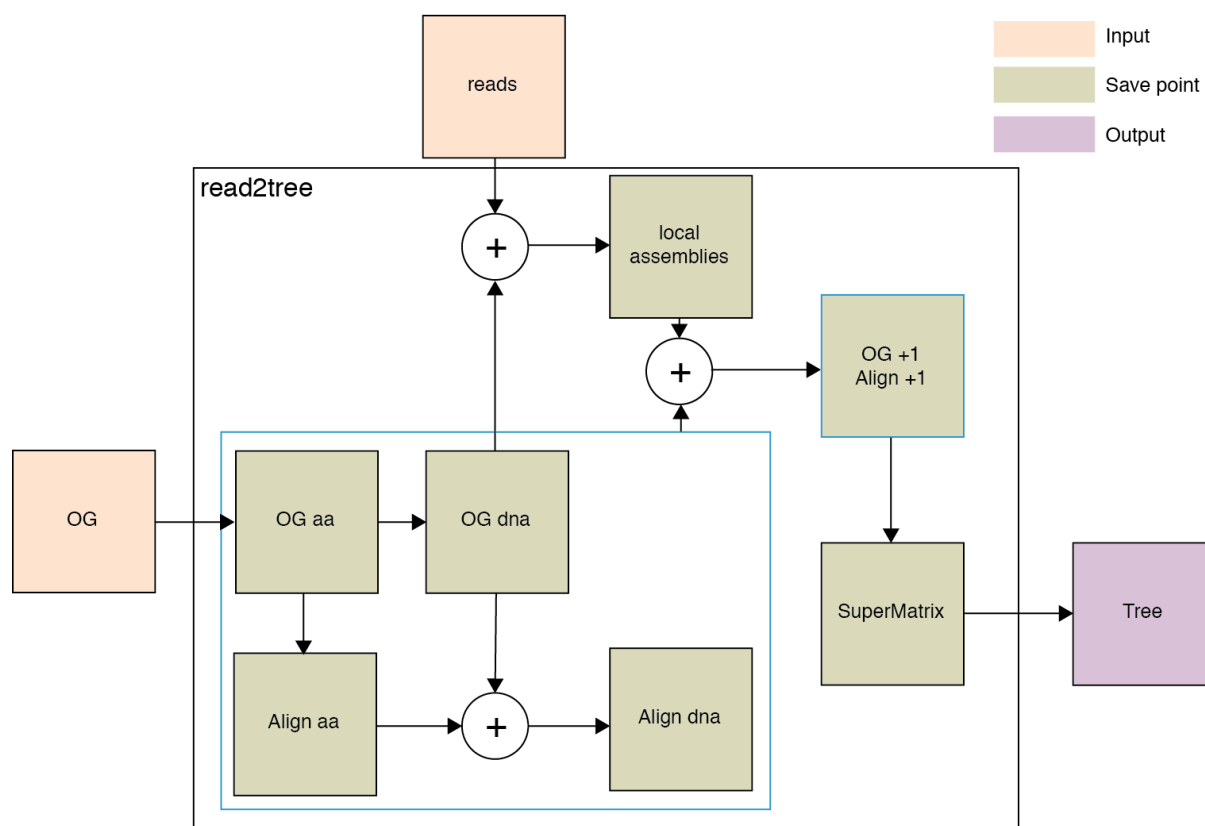
### **Data Inputs**

Read2Tree requires two sets of inputs: 1. A set of reads for a species of interest and 2. Reference orthologous groups.

### **Description of read2tree**

In a first step Read2Tree pulls the DNA sequences for the obtained OGs from the oma database using the available API and separates the OGs into individual folders. Then it collects for each species the DNA sequences and produces one file per species with the relevant sequences. These files are then used for the mapping. To allow for parallelization mapping can be performed on all species sequentially or set individually allowing to span multiple jobs per reference species on an HPC. Mapping of sequences is performed using NextGenMap (Sedlazeck et al., 2013) for short reads and NextGenMap-LR for long reads (Sedlazeck et al., 2018). Mapped reads are then post processed using samtools to extract the bases in regards to the sequences. Once the bases for a given reference sequence are extracted the consensus sequence is built based on the majority base at a position requiring at least 3 reads to be present. The reconstructed sequences are then placed back to the original OG and alignment. All alignments of OGs are then concatenated and result in the final supermatrix that serves as input for any tree inference method.

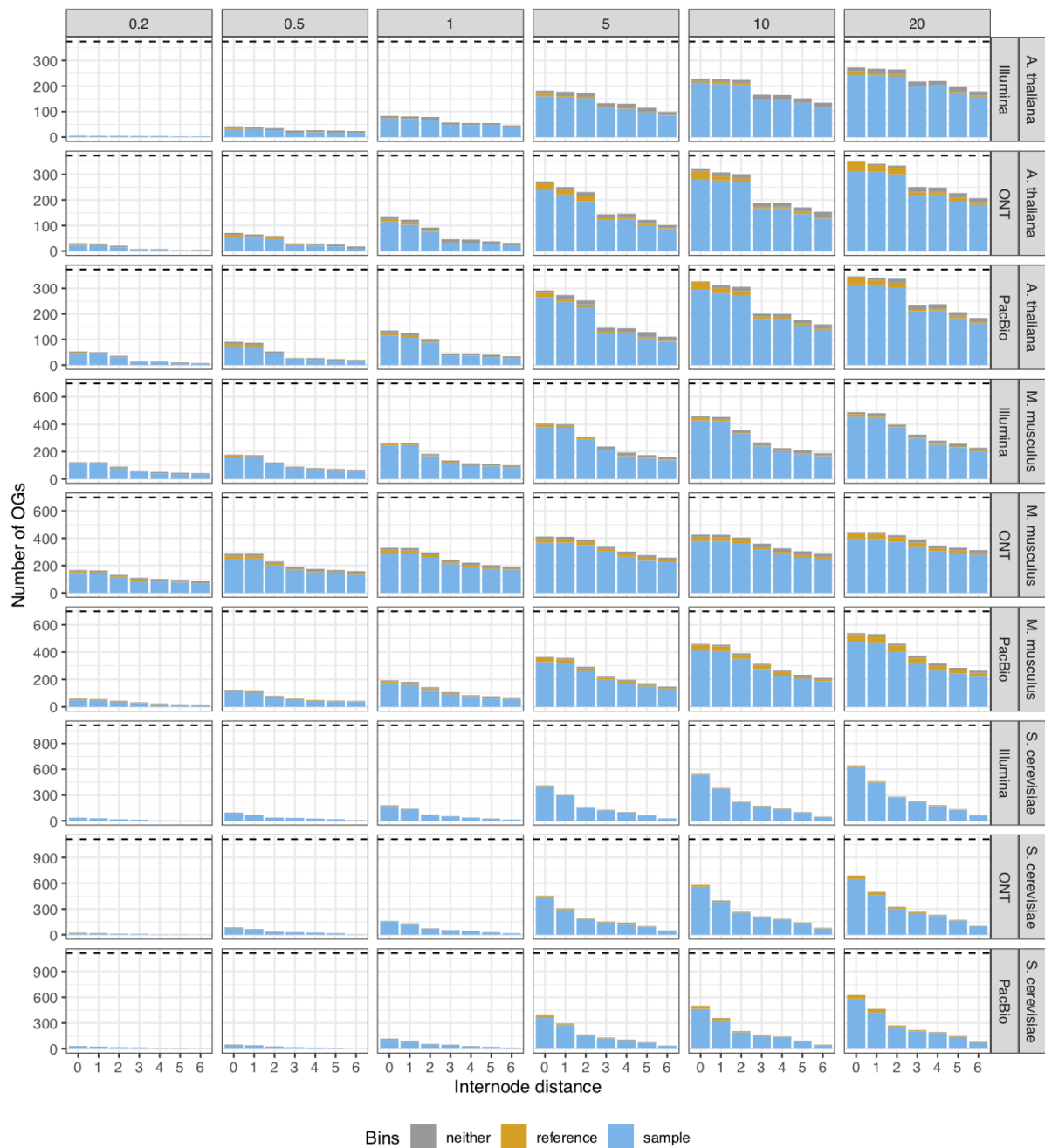
## Supplementary Figures



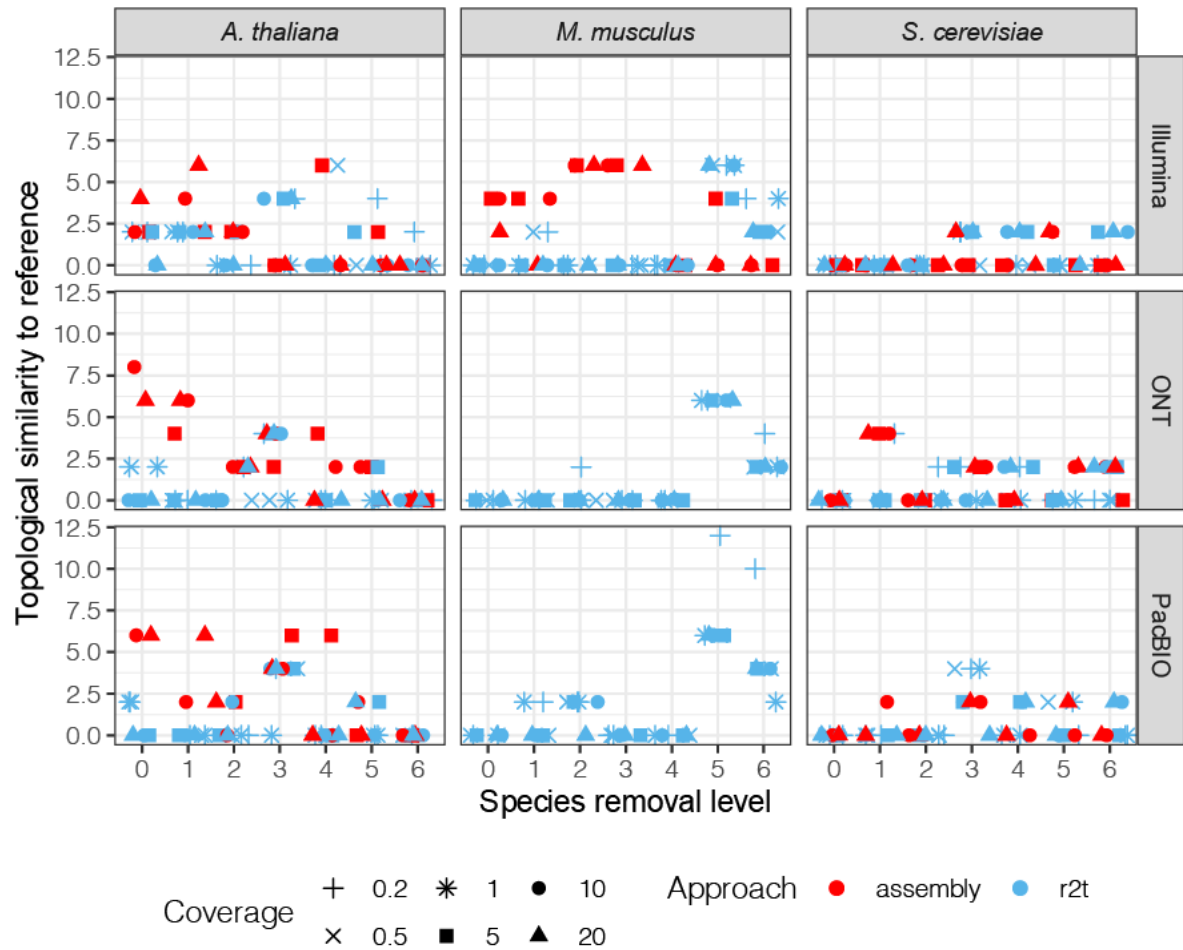
**Supplementary Figure 1.** Graphical representation of pipeline. All boxes in green are stored by Read2Tree. Inputs are reads and a set of reference orthologous groups that can be selected from over 2000 species from the OMA database. Local assemblies here as reconstructed sequences using the bases placed against the reference.

		RANDOM										COV										SC											
	90-100 -	0	0	0	0	0	0	0	0	4	134	-	0	0	0	0	5	0	0	0	6	140	-	0	0	0	0	5	0	0	0	6	138
	80-90 -	0	0	0	0	3	0	0	0	2	8	-	0	0	0	0	0	0	0	1	4	-	0	0	0	0	0	0	0	0	1	5	
	70-80 -	0	0	0	0	0	0	0	0	0	2	-	0	0	0	0	0	0	0	0	1	-	0	0	0	0	0	0	0	0	1	1	
	60-70 -	0	0	0	0	1	0	0	0	1	2	-	0	0	0	0	0	0	0	0	1	-	0	0	0	0	0	0	0	0	0	1	
	50-60 -	0	0	0	0	0	0	0	0	1	0	-	0	0	0	0	0	0	0	0	0	-	0	0	0	0	1	0	0	0	0	1	
	40-50 -	0	0	0	0	0	0	0	0	0	2	-	0	0	0	0	1	0	0	0	1	0	-	0	0	0	0	0	0	0	0	0	
	30-40 -	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	
	20-30 -	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	
	10-20 -	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	
	0-10 -	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	
BCN	90-100 -	0	0	0	0	1	3	0	0	5	125	-	0	0	0	0	2	6	0	1	5	127	-	0	0	0	0	8	1	0	1	8	123
	80-90 -	0	0	0	0	1	1	1	1	3	6	-	0	0	0	0	0	0	0	3	3	-	0	0	0	0	1	0	0	0	2	6	
	70-80 -	0	0	0	0	0	0	0	0	1	4	-	0	0	0	0	1	0	0	1	0	2	-	0	0	0	0	0	0	0	0	3	
	60-70 -	0	0	0	0	0	0	0	2	0	1	1	-	0	0	0	0	0	0	0	1	2	-	0	0	0	0	1	0	0	0	0	1
	50-60 -	0	0	0	0	0	0	0	0	1	0	1	-	0	0	0	0	2	0	0	0	1	-	0	0	0	0	1	0	0	0	0	3
	40-50 -	0	0	0	0	0	0	0	0	1	0	1	-	0	0	0	0	0	1	0	0	2	-	0	0	0	0	0	0	0	0	0	1
	30-40 -	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	
	20-30 -	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	
	10-20 -	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	
	0-10 -	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	
	0-10 -	10-20 -	20-30 -	30-40 -	40-50 -	50-60 -	60-70 -	70-80 -	80-90 -	90-100 -		0-10 -	10-20 -	20-30 -	30-40 -	40-50 -	50-60 -	60-70 -	70-80 -	80-90 -	90-100 -		0-10 -	10-20 -	20-30 -	30-40 -	40-50 -	50-60 -	60-70 -	70-80 -	80-90 -	90-100 -	
		bootstrap																															

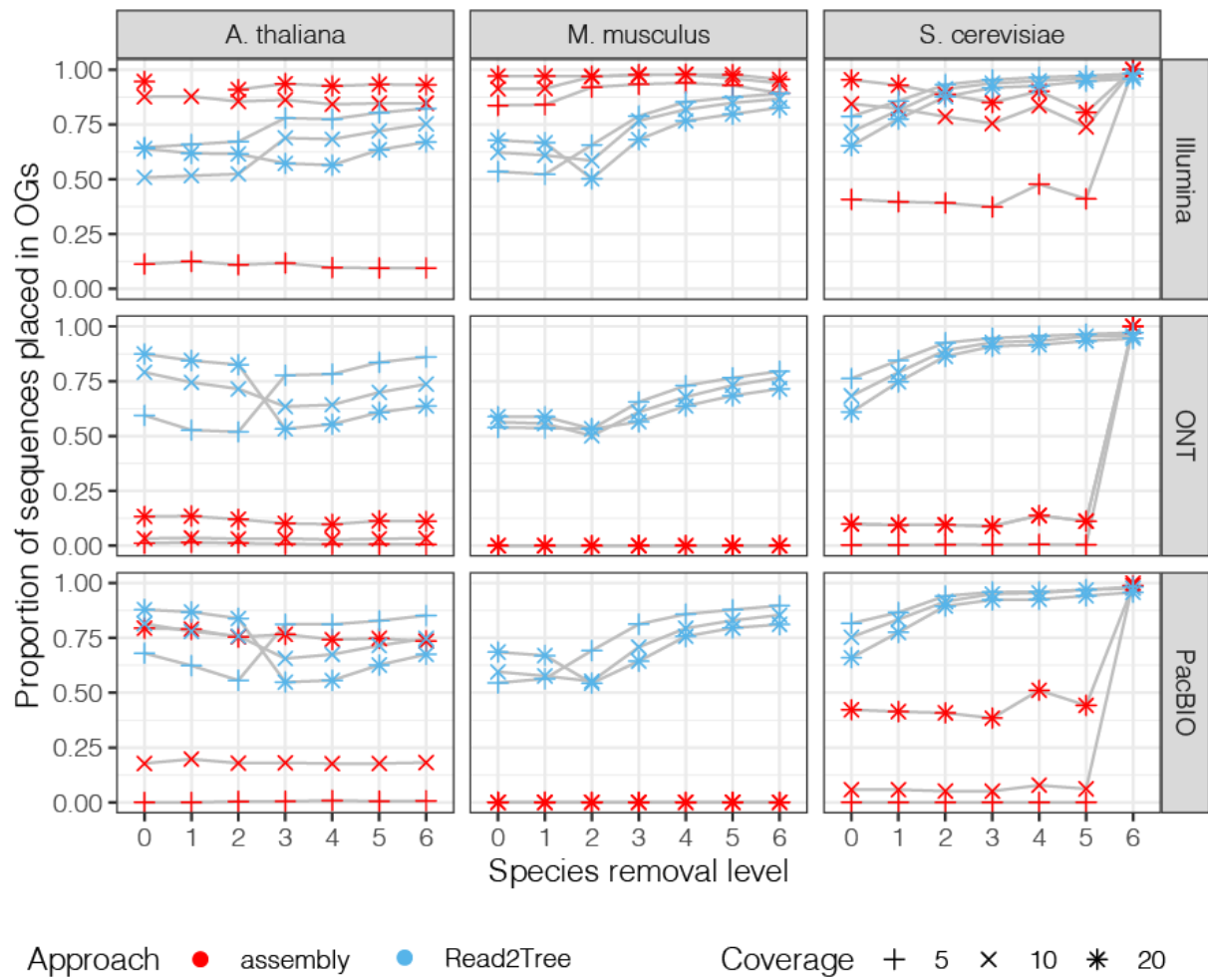
**Supplementary Figure 2.** Comparison of different gene selection methods in 20 times 8out/8in test show differences in relation between bootstrap on nodes of mapped species and its jaccard similarity with reference tree (Best Corresponding Node BCN) value. Top row shows values obtained using simulated reads and bottom row shows values obtained using real reads. In both cases we see that using coverage as a selection method shows the highest number of nodes where mapped species has high bootstrap and high BCN value.



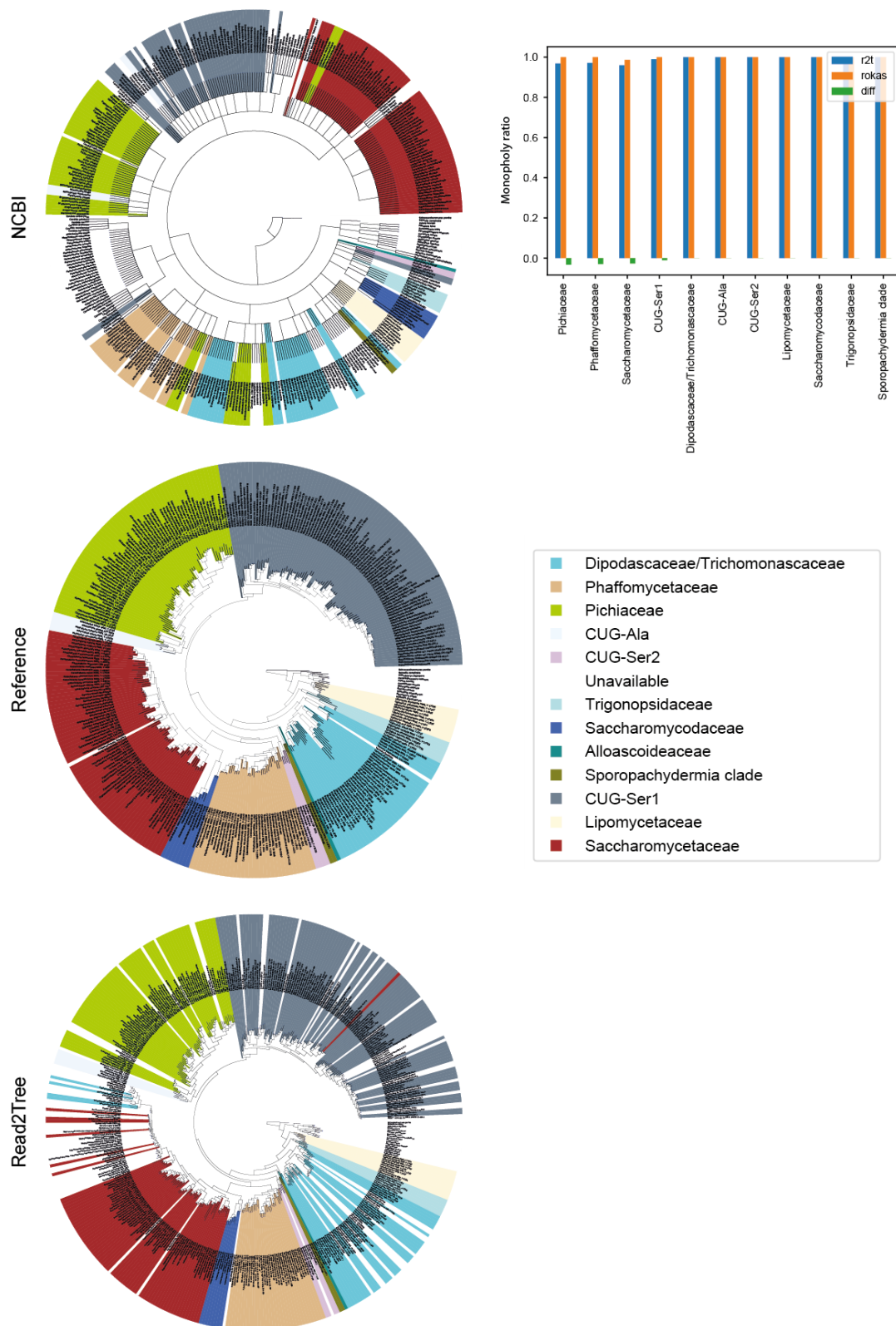
**Supplementary Figure 3.** Binning of top blastp results of r2t-sequence of selected species against their original OG (including the removed species) in either being most similar to its assembled counterpart (blue), to its reference used for reconstruction (yellow) or to any other sequence (grey). Results show that Read2Tree if reconstructing a sequence in most cases reconstructs a sequence that shows highest similarity to its assembled counterpart although this sequence was not present in the reference dataset when running Read2Tree.



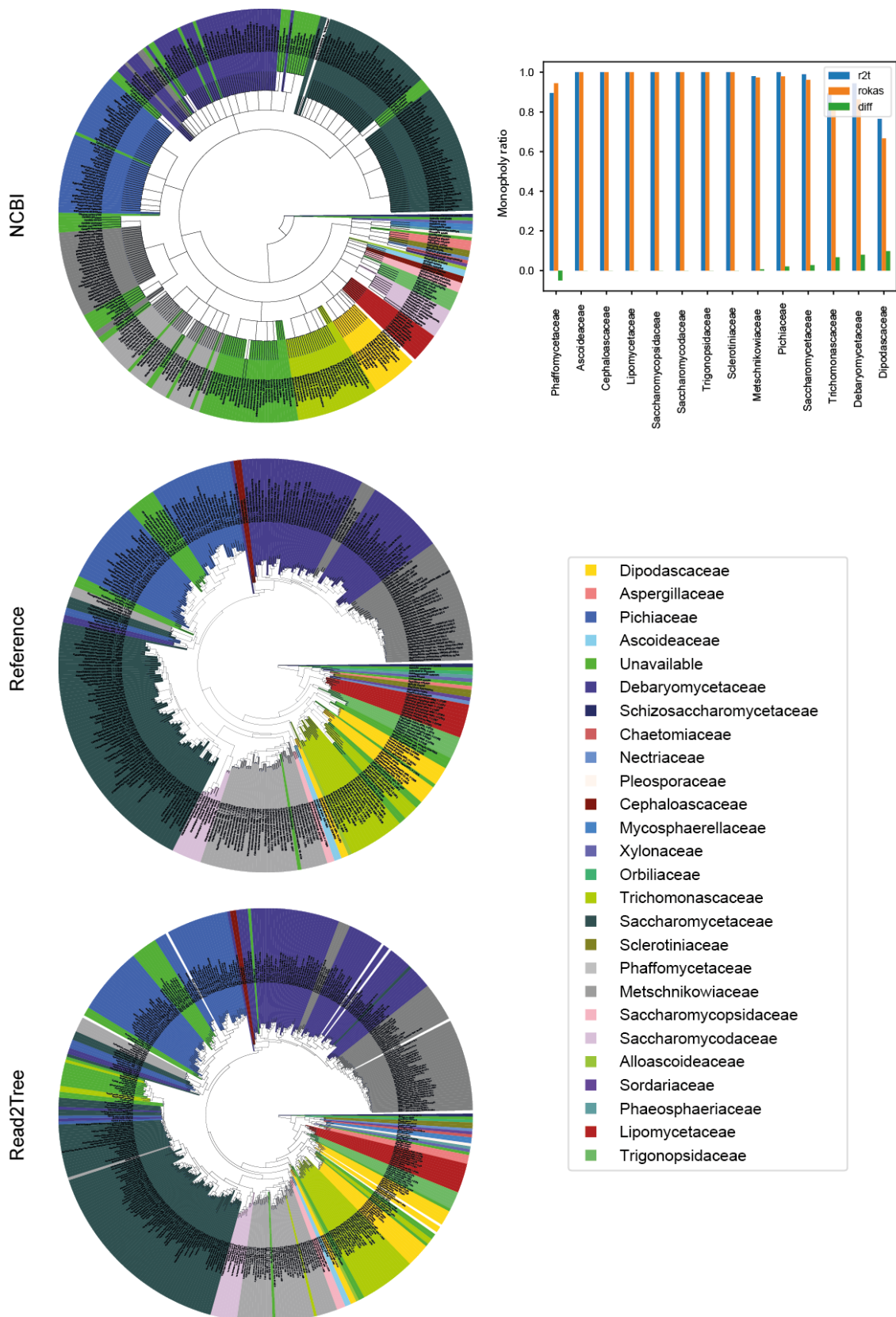
**Supplementary Figure 4.** Comparison of robinson foulds tree distance of Read2Tree against reference tree and assembly obtained tree against reference tree. Read2Tree shows similar performance across technologies, coverage levels and distance to the closest remaining ancestor.



**Supplementary Figure 5.** Proportions of sequences placed into the total number of OGs when selecting OGs with at least 80% taxa.

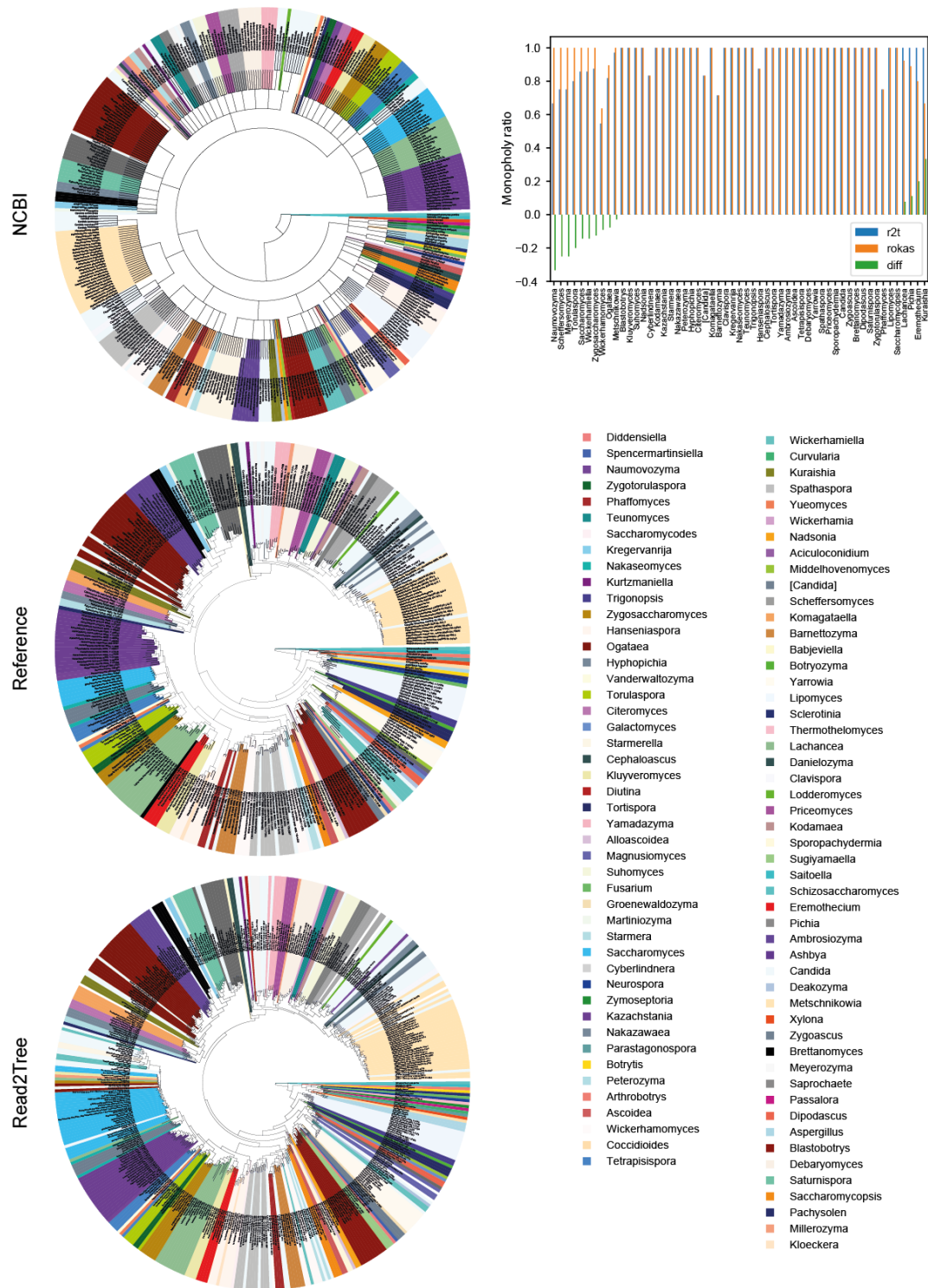


**Supplementary Figure 6.** Comparison of Trees by classification based on (Shen et al., 2018). Left NCBI, reference and Read2Tree inferred trees. Right Monophyly score computed for Shen *et al.* classification. Read2Tree is nearly as precise and the standard pipeline in recapitulating the right monophyletic groupings.

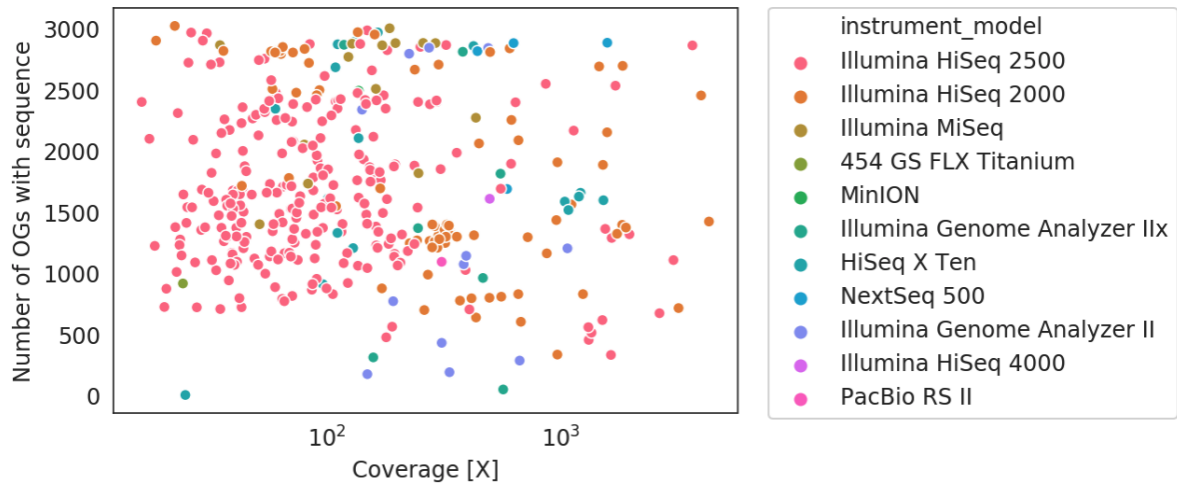


**Supplementary Figure 7.** Comparison of Trees by classification based on (Shen et al., 2018). Left NCBI, reference and Read2Tree inferred trees. Right Monophyly score computed for Shen *et al.* classification. Read2Tree is nearly as precise and the standard pipeline in recapitulating the right monophyletic groupings.

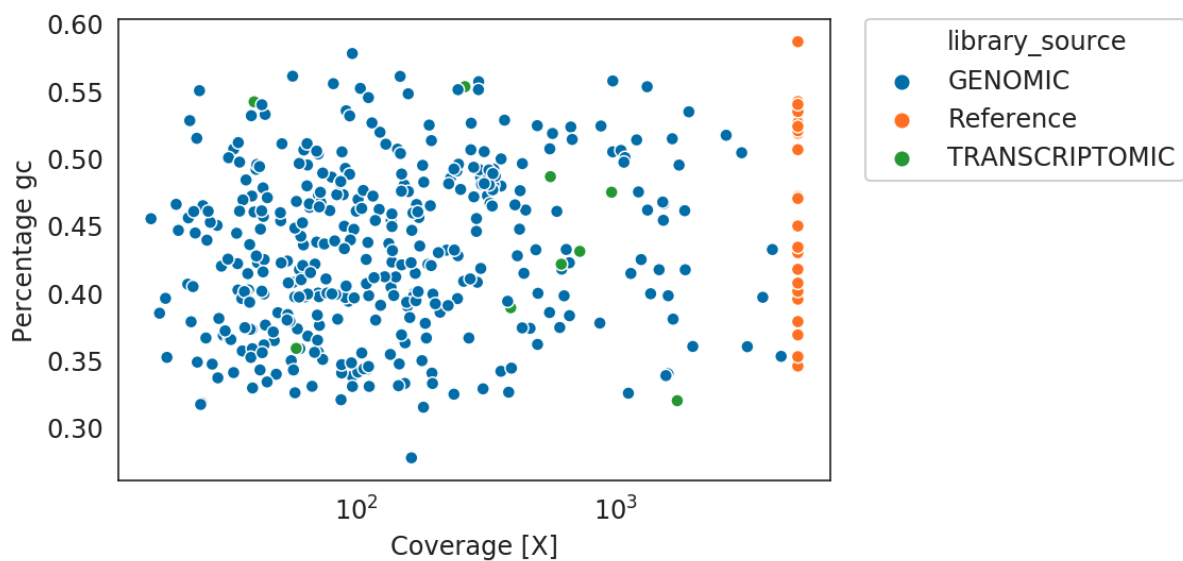




**Supplementary Figure 8.** Comparison of Trees by classification based on (Shen et al., 2018). Left NCBI, reference and Read2Tree inferred trees. Right Monophyly score computed for Shen *et al.* classification. Read2Tree is nearly as precise and the standard pipeline in recapitulating the right monophyletic groupings.



**Supplementary Figure 9.** Coverage to the number of obtained sequence relationships. Even for low coverage we see a large number of obtained sequences. No clear relationship is present between input coverage and number of sequences placed in OGs. In purple at 100X we display the number of sequences present in all ~3000 OGs for the reference species.



**Supplementary Figure 10.** Average percentage GC content in obtained sequences for yeast tree obtained using Read2Tree in comparison to given references. In most cases the reconstructed sequences are within the range of GC content as present in the reference sequences. References sequence artificially set to 5000X coverage for display purposes.

# Tables

## Reads used for species placement test

Species	Machine	BasePairs	Source	Accession
<i>A. thaliana</i>	PacBio	7.0 Gbp	DNA	ERR2173371
<i>A. thaliana</i>	MinION	3.4 Gbp	DNA	ERR2173373
<i>A. thaliana</i>	Illumina MiSeq	8.4 Gbp	DNA	ERR2173372
<i>M. musculus</i>	Illumina HiSeq	5.1 Gbp	mRNA	SRR5171076
<i>M. musculus</i>	PacBio	2.716 Gbp	mRNA	SRR5314792 SRR5314793 SRR5314794 SRR5314795 SRR5314796 SRR5314797 SRR5314798
<i>M. musculus</i>	MinION	0.7892 Gbp	mRNA	SRR4048177 SRR4048178 SRR4048179 SRR4095033 SRR4095035 SRR5286956 SRR5286957 SRR5286958 SRR5286959 SRR5286960 SRR5286961 SRR5286962 SRR5286963
<i>S. cerevisiae</i>	Illumina HiSeq	2.5 Gbp	DNA	SRR5892450
<i>S. cerevisiae</i>	PacBio	4.9 Gbp	DNA	SRR5989371
<i>S. cerevisiae</i>	MinION	1.129 Gbp	DNA	SRR5892449 SRR5924195

## Reference datasets

Species	#Ref	#OGs	#min species OG	Outgroup	#OGs with Species
MOUSE	30	699	26	<i>Amphimedon queenslandica</i>	698
YEAST	26	1188	21	<i>Monosiga brevicollis</i>	1133
ARATH	35	372	32	<i>Chlorella variabilis</i>	369

## Reads used for Figure 2 and 3

See additional file supplementary\_file\_1.xls (can be found here:  
[https://github.com/dvdylus/read2tree\\_paper/blob/main/supplement/](https://github.com/dvdylus/read2tree_paper/blob/main/supplement/))

## Reads used for Figure 4

See additional file supplementary\_file\_1.xlsx (can be found here:  
[https://github.com/dvdylus/read2tree\\_paper/blob/main/supplement/](https://github.com/dvdylus/read2tree_paper/blob/main/supplement/))

## All nwk strings for Figure 2 and 3

See additional file supplementary\_file\_2.xlsx (can be found here:  
[https://github.com/dvdylus/read2tree\\_paper/blob/main/supplement/](https://github.com/dvdylus/read2tree_paper/blob/main/supplement/))

## Reference dataset for mapping of 435 budding yeast species

Reference dataset was obtained from the OMA database (Altenhoff et al., 2021) using marker gene export function for 31 species with minimum coverage threshold of 0.8.

(supplementary\_file\_3.xlsx can be found here:  
[https://github.com/dvdylus/read2tree\\_paper/blob/main/supplement/](https://github.com/dvdylus/read2tree_paper/blob/main/supplement/))

## References

- Altenhoff AM, Train C-M, Gilbert KJ, Mediratta I, Mendes de Farias T, Moi D, Nevers Y, Radoykova H-S, Rossier V, Warwick Vesztrocy A, Glover NM, Dessimoz C. 2021. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res* **49**:D373–D379. doi:10.1093/nar/gkaa1007
- Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**:461–468. doi:10.1038/s41592-018-0001-7
- Sedlazeck FJ, Rescheneder P, von Haeseler A. 2013. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**:2790–2791. doi:10.1093/bioinformatics/btt468
- Shen X-X, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, Haase MAB, Wisecaver JH, Wang M, Doering DT, Boudouris JT, Schneider RM, Langdon QK, Ohkuma M, Endoh R, Takashima M, Manabe R-I, Čadež N, Libkind D, Rosa CA, DeVirgilio J, Hulfachor AB, Groenewald M, Kurtzman CP, Hittinger CT, Rokas A. 2018. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* **0**. doi:10.1016/j.cell.2018.10.023